Digitization of High-Stakes Exams

-

Empirical Insights and Design Recommendations for the Digital Execution and Scoring of Exams

Dissertation
for the award of the degree
Doctor rerum politicarum (Dr. rer. pol.)
of the Georg-August-University Göttingen

within the doctoral program *Wirtschaftswissenschaften*
of the Göttingen Graduate School of Social Sciences (GGG)

Submitted by
Philipp Hartmann, M.Sc. in Management
born in Linz am Rhein

Göttingen, 2023

Examination Committee

First Supervisor	*Prof. Dr. Matthias Schumann*
Second Supervisor	*Prof. Dr. Susan Seeber*
Third Supervisor	*Prof. Dr. Manuel Trenz*

Date of the disputation 29.08.2023

Acknowledgements

I would like to thank my doctoral supervisor, *Prof. Dr. Matthias Schumann*, for giving me the opportunity to pursue my doctorate and for all the invaluable knowledge I gained during my time at the Chair of Application Systems and E-Business. Furthermore, I would like to thank my second reviewer, *Prof. Dr. Susan Seeber*, from whom I was also able to learn a lot during the joint project work, as well as *Prof. Dr. Manuel Trenz* for the role of the third reviewer.

In addition, I am grateful for my colleagues at the Chair of Application Systems and E-Business, who have become good friends, if not a second family, to me over the years. I would like to thank *Dr. Jan Moritz Anke, Christian Finke, Michael Groth, Christine Jokisch, Dr. Kevin Koch, Aline Lange, Mustafa Pamuk, Dr. Tobias Nießner, Dr. Henrik Wesseloh, Lars Wilhelmi,* and *Dr. Steffen Zenker*. Special thanks go to my colleagues *Prof. Dr. Sebastian Hobert* for his professional support, as well as *Dr. Julian Busse, Dr. Pascal Freier,* and *Dr. Raphael Meyer von Wolff* for their mental support. Especially in the phases of a doctorate when things were not going well, I could always rely on you. Furthermore, special thanks go to our secretaries, *Nicole Fiedler-Gries* and *Paula Elena Pascaru*, for their invaluable help with administrative tasks throughout these years. I extend my gratitude to my colleagues from the field of Business Education, *Prof. Dr. Robin Busse*, *Dr. Carolin Geiser*, and *Hanna Meiners*, for our collaboration on various research projects.

Moreover, I am grateful for all the people I met during my time at the University of Göttingen, who enriched my time during the bachelor’s and master’s degree. I particularly want to thank *Björn Bos* and *Adriana Niechoy*, without whom I wouldn't be where I am today, and who made sure that we had some fun along the way.

The biggest thank goes to my family, especially my parents, *Marlies and Johannes*, who didn't always have it easy with me but always supported me. Thank you for always believing in me and encouraging me.

Lastly, I want to express my heartfelt gratitude to my better half, *Lilia*. Without my doctorate, we might not have crossed paths, and your unwavering support helped me to get through the challenges of this journey. Thank you for the countless beautiful moments we shared together, which gave me strength throughout these years.

Göttingen, 29.08.2023 Philipp Hartmann

Abstract of Content

Abstract of Content IV

Table of Content V

List of Figures VII

List of Tables VIII

List of Abbreviations X

A Foundations 1

1 Motivation and Relevance 1

2 Theoretical Foundations 3

3 Research Objectives and Structure 16

B Conducted Research Studies 20

1 Execution of Emergency Remote Exams 20

2 Lessons Learned from Emergency Remote Exams 34

3 Examinee's Trust in AI-based Essay Scoring 54

4 Design of AI-based Essay Scoring Systems 68

5 Design of AI-based Scoring Explanation 86

C Contributions 100

1 Discussion and Implication 100

2 Conclusion 110

3 Limitations and Future Perspectives 113

Appendix 117

References XII

Assurance Upon Admission of the Doctoral Examination XXX

Overview of Author Contribution on the Conducted Research Studies XXXI

Table of Content

Abstract of Content .... **IV**

Table of Content .... **V**

List of Figures .... **VII**

List of Tables .... **VIII**

List of Abbreviations .... **X**

A Foundations **1**

1 Motivation and Relevance 1

2 Theoretical Foundations 3

2.1 Foundations of Digital High-Stakes Exams in Education 3

2.2 Related Research 6

2.2.1 Methodological Approach 6

2.2.2 Literature Review and Discussion 8

3 Research Objectives and Structure 16

B Conducted Research Studies **20**

1 Execution of Emergency Remote Exams 20

1.1 Introduction 21

1.2 Related Research and Hypotheses Development 22

1.3 Research Design 24

1.4 Results 27

1.5 Discussion and Implication 29

1.6 Limitation 32

1.7 Conclusion 33

2 Lessons Learned from Emergency Remote Exams 34

2.1 Introduction 35

2.2 Related Research 36

2.3 Research Design 38

2.4 Results 40

2.5 Discussion, Implication & Limitation 48

2.6 Conclusion 52

3 Examinee’s Trust in AI-based Essay Scoring 54
3.1 Introduction 55
3.2 Related Research and Hypotheses Development 56
3.3 Research Design 59
3.4 Results 62
3.5 Discussion and Implication 64
3.6 Limitation 65
3.7 Conclusion 66
4 Design of AI-based Essay Scoring Systems 68
4.1 Introduction 69
4.2 Related Research on Semi-Automatic Essay Scoring 70
4.3 Research Design 72
4.4 Design and Evaluation 73
4.5 Discussion and conclusion 83
5 Design of AI-based Scoring Explanation 86
5.1 Introduction 87
5.2 Related Research and Hypotheses Development 87
5.3 Research Design 90
5.4 Results 94
5.5 Discussion, Implications & Limitations 96
5.6 Conclusion 98
C Contributions 100
1 Discussion and Implication 100
2 Conclusion 110
3 Limitations and Future Perspectives 113
Appendix 117
References XII
Assurance Upon Admission of the Doctoral Examination XXX
Overview of Author Contribution on the Conducted Research Studies XXXI

List of Figures

Figure 1. CIA Triade 1

Figure 2. Digital Assessment Process 5

Figure 3. Literature Review Framework 6

Figure 4. Literature Search Process 7

Figure 5. Structure of the Dissertation 17

Figure 6. Knowledge and Question Types 41

Figure 7. Trust Model Used for this Research Study 56

Figure 8. Results of the Structural Equation Model 63

Figure 9. Adapted Research Approach 73

Figure 10. Overview of Requirements and Design Principles 74

Figure 11. Screenshot of the Manual Scoring Front-end 77

Figure 12. Results of the UEQ (n = 13) 78

Figure 13. Implementation of the Scoring Mechanism 79

Figure 14. Success Quotas of the Cross-Validation 81

Figure 15. Conceptual Research Model 90

Figure 16. User Interface Designs 92

Figure 17. Mean Plot 95

List of Tables

Table 1. Selected Definition of the Term Digital Assessment 3

Table 2. Positioning of the Literature Analysis 6

Table 3. Relevance Criteria of the Literature 7

Table 4. Results of the Literature Review 8

Table 5. Overview of the Conducted Research Studies 19

Table 6. Summary of Participants 27

Table 7. Results of the Factor Analysis 28

Table 8. Results of the Correlation Analysis 29

Table 9. Summary of Participants 40

Table 10. Descriptive Results 45

Table 11. Correlation Matrix 46

Table 12. Recommendations for Action 51

Table 13. Reliability Coefficients of the Factors and Items Used 61

Table 14. Descriptive Results of AI-Related Factors 63

Table 15. Documentation of the Design Principles based on Gregor et al. (2020) 83

Table 16. Reliability Coefficients of the Factors and Items Used 94

Table 17. Descriptive Results of Trustworthiness-Related Factors 94

Table 18. ANOVA Results 95

Table 19. Results of the Moderated Parallel Mediation Analysis 96

Table 20. Recommendations for Practice - Studies I and II 104

Table 21. Recommendations for Practice - Study III 106

Table 22. Recommendations for Practice - Study IV 107

Table 23. Recommendations for Practice - Study V 109

Table 24. Recommendations for Practice - Overview 110

Table 25. Questionnaire Studies I and II - Part 1 118

Table 26. Questionnaire Studies I and II - Part 2 119

Table 27. Questionnaire Studies I and II - Part 3 120

Table 28. Guideline Follow-Up Interview Study II 121

Table 29. Questionnaire Study III - Part1 122

Table 30. Questionnaire Study III - Part 2 ... 123

Table 31. Questionnaire Study III - Part 3 ... 124

Table 32. Questionnaire Study III - Part 4 ... 125

Table 33. Questionnaire Study V - Part 1 ... 126

Table 34. Questionnaire Study V - Part 2 ... 127

List of Abbreviations

AES	Automated Essay Scoring
AI	Artificial Intelligence
AIS	Association for Information Systems
AMCIS	Americas Conference on Information Systems
API	Application Programming Interface
ATI	Affinity for Technology Interaction
BYOD	Bring-Your-Own-Device
C	Challenge
CA	Cronbach’s Alpha
CAT	Computerized Adaptive Test
CBA	Computer-based Assessment
CBAAM	Computer-based Assessment Acceptance Model
CIA	Curriculum-Instruction-Assessment
CSS	Cascading Style Sheet
DP	Design Principle
DSR	Design Science Research
DUE	Download-Upload Exams
ECIS	European Conference on Information Systems
GDPR	General Data Protection Regulation
HTML	Hypertext Markup Language
ICT	Information and Communication Technology
KMO	Kaiser-Meyer-Olkin
LLM	Large Language Model
LSTM	Long Short-Term Memory
MD	Mean Deviation
MOOC	Massive Open Online Course
NLP	Natural Language Processing
PBA	Paper-based Assessment
PC	Personal Computer
R	Requirement

RA	Recommendations for Action
RP	Recommendation for Practice
RQ	Research Question
SD	Standard Deviation
SE	System Exams
TAM	Technology Acceptance Model
TEA	Technology-enhanced Assessment
TPB	Theory of Planned Behavior
U	User Story
UEQ	User Experience Questionnaire
UTAUT	Unified Theory of Acceptance and Usage of Technology
VBA	Visual Basic for Applications
XAI	Explainable Artificial Intelligence

A Foundations

1 Motivation and Relevance

Digitization in the field of education has long been a topic of interest in research and practice. According to the Curriculum-Instruction-Assessment (CIA) triad of PELLEGRINO (2010), educational systems (to be understood in this context as a social system, not a technical system) can be divided into three linked core components (see Figure 1). First, the curriculum anchored subject matter areas, which include the learning content, and the competencies to be promoted. Second, the instruction for promoting these content-related competencies, representing the teaching methods as well as the learning activities. Third, the assessment of the learning outcome based on the learning content and competencies addressed. Since these three components are linked, their specific design must be aligned.

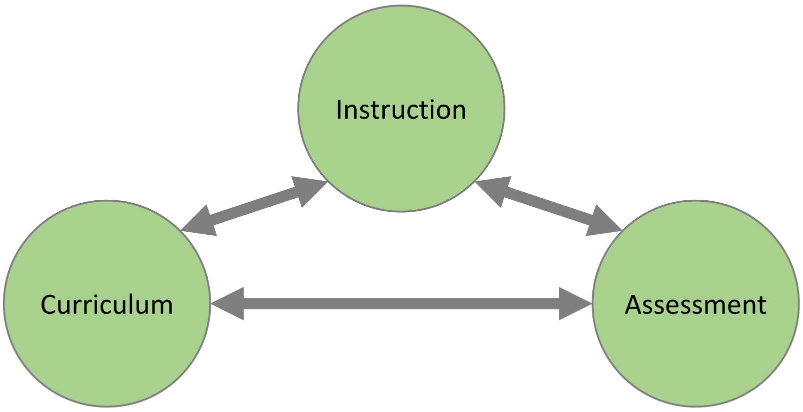


Figure 1. CIA Triade

The impact of digitization on these three components of the CIA triad has changed over time. At the beginning of Information and Communication Technology (ICT) use in education, efficiency advantages were primarily investigated in the context of instruction and assessment (Nicholson 2007; Parhizgar 2012). In the area of instruction, the distribution and management of learning content were addressed in particular. Digital delivery allows content to be (re-)used regardless of time and place, which reduces costs compared to in-class teaching (Weller 2004). Furthermore, it enables learners to access the content flexibly and institutions to offer their courses to a wider audience without capacity limits (Nicholson 2007). In the area of assessment, a first attempt was made by IBM in 1935 to have computers evaluate closed question types (especially multiple-choice questions) in Paper-Based Assessment (PBA) to reduce costs (Parhizgar 2012). Due to technological progress, further steps of the Computer-Based Assessment (CBA) processes, such as exam distribution, scoring, and reporting, were investigated (Burstein et al. 1996; Fulcher 2000). However, the first CBA were "simply paper-and-pencil tests delivered through the new electronic medium" (Fulcher 2000). In addition to the cost advantages, content-related potentials have become increasingly important in recent years. For example, the accelerating digital transformation is changing the qualification requirements in the professional environment (Bertelsmann Stiftung 2020; German Chamber of Commerce and Industry 2020). Routine tasks are more and more being taken over by (partially) automated technologies (Bach et al. 2022; Dengler / Matthes 2015, 2018). Thus, employees must cope with increasingly complex occupational problems (Becker 2015; Hermann et al. 2017; Rausch et al. 2019). These developments are expected to accelerate due to significant technological progress in Artificial Intelligence (AI; for example, through Large Language Models (LLM)). To promote these problem-solving competencies, digitization is increasingly being addressed in the curriculum anchored subject matter areas and the associated instruction (Standing Conference of the Ministers of Education and Cultural Affairs 2022). In instruction, the digital implementation of learning content enables the use of multimedia elements and simulations. This is said to have the potential to improve the learning outcome by providing an authentic setting (Lombardi / Oblinger 2017). Even though, according to the CIA triad, all three components must be aligned, the focus in both practice and research continues to be primarily on digital teaching and learning. Digitization continues to be addressed only superficially in practice, even though many of these content-related potentials apply not only to digital learning but also to digital assessment (Alruwais et al. 2018; Butler-Henderson / Crawford 2020). In its annual report, the STANDING CONFERENCE OF THE MINISTERS OF EDUCATION AND CULTURAL AFFAIRS (2022) highlighted the need for increased evaluation and introduction of digital devices in assessments. While there is a need to advance digital assessment capabilities, the legal framework and the competencies of examiners to execute digital assessments are often still lacking (Ständige Wissenschaftliche Kommission 2022).

An important aspect in this context was the COVID-19 pandemic, which proved to be a catalyst for the digitization of assessments. The temporary closure of educational institutions and the resulting suspension of in-class education led to the rapid establishment of digital infrastructures and legal frameworks for digital assessments. The questions that arise from this rapid change are to what extent processes and infrastructures implemented during the pandemic will be retained in the long term. This results in a unique situation where institutions do not have to discuss the introduction of new processes and systems, but rather the continued use and further development of existing ones. However, whether the potentials of digital assessments will be realized depends on the extent to which they are actually used. In addition to the fundamental decision on use by institutions, usage also depends on the acceptance of examiners and examinees. Therefore, this dissertation will focus on digital assessments from the perspective of examiners as well as examinees. The goal is to provide empirical insights and recommendations for all stakeholders on promoting the use of digital high-stakes exams.

2 Theoretical Foundations

This section will introduce the theoretical foundations of this dissertation. First, the foundations of digital assessments in education are addressed to ensure a uniform understanding of the topic. For this purpose, a definition is derived, and a distinction is made regarding different approaches and characteristics of assessment in education (Section A.2.1). Second, the related research is presented to determine the core topics of the research and to identify research gaps. For this purpose, the methodological approach of the literature review is presented, followed by classification and discussion of the identified literature (Section A.2.2).

2.1 Foundations of Digital High-Stakes Exams in Education

The term digital assessment is often used in education to refer to examinations that are conducted entirely with the aid of ICT (Alruwais et al. 2018). In this context, there is a variety of related terms, some of which are used as synonyms (or terminological adaptation over time) or differ only in terms of specific characteristics (Timmis et al. 2016). These variations relate to both digitization (e.g., computer-based, electronic, technology-enhanced, and digital) and to assessment (e.g., assessment, exam, and test). In the following, an overview of selected definitions of digital assessments is given (see Table 1), and their core characteristics are discussed to achieve a common understanding of the term digital assessment.

Author	Definition
JOINT INFORMATION SYSTEMS COMMITTEE (2007)	"E-Assessment is the **end-to-end electronic assessment processes** where ICT is used for the **presentation of assessment activity, and the recording of responses**. This includes the end-to-end assessment process from the perspective of learners, tutors, learning establishments, awarding bodies and regulators, and the general public."
JOHAR / KUMAR (2016)	"Computer-based assessment (CBA) is a **method of administering tests** in which the **responses are electronically recorded and assessed** with the aid of dynamic visuals, sound, user interactivity, as well as adaptivity to individual test-takers and near real-time score reporting."
TIMMIS ET AL. (2016)	"We define TEA [technology-enhanced assessment] to include **any use of digital technologies** for the purposes of **enhancing formal or informal educational assessment** for both formative and summative purposes."

Table 1. Selected Definition of the Term Digital Assessment

The JOINT INFORMATION SYSTEMS COMMITTEE (2007) defines e-assessment as the complete electronic execution of assessments. They explicitly mention the distribution of assessment tasks from the examiner to the examinee and the distribution of answers from the examinee to the examiner. However, it is also stated that the entire assessment process should take place electronically for all stakeholders involved. JOHAR / KUMAR (2016) define CBA as a general method for administering tests in which the answers are recorded and evaluated electronically. The digital implementation of assessments consists of dynamic and adaptive media elements. The scoring of the answers is done in near real-time. TIMMIS ET AL. (2016) use the term technology-enhanced assessment to refer to any use of digital technologies in educational assessments. While the use of technology to improve the assessment is not further specified, the use for formative and summative assessment is addressed.

Despite the differences in terminology, the definitions show that the core characteristics of digital assessment in education are largely consistent. The first core characteristic is the use of digital technologies, although no specifics are given regarding the nature of these technologies. The second core characteristic is the support digital technologies offer in the assessment process, which varies between definitions. While the JOINT INFORMATION SYSTEMS COMMITTEE (2007) focuses on the digital distribution of exam tasks and answers, JOHAR / KUMAR (2016) also include support for scoring through the use of technology. TIMMIS ET AL. (2016) simply state that the use of technology improves assessment processes. It is important to note that although the JOINT INFORMATION SYSTEMS COMMITTEE (2007) only refers to the distribution of exam tasks and answers between examinees and examiners, they emphasize that the entire assessment process for other stakeholders is also included in their understanding of e-assessment. Therefore, scoring by examiners is also included in this case. Based on the definitions of JOHAR / KUMAR (2016) and TIMMIS ET AL. (2016), in the remainder of this dissertation, the term digital assessment can be understood as follows:

> Digital assessment is defined as any use of digital technologies to enhance the recording and scoring of responses in educational assessments.

Assessment Approaches

Regardless of the use of technology, assessments can differ in purpose. A distinction can be made between diagnostic, formative, and summative educational assessment (Joint Information Systems Committee 2007).

- **Diagnostic Assessment** describes the evaluation of individual performance before the start of the learning process to plan the required learning content (Joint Information Systems Committee 2007). It serves as a reference point for further assessment of performance as part of the educational process (Chufama / Sithole 2021).

- **Formative assessment** describes the evaluation of individual performance at a specific point in time during the learning process and serves to support the learner (Chufama / Sithole 2021; Harlen 2005; Taras 2005). It is not only based on certain objective performance criteria but also on individual progress (Harlen / James 1997). Therefore, formative assessment is also referred to as assessment for learning (Joint Information Systems Committee 2007).

- **Summative Assessment** describes the evaluation of individual performance at the end of a learning process and is used, among other things, for reporting to different stakeholders (Chufama / Sithole 2021; Harlen 2005; Taras 2005). It is based on defined public criteria without considering individual progress (Harlen / James 1997). Therefore, summative assessment is also called assessment of learning (Joint Information Systems Committee 2007).

Diagnostic and formative assessment thus serve examinees during the instruction phase as preparation for the assessment phase. The remainder of this dissertation will focus on digital exams in summative assessments.

Due to their reporting function based on public criteria, summative assessments are usually given higher importance and are often referred to as high-stakes exams (Joint Information Systems Committee 2007). The stake of an exam is determined by the relevance of its results for the examinee. A distinction can be made between low-stakes and high-stakes exams (Joint Committee on Standards for Educational and Psychological Testing 2014). Due to the increased importance of high-stakes exam scores, examiners must show additional evidence that the assessment serves its intended purpose (Joint Committee on Standards for Educational and Psychological Testing 2014). This may involve collecting collateral information (i.e., factors that influence the exam result) so that the appropriate interpretation of the exam results is possible (Harlen 2005). Here, the digital execution and scoring of exams enable the standardization of planning, execution, and assessment of large-scale exams, thereby improving both comprehensibility and test efficiency (Schmidgall / Powers 2017). Therefore, the following of this dissertation will focus on written high-stakes exams.

Digital Assessment Process

To systematize the related research in Section A.2.2 and derive the recommendations in Section C.1, the process of digital high-stakes exams is interpreted as a three-stage process from the examinee's perspective (see Figure 2).

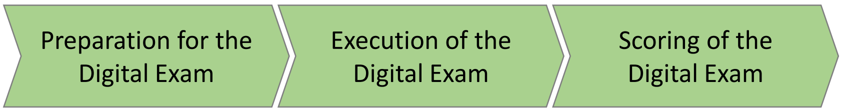

Preparation for the Digital Exam
Execution of the Digital Exam
Scoring of the Digital Exam

Figure 2. Digital Assessment Process

The first stage of the process is the preparation for the digital exam, which includes both the preparation of the content and the organizational preparation of the examiners as well as the examinees. This stage is excluded from the literature review since this preparation mainly addresses the instructional phase and this dissertation focuses on summative assessments. Nevertheless, single recommendations for this stage are also derived from studies presented in Part B, in so far as this is suitable. However, these recommendations will not relate to content but to organizational preparation. The second stage consists of executing the written digital exam. Here, the exam tasks are presented to and answered by the examinees. The correctness of the answers determines whether the learning objective has been achieved. The third stage is the digital scoring of exams, which includes both the computer-based scoring and digital presentation of the results to the examiners and examinees.

2.2 Related Research

In this section, research related to the execution and scoring of digital high-stakes exams is identified. Section A.2.2.1 includes the underlying methodological procedure of the structured literature analysis. Then, in Section A.2.2.2, the related research is presented and discussed.

2.2.1 Methodological Approach

The determination of the related research of the execution and scoring of digital high-stakes exams is based on a structured literature review following the approach of vom Brocke et al. (2009), Fettke (2006), and Webster / Watson (2002). According to Webster / Watson (2002), an effective literature review enables the identification of research gaps and thus forms the basis for building further knowledge. The process is shown in Figure 3.

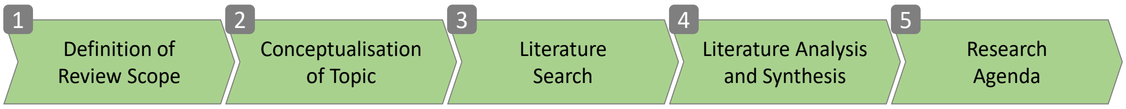

1
Definition of Review Scope
2
Conceptualisation of Topic
3
Literature Search
4
Literature Analysis and Synthesis
5
Research Agenda

Figure 3. Literature Review Framework

In the first phase, the definition of the review scope was determined following Fettke (2006) and vom Brocke et al. (2009) (see Table 2). The goal was to capture central issues related to the digitization of high-stakes exams to present a cross-section of the research domain. Even though the categories of integration and criticism were not explicitly identified as a central goal, a critical reflection of the literature and integration into the examination process took place. By identifying central issues, a selective coverage of the related literature was aimed. A comprehensive presentation of the specific selection aspects is shown in the third phase of this process. The focus of the review was on the research outcomes, which are needed to identify existing research gaps and to derive the need for research. The structure of the literature analysis was based on the digital assessment process presented in Section A.2.1.

<table>
<tr><th>Characteristic</th><th colspan="12">Categories</th></tr>
<tr><td><b>Goal</b></td><td colspan="4"><i>Integration</i></td><td colspan="4"><i>Criticism</i></td><td colspan="4"><i>Central Issues</i></td></tr>
<tr><td><b>Coverage</b></td><td colspan="3"><i>Central/Pivotal</i></td><td colspan="3"><i>Representative</i></td><td colspan="3"><i>Selective</i></td><td colspan="3"><i>Exhaustive</i></td></tr>
<tr><td><b>Focus</b></td><td colspan="3"><i>Applications</i></td><td colspan="3"><i>Theories</i></td><td colspan="3"><i>Research methods</i></td><td colspan="3"><i>Research outcomes</i></td></tr>
<tr><td><b>Structure</b></td><td colspan="4"><i>Historical</i></td><td colspan="4"><i>Conceptual</i></td><td colspan="4"><i>Methodological</i></td></tr>
</table>

Table 2. Positioning of the Literature Analysis

In the second phase, the topic was conceptualized. To structure the results of the literature analysis thematically, the domain was systematized based on the digital assessment process (see Figure 2) defined in Section A.2.1. In addition, further exploratory structuring into the technical and the user-centric perspectives was undertaken based on the identified literature.

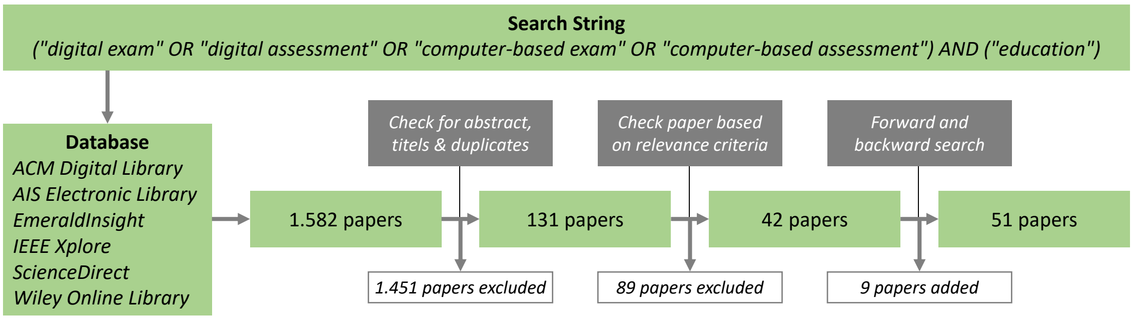
The third phase involved the literature search. The domain of digitization in high-stakes exams was addressed using the search string *("digital exam" OR "digital assessment" OR "computer-based exam" OR "computer-based assessment")*. In addition, the search was restricted to the education sector (search term: *"education"*). The resulting search string was used to search for literature in the databases shown in Figure 4. Furthermore, a backward and forward approach was conducted by reference and author searches. Figure 4 presents an overview of the literature search process, which took place in December 2019. A total of 1,582 contributions were identified and evaluated based on their relevance. The relevance of the literature was evaluated based on specified relevance criteria (see Table 3) following the approach of VOM BROCKE ET AL. (2015). A paper was considered relevant if it met all four criteria.

Relevance Criteria	**Description** *Relevant are articles that focus on…*
1	… the assessment of skills and competencies in an educational context.
2	… the digitization of the assessment process.
3	… high-stakes exams.
4	… written exams.

Table 3. Relevance Criteria of the Literature

Relevant articles were those that dealt with the assessment of skills and competencies in an educational context. The intention was to exclude literature in which digital assessment refers to other areas of application (e.g., during patient examination in the medical context). In addition, the search targeted research on the digitization of the assessment process to delineate articles that target digital literacy. Lastly, a focus was placed on high-stakes and written exams to exclude formative application scenarios that are more likely to be assigned to digital learning.

Overall, 51 contributions remained as the basis for determining the related research. The analysis of the identified literature (fourth phase) is presented in Section A.2.2.2. The research gaps, and the further research agenda for this work (fifth phase), are presented in Section A.3.



Search String
("digital exam" OR "digital assessment" OR "computer-based exam" OR "computer-based assessment") AND ("education")
Database
ACM Digital Library
AIS Electronic Library
EmeraldInsight
IEEE Xplore
ScienceDirect
Wiley Online Library
Check for abstract, titels & duplicates
Check paper based on relevance criteria
Forward and backward search
1.582 papers
131 papers
42 papers
51 papers
1.451 papers excluded
89 papers excluded
9 papers added

Figure 4. Literature Search Process

2.2.2 Literature Review and Discussion

Following, the results of the literature review will be discussed, and structured according to the two addressed stages of the digital assessment process (see Section A.2.1). Table 4 provides an overview of the identified literature.

	Content Focus	
Process Stage	Technical	User-Centric
Execution	Burlak et al. (2006); Fluck et al. (2009); Higgins et al. (2002); Higgins et al. (2006); Hillier / Fluck (2013); Ju et al. (2018); Kaya / Özel (2014); Kleerekoper / Schofield (2019) Kleinhans / Schumann (2015); Laubscher et al. (2005); Lazarinis et al. (2010); Opgen-Rein et al. (2018); Piech / Gregg (2018); Sindre / Vegenda (2015); Tinoca (2012); Wiannastiti et al. (2018)	Backes / Cowan (2019); Boevé et al. (2015); Čandrlić et al. (2014); Delotach et al. (2016); Dermo (2006); Grissom et al. (2016); Hillier (2014); Hillier (2015); Jeong (2014); Kalogeropoulos et al. (2013); Maguire et al. (2010); Matthíasdóttir / Arnalds (2016); Miller (2011); Piaw (2012); Prisacari / Danielson (2017); Shermis / Lombard (1998); Stephenson (2018); Terzis / Economides (2011); Terzis et al. (2012); Terzis et al. (2013); Wise (2019)
Scoring	Attali / Burstein (2005); Attali / Powers (2008); Burstein et al. (2004); Buyrukoglu et al. (2019); Higgins et al. (2002); Higgins et al. (2005); Kerr et al. (2013); Lajis / Aziz (2010); Lajis / Aziz (2012); Mohler / Mihalcea (2009); Prados et al. (2011); Quinlan et al. (2009); Shermis (2014); Shermis (2015); Summons (1997)	---

Table 4. Results of the Literature Review

In the area of the execution of digital exams, both the technical perspective and the user-centric perspective were identified in the relevant literature. The technical perspective was addressed in 16 research articles, while the user-centric perspective was addressed in 22 research articles. For the scoring of digital exams, 15 research articles from the technical perspective were identified, whereas no research contributions were found from the user-centric perspective.

Execution of the Digital Exam – Technical Perspective

The focus of the related research in this area is on the development of different digital examination systems. Three publications deal with design recommendations for e-assessment systems in general. Tinoca (2012) developed and discussed a conceptual framework for e-assessment with the goal of improving exam quality and simplifying the design. The framework consists of four dimensions, namely, authenticity, consistency, transparency, and practicability. Hillier / Fluck (2013) identified drivers and requirements for e-assessment systems for high-stakes exams. In this context, the term "system" includes not only the technical component but also the associated processes and policies. A total of 13 requirements were assigned to the areas of students, teaching and pedagogy, as well as institutions. Based on these requirements, 20 functionalities and strategies were derived for the system implementation. Fluck et al. (2009)

investigated the general use of a CD-based e-assessment system for Bring-Your-Own-Device (BYOD) exams, which could be taken on students' own devices without internet access. The researchers found that examinees were indifferent concerning the use of paper-based and computer-based exams. In addition, it was shown that a positive previous experience regarding CBA positively influences the preference for future use. Five other publications focus on individual application domains in system development, in which a specific competence is tested. The primary application domain in the identified literature is the digital assessment of ICT-related competencies. PIECH / GREGG (2018) developed an assessment tool inspired by an authentic coding environment. The tool replicated the basic features of a coding tool, without having a compile or run feature. The authors cited the potential use of an automatic scoring mechanism as a long-term advantage of conducting digital exams. A similar approach was taken by HIGGINS ET AL. (2002) and HIGGINS ET AL. (2005) with their (further) development of a combined exam and scoring tool for coding tasks. Again, examinees were provided an authentic coding environment through which answers for coding tasks could be entered into the system. JU ET AL. (2018) also aimed to provide a coding exam system where examinees would have access to internet resources, but direct communication with other persons would be blocked. In addition, hints for the correct answer could be used against the deduction of points. The evaluation showed that more than half of the examinees did not use any or only a few hints. It was also observed that, despite the proctored environment, attempts were made to cheat using the internet. The majority of examinees also stated that the exam score accurately reflected their respective coding skills. By contrast, WIANNASTITI ET AL. (2018) developed and evaluated an integrated multimedia website for a writing test. The focus was on evaluating different design elements of the user interface and their usefulness in writing tests.

In addition to the design of digital exam systems, two identified papers also addressed the structural design of individual examinations. While PBA use a linear procedure, digital exams also allow the use of Computerized Adaptive Tests (CAT), where the questions can be adapted to the examinee's level of competence during the test (Kleinhans / Schumann 2015). KLEINHANS / SCHUMANN (2015) implemented and evaluated a CAT instrument using an example case in the health sector. An increased measurement efficiency could be demonstrated by reducing the required items by 40 %. Deviations arose at high or low competence levels of examinees, which required a high number of items to exclude possible measurement inaccuracies. In this context, LAZARINIS ET AL. (2010) investigated the factors and adaptation rules that influence the selection of the respective following task. In summative assessments, these included prior knowledge and performance.

The last aspect identified in the conduction of digital exams is securing the integrity of exams, which is identified in six publications. Monitoring software should always aim to detect cheating effectively without interfering with the examinee's ability to complete tasks (Laubscher et al. 2005). Sindre / Vegendla (2015) compared the threats and countermeasures against cheating for BYOD exams and PBA. They concluded that neither mode of exam is per se worse nor better in this context. Thus, both modes have specific cheating risks that require respective countermeasures. Many known risks associated with PBA can be transferred to BYOD exams. Burlak et al. (2006) used data mining to identify examinees who cheated during an online assessment. Three examinee types were identified: advanced students, average students, and cheaters. The individual examinee types were characterized by the response quality and the required response time. In addition, several publications investigated different ways to detect cheating in exams. Laubscher et al. (2005) investigated the potential use of keyloggers to identify different types of cheating. A different approach to cheating detection was taken by Opgen-Rhein et al. (2018), who developed an AI-based tool to review written task answers. The tool identified the examinee's writing style based on previously collected reference work and compared the writing style across different answers. This should prove whether the task was completed independently. Kaya / Özel (2015) developed and implemented a code plagiarism detection tool. The tool determined the similarity between the answers of different examinees using k-grams. A higher similarity was considered an indication of cheating. Kleerekoper / Schofield (2019) investigated the false-positive rate of different automated plagiarism detection algorithms for SQL tasks. Due to the shortness and limited variety of answers, false positives are more likely to occur for SQL answers. The results showed that especially shorter answers have a higher false-positive rate than long ones. Even with the strictest method, 15 % of the answers were falsely flagged.

In the following, the results of the literature review on the digital execution of exams from a technical perspective are discussed. In this field, the research focused on ICT-related subjects and detecting cheating in digital exams. Not unexpectedly, the literature on digital exams focused on ICT-related subjects. Here, digital exams make it possible to provide an authentic exam environment. For coding tasks, it was found that PBA is not a suitable approach to assess the respective competencies learned. This is particularly true if the learning phase already involves application-oriented teaching. In the case of cheating, it was shown that in addition to the known types of cheating from PBA, further types arise through the digital conduction of exams. However, CBA also offers the possibility of detecting different types of cheating at this point. It is important to note that digital exams must also implement careful cheating detection mechanisms.

Execution of the Digital Exam – User-Centric Perspective

The user-centric perspective addresses both examiners' and examinees' perceptions of CBA. Eleven publications were found to address individual factors influencing user perception of digital exams. Wise (2019) examined how CBA can help identify and address construct-irrelevant factors in exams. Construct-irrelevant factors are described as influences on the examinee that bias the true exam score (e.g., disengagement, test anxiety, and cheating). Thus, information on the exam environment can be collected and framed through CBA. DERMO (2009) examined examinees' perceptions of e-assessment, finding that participants rated CBA as suitable based on affective factors, validity, practicality, reliability, and security. There were some concerns regarding the fairness of randomized item banks, which were perceived as unfair because the assignment of tasks is random. HILLIER (2014) and HILLIER (2015) showed that ICT-related study programs are more open to digital exams. This contrasted with study programs whose classic exam tasks were difficult to replicate digitally (e.g., due to extensive drawings that must be made). In addition, concerns about cheating were expressed. According to the author, these concerns often arise because tools for formative assessment are frequently used for summative exams without being tested for suitability. Furthermore, examinees are not familiar with the execution of digital exams at first use and therefore report higher levels of stress and difficulty. Another aspect was fairness, which was often questioned by inexperienced examinees. Relevant factors regarding the low level of perceived fairness were cheating, lack of technical equipment, and lack of familiarity. Another focus within the literature in this area is the use of CBA for coding tasks. STEPHENSON (2018) conducted a qualitative study among examinees of a coding course. Overall, CBA was evaluated as a fair and appropriate delivery method. Thus, CBA increased perceived authenticity through the ability to use compile or run features and learning materials in the specific course. However, negative effects of the described exam execution were also reported. Examinees complained about a lack of time to complete the exam tasks, leading to concerns that the exam measured time management rather than competence. Increased stress, anxiety, and pressure were also observed among individual examinees. This was attributed to the permitted aids, as these increased the required quality of the overall solutions. MATTHÍASDÓTTIR / ARNALDS (2016) came to similar conclusions. In their examination of the computer-based execution of a coding exam, examinees were more satisfied with the digital execution, as it was perceived as more authentic for task completion. However, the issue of time management was also raised. DELOATCH ET AL. (2016) focused on anxiety among examinees in coding tasks, comparing PBA and CBA. Their results show no significant differences in anxiety regarding the mode of exam. The researchers also asked examinees about the specific influence of anxiety on the exam process. SHERMIS / LOMBARD (1998) showed that test anxiety partially negatively influences digital assessment performance and that the frequently studied computer anxiety actually expressed general test anxiety. PRISACARI / DANIELSON (2017) examined the differences in cognitive load between PBA and CBA without finding significant differences. The investigation studied algorithmic, conceptual, and definition-based tasks, as well as the effect of scratch paper. The highest rate of scratch paper use was found for algorithmic tasks, with no significant differences between the mode of exam. This finding was explained by the fact that writing out calculations represents a common behavior in exams. A positive correlation was found between mental effort and scratch paper use. MILLER (2011) showed that this cognitive load could be significantly reduced by the aesthetic design of exams. PIAW (2012) observed higher self-efficacy as well as intrinsic and social motivation among examinees for CBA compared to PBA. However, consistent test scores were observed for the two modes of exam. Although a better perception of CBA did not improve the examinees' scores, it showed the suitability of CBA. Therefore, the author recommended that examiners follow PBA as closely as possible when creating a CBA. The reason for the recommendation is the belief that PBA serves as a gold standard for assessments that CBA must match.

In addition, eight studies focus on the examinees' performances as an objective measure for assessing the suitability of CBA. However, former research findings vary widely in this area. JEONG (2014) investigated the extent to which different PBA and CBA scores exist and whether gender or exam subjects influenced these. The results showed that higher scores were obtained in PBA. However, this was only true for half of the exam subjects considered and is not universally applicable. A gender-specific analysis showed that women received a lower score in CBA than in PBA in significantly more subjects than men. This effect was attributed to a higher familiarity with computers among male examinees. KALOGEROPOULOS ET AL. (2013) investigated examinees' PBA and CBA performance in an ICT-related course. No significant differences were observed for closed question types. However, examinees in CBA performed significantly better in application-oriented tasks (e.g., coding tasks). The use of an authentic exam environment (e.g., compiler functions for code testing) was listed as one reason for the observation. GRISSOM ET AL. (2016) also observed significantly higher scores for CBA when authentic exam environments were used. However, they found that almost half of the CBA participants still made errors in their answers. CANDRLIC ET AL. (2014) conducted a comparative study of PBA and online exam scores. Comparable scores were observed for both modes of exam, as long as particular composition decisions were considered. Thus, they stated that CBA must include an adequate question type for each PBA task, or an additional assessment activity must be implemented. CLARIANA / WALLACE (2002) also investigated possible factors for different performances in PBA and CBA and found that content familiarity has a significant influence on the observed performance differences. By contrast, the factors computer familiarity, competitiveness, and gender did not influence the results. BOEVÉ ET AL. (2015) compared the scores of PBA and CBA, with no significant differences observed between the modes of exam. Nevertheless, a higher acceptance of PBA was stated. The reasons given for this were that PBA allow students to work in a structured manner, achieve better concentration, and gain a good overview of their exam progress. By contrast, in a CBA examinees felt less in control compared to a PBA. However, it was also shown that acceptance of CBA increased with experience and familiarity. Backes / Cowan (2019) also addressed the aspect of familiarity by investigating the implementation of CBA and the different effects of online and offline formats. The long-term study showed negative effects on the scores of online exams, especially for the first use. However, exam scores increased due to the increasing familiarity. Maguire et al. (2010), in contrast, observed significantly higher average scores for CBA compared to PBA. Possible reasons were a preferred interaction of examinees with computers and a possible reduction of test anxiety.

In addition to investigating individual factors, the literature also mentions three approaches to adapt existing acceptance models to CBA. Terzis / Economides (2011) developed a model to determine acceptance in the form of behavioral intention. The Computer Based Assessment Acceptance Model (CBAAM) is based on existing models (e.g., Technology Acceptance Model (TAM), Theory of Planned Behavior (TPB), or the Unified Theory of Acceptance and Usage of Technology (UTAUT)) and their interrelationships. The core factors include perceived usefulness, perceived ease of use, content, and perceived playfulness. These factors are influenced by computer self-efficacy, social influence, facilitating conditions, and goal expectancy. Building on this model, Terzis et al. (2012) used the Big Five Inventory (Goldberg 1990) to measure the influence of personality traits, including agreeableness, conscientiousness, extroversion, neuroticism, and openness, on acceptance of CBA. It was found that higher agreeableness had a positive effect on social influence and perceived ease of use. In addition, neurotic examinees found that the CBA was less useful and had more negative expectations regarding their exam performance. Conscientiousness had a significant direct effect only on perceived ease of use. In addition, the perceived performance was included to show the direct influence of personality traits on acceptance. Here, a significant influence on perceived importance was only demonstrated for extroversion and openness. A further study that examined which factors influence the long-term acceptance of CBA (Terzis et al. 2013) found that confirmed positive ease of use and playfulness promoted continual acceptance and, thus, long-term use.

In the following, the results of the literature review on the digital execution of exams from a user-centric perspective are discussed. In this field, the research focused on factors that influence the examinees' behaviors and exam scores. The most frequently considered factors included anxiety, familiarity, stress, and cheating. The observed influence varied between the respective publications. Although the authors largely agreed that increasing familiarity with the mode of exam reduces stress and anxiety, these are also influenced by other factors. Thus, different types of stress and anxiety affect the CBA performance. There are also different research results about the exam score. Although comparable exam performance between PBA and CBA was observed in most studies, this was not the case for all studies, as some observed significantly higher scores for both modes of exam. The aspect of familiarity was also said to be important in this context. In addition, isolated efforts were found to transfer the identified factors and interrelationships into a unified model.

Scoring of the Digital Exam – Technical Perspective

From the technical perspective of scoring digital exams, five papers were identified as focusing on scoring ICT-related assignments. SUMMONS (1997) showed the potential for automatic scoring of ICT-based tasks in the example of Excel worksheets. Thus, the automatic scoring of completed Excel templates was evaluated using Visual Basic for Applications (VBA) programming. HIGGINS ET AL. (2002) and HIGGINS ET AL. (2005) investigated the automatic scoring of more complex coding and modeling tasks. The scoring involved several evaluation criteria, including typographic aspects (e.g., length of identifiers; usage of comments), content evaluation (using a test data set), and verification of task-specific features. A similar evaluation approach for graphical answers (e.g., diagrams or modeling) was also followed by PRADOS ET AL. (2011). Here, the answers were compared to sample solutions in the system, with deviations leading to point deductions. While such fully automated systems are said to have high consistency in scoring, the feedback of results can only be personalized to a limited extent. In this context, BUYRUKOGLU ET AL. (2019) pursued a semi-automatic approach to score coding tasks. Thus, individual answer components were scored manually, and the scoring comments were transferred to other semantically identical answers. The intention was to speed up the scoring process while enabling individualized and consistent feedback.

The second focus of digital exam scoring, the automatic scoring of textual answers (e.g., short answers or essays), is addressed in 10 identified papers. LAJIS / AZIZ (2010) and LAJIS / AZIZ (2012) investigated the automatic scoring of short answers using an AI-based approach. They showed that such an approach is particularly suitable for lower skill levels (e.g., remembering and understanding) according to BLOOM ET AL. (1956). KERR ET AL. (2013) investigated the use of proposition extraction for domain-independent, deep natural language processing. Here, the answers were first structured using grammatical rules and then compared with a template. The structural approach also allowed a content-based assessment of text quality. However, problems arose for texts with less common grammatical structures and texts in which the evaluated information was distributed using several sentences. SHERMIS (2014) and SHERMIS (2015) showed a higher scoring accuracy for essays compared to short answers. In these two studies, the focus was not on evaluating writing ability but on evaluating content quality. However, the considered machine scoring algorithms did not reach the same level as human raters due to the amount of information needed for the evaluation, which was lower for short answers than for essays. Moreover, the automatic systems could not evaluate the quality of argumentation and conclusion but only performed structural evaluations. Therefore, he recommended semi-automatic scoring systems. MOHLER / MIHALCEA (2009) investigated the use of unsupervised techniques for short answer scoring. The authors showed that knowledge-based and corpus-based measures of similarity performed comparably in scoring short answers. However, corpus-based measures have a greater potential due to their extensibility and domain-relatedness. One frequently studied system for scoring textual answers is the so-called e-rater, as presented by BURSTEIN ET AL. (2004). The system evaluates the quality of a text based on different features, including grammatical and linguistic aspects like missing punctuation or typographical errors. Based on these errors, a score is calculated which reflects the text quality. ATTALI / BURSTEIN (2005) modified the set of features and the model-building approach. The adapted set of features was based on human rubrics for scoring essays and allowed standardized essay scoring across different prompts. This improved the performance expressed by the agreement with the human rater. ATTALI / POWERS (2008) investigated the development of scoring measures for persuasive and descriptive modes of writing. Three classes of features, namely, fluency (essay length and style), conventions (grammar, usage, and mechanics), and word choice (vocabulary and word length) were identified as appropriate evaluation measures. The researchers observed only minor differences in the scores between the two modes of writing. Here, the scores for persuasive essays were lower than for descriptive essays, mainly due to the word choice features. QUINLAN ET AL. (2009) studied the scoring accuracy on the level of microfeatures, which were assigned to the features mentioned above. Their results showed partly low accuracies, especially on the level of the microfeatures, as the cause of some errors could not be assigned to individual microfeatures.

In the following, the results of the literature review on the digital scoring of exam tasks are discussed. In this field, the research focused on open question types since closed question types do not pose problems for scoring systems due to their explicit solutions. The focus on the digital scoring of ICT-related tasks results from the widespread use of digital exams in this subject area. Thus, it is obvious that digitally recorded exam answers are also evaluated automatically as much as possible. In addition, coding and modeling tasks have a clear structure and clearly defined characteristics. Here, existing research showed that the evaluation of the structure of answers can already be carried out with high accuracy by scoring systems. The focus of scoring textual answers differed in individual parts. Thus, textual answers can be scored according to different evaluation criteria. A large part of the research dealt with evaluating writing ability and semantic criteria. In this process, the textual answers were checked for structure, length, punctuation, and vocabulary. According to QUINLAN ET AL. (2009), two problems arise here. First, correlations between features can occur, and second, writing styles that deviate from a particular specification are sometimes incorrectly evaluated. An actual evaluation of the content does not occur with this scoring type. Systems often reach their limits when it comes to assessing content. It was shown that most publications deal with the feasibility of content-based scoring of essays, with the determination of scoring features and accuracy being a core element.

3 Research Objectives and Structure

In this section, the research objectives and research questions (RQ) are derived based on the results of the related research from Section A.2.2. In the following, the further structure of the dissertation as well as the conducted research studies are presented.

Research Objectives

First, the technical perspective on executing digital exams showed that the focus is primarily on creating authentic exam environments. Therefore, digital exams are frequently used for ICT-related subjects. The topic of cheating and cheating detection is also considered relevant in this area. As the research results in Section A.2.2.2 consistently showed, digital execution enables both new opportunities for cheating and new ways of detecting it. However, the results regarding the user-centric perspective of executing digital exams are much more contradictory. Core topics in the former research were anxiety, familiarity, and exam scores. In addition, it was shown that these factors change over time. Thus, continuous use of CBA leads to increased familiarity and reduces CBA-specific anxiety. Therefore, previous research often emphasized the extensive preparation of examinees when introducing digital exams.

However, the beginning of the COVID-19 pandemic in 2020 showed that there is not always enough time for this introduction. In this context, digital execution was implemented without much lead time and mostly in legal gray areas. RQ1 thus addresses the area of user-centric implications for introducing the execution of digital exams.

Research Field: Execution of Digital Exams – User-Centric Perspective	
RQ1	Which factors must be considered in the spontaneous introduction of the execution of digital exams?

Second, the user-centric perspective on scoring digital exams has been given little to no consideration in current research. In particular, in ICT-use trust in the systems used is considered to be of great importance for acceptance and thus use (Kocielnik et al. 2019). Therefore, whether and how such systems are ultimately used depends on the users. The use of scoring systems includes the active use by examiners and the passive use by examinees. For examiners in particular, the advantages of digital scoring seem to increase their motivation to use the system, since the overall scoring effort can be reduced, even when using semi-automatic instead of automatic systems. However, the advantages for examinees who are only confronted with the scores are limited. Therefore, RQ2 addresses the use of AI-based (semi-)automatic essay scoring (AES) systems from the examinees' perspective.

Research Field: Scoring of Digital Exams – User-Centric Perspective	
RQ2	Which factors influence the trust of examinees in AI-based (semi-)AES systems?

Third, the technical perspective on scoring digital exams showed that the research focuses on improving scoring accuracy. This concerns both the features and corpora with which texts are evaluated as well as the underlying scoring models. AI-based systems are increasingly being used in this context. Understandably, such systems are only used if they also provide added value and reliably perform the intended task. One point that has been ignored in current research is the design of the systems in which scoring takes place. Although there are efforts to implement an automatic system as the last stage of development, there is still a need for semi-automatic systems, especially because of the low accuracy and limitations of such systems. Here, it is especially important to consider the required functionalities as well as the presentation of the scoring results. RQ3 thus addresses the technical perspective of designing AI-based (semi-)AES systems.

Research Field: Scoring of Digital Exams – Technical Perspective	
RQ3	How must AI-based (semi-)AES systems be technically designed to be perceived as useful?

Structure of the Dissertation

This dissertation is divided into three parts (see Figure 5). Part A gives an introduction to the thesis. After showing the motivation and relevance of the dissertation topic (Section A.1), the theoretical foundations, including the related research concerning the digitization of high-stakes exams, are presented (Section A.2). Part A concludes with the derivation of the research questions and an overview of the thesis structure (Section A.3). Part B includes the research studies conducted as part of the dissertation. In the three-stage process of digital exams, these studies can be assigned to the steps of digital execution and scoring of exams. Studies I and II address the user-centric perspective of the digital execution of exams (RQ1). Study III looks at the user-centric perspective of the (semi-)automatic scoring of essays in exams (RQ2). Finally, Studies IV and V deal with the technical design of exams scoring systems (RQ3).

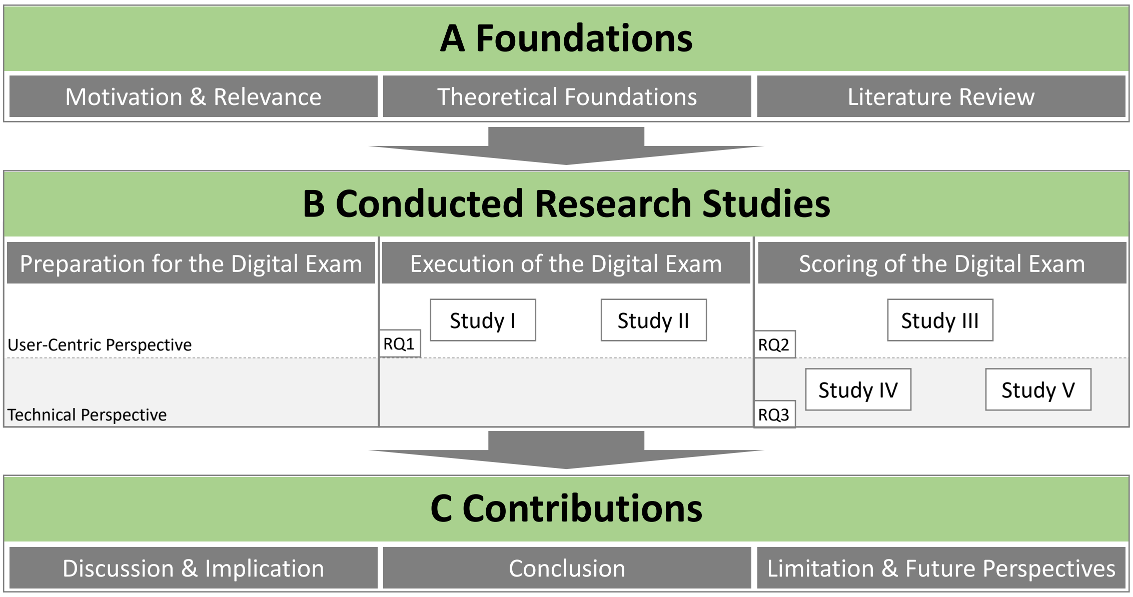


Figure 5. Structure of the Dissertation

Study I (see Section B.1) arose in 2020 from the need to conduct high-stakes exams at universities in a decentralized manner as online exams on short notice due to the emergence of the COVID-19 pandemic. The need for research resulted from the lack of experience of examiners and examinees with online exams, as German universities traditionally focused on in-class PBA. A survey study was performed to address the challenges that examinees faced due to the change in the mode of exam. The aim was to investigate the students' perceptions of the exam and exam environment. Later in the pandemic, the research field of emergency remote assessments emerged for this type of digital assessment. Study II (see Section B.2) builds on and extends the findings from Study I. During the COVID-19 pandemic, the technical and organizational foundations were laid for the long-term use of online exams even after the pandemic. The intention to participate in online exams depends on the extent to which institutions and examiners apply the lessons learned from the pandemic. To derive these learnings, a mixed-method approach was followed. First, a quantitative survey study was conducted among those who took online exams during the pandemic. In addition, in-depth semi-structured follow-up interviews were carried out with examinees. The aim was to derive recommendations for action. In Study III (see Section B.3), the trust of examinees towards AI-based AES systems was investigated. Following the trust model of MAYER ET AL. (1995), examinee characteristics, system characteristics, and environmental characteristics were addressed. The study was conducted as a scenario-based experiment identify differences in examinee trust towards the (semi-)automatic implementation of AI-based essay scoring systems. The study aimed to evaluate initial trust as crucial aspect of supporting the adoption of such systems. Based on the results of Study III, Study IV (see Section B.4) addresses the design of a semi-automatic machine-learning-based essay scoring system. While previous research often focused on scoring quality, the design of such systems has been addressed little or not at all. This study investigates the design of a functional software artifact to support essay scoring. For this purpose, a fully functional user interface artifact (including a first version of a machine learning-based scoring algorithm) was implemented. Methodologically, a Design Science Research (DSR) approach was followed (Peffers et al. 2007; Hevner et al. 2004) with the goal of identifying and documenting key design principles for semi-automatic essay scoring systems (Gregor et al. 2020). Study V (see Section B.5) investigates the presentation of scoring results of an AI-based automated essay scoring system from an examinee's perspective. While the use of Explainable AI (XAI) was extensively explored from an algorithmic perspective, there was little actual user-centered consideration given in the educational domain. The research was performed as a survey-based experiment in which visual and content design elements as well as their interactions were evaluated. The goal was to determine how AES systems must be designed, independent of the algorithms, to ensure the local interpretability of individual assessment results. Table 5 gives an overview of the conducted research studies presented in Part B as well as the adaptations made compared to the published version. Furthermore, corresponding appendixes to the respective studies are listed, which were added within the scope of this work.

Study	Adaption
Study I: "The Intention to Participate in Online Exams - The Student Perspective" *(Hartmann et al. 2021) - AMCIS 2021 - published*	
Adaption	- Layout and numbering (tables and figures) to the dissertation style
Appendix	- Appendix A: Questionnaire Studies I and II
Study II: "From Emergency Remote Assessment to a New Status Quo? – Lessons Learned From Online Assessments During the COVID-19 Pandemic" *(Hartmann/Hobert 2023 b) - ECIS 2023 - published*	
Adaption	- In-study references (numbering of sections) to the dissertation structure - Layout and numbering (tables and figures) to the dissertation style - Placement of tables and figures to avoid page breaks
Appendix	- Appendix A: Questionnaire Study I and II - Appendix B: Guideline of the Follow-Up Interviews
Study III: "Trust, but Verify! - An Empirical Investigation of Students' Initial Trust in AI-Based Essay Scoring" *(Hartmann et al. 2022 b) - AMCIS 2022 - published*	
Adaption	- Layout and numbering (tables and figures) to the dissertation style
Appendix	- Appendix C: Questionnaire Study III
Study IV: "(AI)n't Nobody Helping Me? - Design And Evaluation of a Machine-Learning-Based Semi-Automatic Essay Scoring System" *(Hartmann et al. 2022 a) – ECIS 2022 - published*	
Adaption	- In-study references (numbering of sections) to the dissertation structure - Layout and numbering (tables and figures) to the dissertation style
Appendix	---
Study V: "Explain AI-Based Essay Scorings without XAI – Empirical Investigation of an User-Centered UI Design for AI-Based AES Systems" *(Hartmann/Hobert 2023 a) - AMCIS 2023 - published*	
Adaption	- Layout and numbering (tables and figures) to the dissertation style - Placement of tables and figures to avoid page breaks
Appendix	- Appendix D: Questionnaire Study V

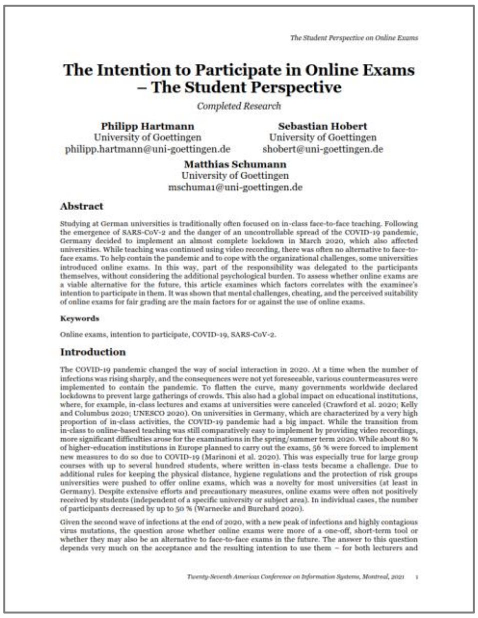
Table 5. Overview of the Conducted Research Studies

Part C discusses the results of the conducted research studies and their relevance to the current state of research. For this purpose, generalizable recommendations regarding the digitization of high-stakes exams are given (see Section C.1). Subsequently, conclusions are drawn based on the research questions posed in Section A.3 (see Section C.2). The thesis closes with the limitations of the thesis and an outlook for future research (see Section C.3).

B Conducted Research Studies

1 Execution of Emergency Remote Exams

The Intention to Participate in Online Exams – The Student Perspective



The Student Perspective on Online Exams

The Intention to Participate in Online Exams – The Student Perspective

Completed Research

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

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

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

Abstract

Studying at German universities is traditionally often focused on in-class face-to-face teaching. Following the emergence of SARS-CoV-2 and the danger of an uncontrollable spread of the COVID-19 pandemic, Germany decided to implement an almost complete lockdown in March 2020, which also affected universities. While teaching was continued using video recording, there was often no alternative to face-to-face exams. To help contain the pandemic and to cope with the organizational challenges, some universities introduced online exams. In this way, part of the responsibility was delegated to the participants themselves, without considering the additional psychological burden. To assess whether online exams are a viable alternative for the future, this article examines which factors correlates with the examinee's intention to participate in them. It was shown that mental challenges, cheating, and the perceived suitability of online exams for fair grading are the main factors for or against the use of online exams.

Keywords

Online exams, intention to participate, COVID-19, SARS-CoV-2.

Introduction

The COVID-19 pandemic changed the way of social interaction in 2020. At a time when the number of infections was rising sharply, and the consequences were not yet foreseeable, various countermeasures were implemented to contain the pandemic. To flatten the curve, many governments worldwide declared lockdowns to prevent large gatherings of crowds. This also had a global impact on educational institutions, where, for example, in-class lectures and exams at universities were canceled (Crawford et al. 2020; Kelly and Columbus 2020; UNESCO 2020). On universities in Germany, which are characterized by a very high proportion of in-class activities, the COVID-19 pandemic had a big impact. While the transition from in-class to online-based teaching was still comparatively easy to implement by providing video recordings, more significant difficulties arose for the examinations in the spring/summer term 2020. While about 80 % of higher-education institutions in Europe planned to carry out the exams, 56 % were forced to implement new measures to do so due to COVID-19 (Marinoni et al. 2020). This was especially true for large group courses with up to several hundred students, where written in-class tests became a challenge. Due to additional rules for keeping the physical distance, hygiene regulations and the protection of risk groups universities were pushed to offer online exams, which was a novelty for most universities (at least in Germany). Despite extensive efforts and precautionary measures, online exams were often not positively received by students (independent of a specific university or subject area). In individual cases, the number of participants decreased by up to 50 % (Warnecke and Burchard 2020).

Given the second wave of infections at the end of 2020, with a new peak of infections and highly contagious virus mutations, the question arose whether online exams were more of a one-off, short-term tool or whether they may also be an alternative to face-to-face exams in the future. The answer to this question depends very much on the acceptance and the resulting intention to use them – for both lecturers and

Twenty-Seventh Americas Conference on Information Systems, Montreal, 2021 1

Abstract: *Studying at German universities is traditionally often focused on in-class face-to-face teaching. Following the emergence of SARS-CoV-2 and the danger of an uncontrollable spread of the COVID-19 pandemic, Germany decided to implement an almost complete lockdown in March 2020, which also affected universities. While teaching was continued using video recording, there was often no alternative to face-to-face exams. To help contain the pandemic and to cope with the organizational challenges, some universities introduced online exams. In this way, part of the responsibility was delegated to the participants themselves, without considering the additional psychological burden. To assess whether online exams are a viable alternative for the future, this article examines which factors correlate with the examinee's intention to participate in them. It was shown that mental challenges, cheating, and the perceived suitability of online exams for fair grading are the main factors for or against the use of online exams.*

Keywords: *Online exams, intention to participate, COVID-19, SARS-CoV-2*

Citation: (Hartmann et al. 2021) Hartmann, P.; Hobert, S.; Schumann, M.: *The Intention to Participate in Online Exams – The Student Perspective*. In: *Proceedings of the 27th Americas Conference on Information Systems*. Montreal, Canada. 2021. pp. 1-10.

1.1 Introduction

The COVID-19 pandemic changed the way of social interaction in 2020. At a time when the number of infections was rising sharply, and the consequences were not yet foreseeable, various countermeasures were implemented to contain the pandemic. To flatten the curve, many governments worldwide declared lockdowns to prevent large gatherings of crowds. This also had a global impact on educational institutions, where, for example, in-class lectures and exams at universities were canceled (Crawford et al. 2020; Kelly / Columbus 2020; UNESCO 2020). On universities in Germany, which are characterized by a very high proportion of in-class activities, the COVID-19 pandemic had a big impact. While the transition from in-class to online-based teaching was still comparatively easy to implement by providing video recordings, more significant difficulties arose for the examinations in the spring/summer term 2020. While about 80 % of higher-education institutions in Europe planned to carry out the exams, 56 % were forced to implement new measures to do so due to COVID-19 (Marinoni et al. 2020). This was especially true for large group courses with up to several hundred students, where written in-class tests became a challenge. Due to additional rules for keeping the physical distance, hygiene regulations and the protection of risk groups pushed universities to offer online exams, which was a novelty for most universities (at least in Germany). Despite extensive efforts and precautionary measures, online exams were often not positively received by students (independent of a specific university or subject area). In individual cases, the number of participants decreased by up to 50 % (Warnecke / Burchard 2020).

Given the second wave of infections at the end of 2020, with a new peak of infections and highly contagious virus mutations, the question arose whether online exams were more of a one-off, short-term tool or whether they may also be an alternative to face-to-face exams in the future. The answer to this question depends very much on the acceptance and the resulting intention to use them – for both lecturers and students. Possible challenges for online exams include the risk of technical problems, cheating, assessment and certification of practical knowledge and skills, and equal exam conditions (OECD 2020). It is the goal of this study to determine the status quo from the examinees' perspective regarding these challenges.

Thus, the principles of fairness and objectivity of the exam must be guaranteed for online exams. Another aspect is the execution of the online exam. While the requirements for participation in a face-to-face exam are mainly that the students are in the right place at the right time, online exams require a higher degree of self-organization. Technical problems, such as computer problems or noise pollution during the exam, are hence no longer the responsibility of the university but of each individual examinee. Therefore, we expect a certain affinity for technology to be required, especially when dealing with online exams. The overall goal of the conducted survey study is to determine the relevant factors that correlate with the students' intention to participate in online exams.

1.2 Related Research and Hypotheses Development

In the following, expected relevant factors for the adoption of online exams from a student perspective are discussed and hypotheses regarding their correlation with the intention to participate in online exams are derived. We selected the factors technology affinity, mental challenges, cheating, and suitability of online exams for fair grading since these challenges are named in literature (OECD 2020) and could be confirmed in informal conversations with students.

Technology Affinity

The use of technology affects many components of an online exam. These can range from handling the exam system before and during the exam to solving technical problems. Therefore, the importance of technology affinity for online exams is expected to be particularly important from many perspectives. For example, technical challenges in time-limited exams can further shorten the processing time (Gamage et al. 2020). Especially, the ability of touch typing plays an important role in the processing of the exam tasks (Thomas et al. 2002). Previous research has shown that the use of computers in essay exams depends on the perceived speed and accuracy of typing (Mogey / Fluck 2015). In the present scenario, we expect that students with a higher affinity for technology will have fewer problems due to the change in exam type and will have a higher intention to participate in online exams.

H1: *Students with a higher technology affinity show a higher intention to participate in online exams.*

Mental Challenges

Exam anxiety is an important component of research in the field of education. In particular, a high level of additional exam anxiety has already been demonstrated during the Corona pandemic (Alsaady et al. 2020). The factors that can influence exam anxiety can be very multifaceted. Since we want to concentrate on the differences between face-to-face and online exams, we will focus on the online specific mental challenges, namely the concentration in the new exam environment and additional stress due to the transformed processes of the exams.

Among the mental challenges, the aspects of stress and concentration in online exams, which are considered in the following, can influence the participants in addition to the actual exam performance. For example, online exams represent a deviation from the usual face-to-face exam processes. Students need time to become familiar with the new situation and to get used to the procedure and the execution at home. It has been shown that a missing familiarization is expected to slow down the task response so that the examinees find themselves under additional time pressure (Thomas et al. 2002). In addition, the students' own responsibility for the device on which they take the exam can lead to stress. For example, students fear that they will be at a significant disadvantage compared to other participants in the event of technical problems (Küppers / Schroeder 2018). In addition, the decentralized nature of the exam means that only limited support can be provided when problems arise, which can further reinforce this effect (Gamage et al. 2020). However, not all individuals are equally affected by stress. Whereas some students prefer online exams to traditional face-to-face exams and thus have lower exam anxiety (Stowell / Bennett 2010), there are other students for whom online exams increase exam anxiety. In the latter case, it is assumed that the additional responsibility, additional procedures for registration and execution, and possibly associated technical problems cause this effect (Stowell / Bennett 2010). Another important factor is the availability of a quiet exam environment. Not every student has access to an undisturbed exam environment at home, which influences the concentration during the exam (OECD 2020). Therefore, one's own home as an exam environment is sometimes considered inappropriate (Elsalem et al. 2020). In the present scenario, we expect that students with a lower perceived mental challenge also have a higher intention to participate in online exams.

H2: *Students with a lower mental challenge at online exams also show a higher intention to participate in online exams.*

Cheating

The issue of cheating in exams is a major challenge for all universities (McCabe 2005), and while supervision in centralized face-to-face exams is already difficult with the use of modern technology, decentralized exams pose a much greater challenge. There have been extensive studies on digital in-class exams showing that the use of one's own devices (e.g., bring your own device) already causes a higher expected risk of cheating among examinees (Küppers / Schroeder 2018). If these online exams are also written decentrally as take-home exams, the potential for cheating increases, especially in exams without proctoring procedures (Harmon / Lambrinos 2008). For example, the unsupervised execution of online exams allows the use of unauthorized aids and sources as well as communication with third parties (Gamage et al. 2020). The reasons for cheating are manifold. Often the pressure to perform in the competition with other graduates plays an important role (Gamage et al. 2020). Cheating may also influence the behavior of the other students as students may feel compelled to cheat in order not to be at a disadvantage (Owunwanne et al. 2010). Especially in Germany, where there are narrowed legal limits to surveillance and data collection at private computers in general, online proctoring is no option for many universities (Dieckmann 2021). In our scenario, we, therefore, assume that an equal or lower expected level of cheating in online exams compared to face-to-face exams occurs with a higher intention to participate in online exams.

H3: *Students expecting an equal or lower level of fellow students to cheat in online exams show a higher intention to participate in future online exams.*

Perceived Suitability of Online Exams for Fair Grading

As already mentioned, the fair assessment and certification of knowledge and skills can become a challenge in online exams. Grading systems can be divided into standard-based systems, where the goal of grading is to establish a certain level of knowledge and skills, and the norm-referenced system, where an individual's performance is considered in relation to other individuals (Tierney et al. 2011). While all students have nearly similar conditions at in-class face-to-face exams, the exam environment in online exams may differ, for example, with regard to the participants' technical equipment (OECD 2020). Studies show that students without experience in computer-based exams believe that these favor individual participants, although these prejudices lose their significance with increasing experience (Hillier 2014). Consequently, it is particularly important to consider the perceived suitability of online exams for fair grading for the fair assessment of knowledge and skills. In the present scenario, we therefore expect that a higher perceived suitability of online exams for fair grading will occur with a higher intention to use them.

H4: *Students expecting the online exam to be suitable to assess their knowledge and skills show a higher intention to participate in online exams.*

1.3 Research Design

To answer the research questions and to analyze the hypotheses, a survey among students at a large university was conducted. In the following, we first describe the case of the introduction of the online exams in comparison to the initial situation before the COVID-19 pandemic. We then give an overview of the questionnaire used and the participants.

The Case of a German University: Face-to-Face vs. Online Exams

In this study, we focus on a large German University whose teaching is strongly oriented towards in-class teaching in order to promote the interaction and knowledge exchange among students. Before the start of the SARS-CoV-2 disease, examinations were most often offered in the form of face-to-face exams, in which the participants had to work on the assignments in a lecture hall. Seminar papers, term papers, and practical exams (e.g., in labs) were an exception to this. In the exams, students were given printed exam tasks at the beginning of the processing time. The used task types typically ranged from open to closed questions (e.g., multiple-choice) and included mathematical tasks. The students had a previously defined processing time to work on the assignment under the supervision of university staff. At the end of this processing time, the exam tasks and the solutions worked out by the individual examinee had to be handed in to the exam supervisor. The permitted aids and processing time (usually between 60 and 120 minutes) may vary for each exam. An identity check is also carried out during each exam. Therefore, the participants must identify themselves with a photo ID and are then checked off on the list of registered examinees.

When transferring the face-to-face exam to an online exam, there were no central specifications regarding the design of the exam form. However, two forms were primarily used for online exams. First, there was the possibility of downloading the assignment as a text file (e.g., MS Word). In this case, the students were enabled to download the file at a certain point in time. The downloaded assignment was then processed on the student's computers, saved as a PDF file, and uploaded at the end of the processing time. Second, there was an electronic exam system based on the learning management system ILIAS (ILIAS 2021). This had already been used before in a similar form for face-to-face e-exams at the given university. The exam was activated in the system at a given time and could be started. The exam time was automatically recorded by the system, and the answers were automatically saved at the end of the official processing time at the latest. A further upload was not necessary at this point.

In addition to the exam process, a video identification procedure was available for identity verification (i.e., a replacement of the former photo ID checks in face-to-face exams). Even though many lecturers used the identification system, it was not mandatory in all university departments. During the video identification procedure, the examinees had to log in via a link before the start of the exam and had to take a photo using a webcam or smartphone, on which both their face and a photo ID could be recognized. Due to legal considerations (e.g., privacy issues), there was no complete supervision during the processing time, for instance, in the form of keylogging or video monitoring.

Questionnaire

To test the relationships postulated in the hypotheses, a four-part questionnaire was designed. The four parts covered the topics of demographics, exam procedure, general aspects about online exams as well as standardized questionnaires (e.g., about technology affinity and user experience). To measure the hypothesis-related items for the intention to participate, the mental challenges, the suitability of online exams for fair grading, and cheating, the participants were asked about their level of agreement with pre-formulated statements. The agreement with the respective statements was measured on the basis of a 7-point Likert scale. A detailed list of the considered items as well as the assignment to the factors can be found in Table 7 in the results section.

To determine the intention to participate, it was postulated that the participants were happy that the exam was conducted as an online exam and that they would like to continue having online exams after the Corona pandemic. To examine the expected relationship between technology affinity and intention to participate formulated in hypothesis 1, the technology affinity was examined using the ATI scale (Franke et al. 2019). The ATI scale consists of a standardized 9 items covering questionnaire that tendency to engage or avoid the technology interaction. The rating is based on a scale from 1 (avoid) to 6 (engage). To investigate the relationship between mental challenges and the intention to participate from hypothesis 2, two statements were provided to evaluate the mental challenges. In the first statement, participants had to indicate to what extent they agree that online exams cause the same or less stress compared to face-to-face exams. In the second statement, we stated that the concentration in online exams is at least as great as in face-to-face exams. The statements for the determination of the factor cheating indicated that students do not expect a higher cheating rate in online exams and that with a higher rate of cheating the reasons for this are neither the lack of supervision nor a disadvantage compared to the other students. To determine the suitability of online exams for fair grading, two statements were made about the fairness of the assessment of online exams. First, students were asked for their agreement with the statement that online exams reflect a fair assessment of their knowledge skills. Following this, students were asked to agree with the statement that an online exam provides a fair assessment of one’s knowledge and skills compared to fellow students.

The questionnaire was piloted by students before being sent out to check its comprehensibility among the target group. Afterward, the online survey was forwarded by the lecturers to the examiners after the exam and was conducted completely anonymously. Participation was voluntary and no participation incentives were offered.

Participants

All lecturers who had offered an online exam in the spring/summer term 2020 were contacted via e-mail by the first author with the request to forward the questionnaire to the participants of the exam. To what extent the lecturers complied with this request cannot be determined. A weighting of the respondents according to gender, age, degree program, or other criteria was not carried out due to data protection regulations and the anonymous participation. Overall, 171 students participated in the survey, of which 66.09 % were female, 28.65 % male, and less than 1.00 % diverse. The age varied from 18 to 54 years with an average of 22.4 years (SD = 4.7). 92.41 % of the participants studied at the faculties of law, business and economics, social science, and philosophy.

Faculty	Participants	Male	Female	Others	No Answer
Law	38.60 %	10.53 %	26.32 %	0.58 %	1.17 %
Business & Economics	11.70 %	5.26 %	5.85 %	0.00 %	0.58 %
Social Science	11.11 %	2.92 %	8.19 %	0.00 %	0.00 %
Philosophy	31.00 %	5.85 %	22.22 %	0.00 %	2.92 %
Others	7.60 %	4.09 %	3.51 %	0.00 %	0.00 %
Total	100 % (n = 171)	28.65 %	66.09 %	0.58 %	4.67 %

Table 6. Summary of Participants

1.4 Results

Descriptive Analysis

57.86 % of the participants stated that their exams were conducted by downloading the exam task and uploading a document with the answers at the end of the exam, while 35.71 % took their exam in the electronic examination system. The remaining 6.43 % named a different online exam type. The students were also asked to what extent they agreed with the statement that they preferred the online exam to the face-to-face exam. Here the respondents agreed with a mean of 4.16 (SD = 2.38). However, the desire to have online exams even after the pandemic was not clearly expressed (MD = 3.35; SD = 2.42).

Even though the university had no previous experience in conducting online exams, the students reported no major technical problems. However, some of them remarked that there was sometimes a high load on the examination system, particularly at the beginning or end of the exam. On average, 73.60 % of the participants stated that they had no or only slight problems (represented by a Likert scale value of 1 or 2) in general. In contrast, only 5.38 % mentioned more significant problems (represented by a Likert scale value of 6 or 7) during the exam.

Regarding the hypothesis-related items, the results showed a medium average technology affinity of the students (MD = 3.31; SD = 1.14), according to the 6-point ATI scale. For the mental challenges, it can be shown that with an average score of 3.68 (SD = 2.45), the students perceived a comparable (slightly lower) level of concentration in online exams as in face-to-face exams. In addition, the students stated that they felt a comparable (slightly lower) level of stress during online exams compared to face-to-face exams (MD = 4.50; SD = 2.37). Looking at the items on cheating, the examinees disagreed with the statement that fewer cases of cheating occur in online exams (MD = 2.44; SD = 1.92). The primary cause of cheating is the lack of supervision (MD = 3.29; SD = 2.19), whereas cheating as a reaction to the behavior of fellow students plays a subordinate role (MD = 4.09; SD = 2.32). For the evaluation of the suitability of online exams for fair grading, the students declared online exams a fair means of assessing individual knowledge and skills by itself (MD = 4.31; SD = 1.91) and in comparison to fellow students (MD = 4.37; SD = 1.81).

Statistical Analysis

Within the data analysis, first, a factor analysis was conducted for dimensional reduction. Since the items were collected using a Likert scale and thus have an ordinal scale, the relationship between the individual factors and the intention to participate was determined by a correlation analysis using the Kendall rank correlation coefficient.

During the factor analysis, the items used were able to identify the factors of mental challenges (Cronbach's Alpha = 0.865), cheating (Cronbach's Alpha = 0.860), the suitability of online exams for fair grading (Cronbach's Alpha = 0.818), and the intention to participate (Cronbach's Alpha = 0.773) and were evaluated as reliable. The sample has a KMO-value of 0.785 and can be considered somewhere between middling (Kaiser 1974) and good (Hair et al. 2010). Since the technology affinity was determined with the help of the ATI-score (Franke et al. 2019), these items were excluded from the factor analysis. Table 7 shows the results of the conducted factor analysis as well as the included items.

Factor	Cronbach's Alpha	Items
Intention to Participate	0.773	I wanted to write the exam as an online exam rather than a face-to-face exam.
		After the end of the COVID-19 pandemic, online exams should not be replaced again by face-to-face exams.
Technology Affinity	ATI-Scale based on 9 items (Franke et al. 2019)	
Mental Challenges	0.865	Taking exams online causes less stress than taking them at the university.
		I find it easier to concentrate in online exams than in face-to-face exams.
Cheating	0.860	The percentage of cheating examinees is lower for online exams than for face-to-face exams.
		Students do not cheat more on online exams due to a lack of supervision.
		Students do not cheat more on online exams due to the risk of disadvantage in online exams.
Perceived Suitability of Online Exams for Fair Grading	0.818	The grading of the online exam will reflect my actual knowledge and skills on the topic.
		The grading of the online exam will reflect my actual knowledge and skills in comparison to my fellow students.

Table 7. Results of the Factor Analysis

After the identification of the factors and the calculation of the factor values, the Kendall rank correlation coefficient between the potentially correlated factors and the intention to participate was determined. Since we have determined the factor values as an equally-weighted average over the respective items, the number of values for every factor may vary. Therefore, the correlation coefficient Kendall tau-b is used.

Correlation Factors	Kendall-Tau-b Value	Significance
Intention to Participate & Technology Affinity	0.108	0.061
Intention to Participate & Mental Challenges	0.606	< 0.001
Intention to Participate & Cheating	0.301	< 0.001
Intention to Participate & Perceived Suitability of Online Exams for Fair Grading	0.253	< 0.001

Table 8. Results of the Correlation Analysis

For the interpretation of the correlation coefficients, the work of Cohen (1992) was used. Therefore, the pairwise correlation between the intention to participate and the suitability of online exams for fair grading (Tau-b = 0.253) showed a small to medium correlation, while the correlation between the intention to participate and cheating (Tau-b = 0.301) can be described as medium. For the correlation between the intention to participate and mental challenges (Tau-b = 0.606) a significant, large correlation ($p < 0.001$) could be proven. Only for the correlation between the intention to participate and the technology affinity no significant correlation ($p = 0.061$) could be shown.

1.5 Discussion and Implication

In this section, the results of the survey are discussed with regard to the hypotheses. Further, possible implications are outlined. In the following, we particularly focus on the considered factors technology affinity, mental challenges, the suitability of online exams for fair grading, and cheating.

Technology Affinity

Hypothesis 1 hypothesized a positive correlation between the technology affinity, as measured by the ATI-Score, and the intention to participate.

Surprisingly, no significant correlation between the technology affinity and the intention to participate in online exams could be detected. On average, the participants' ATI score was 3.30 on a scale of 1 to 6, which is relatively close to the middle. On the one hand, this may indicate that the technological handling or the introduction to the process was sufficiently communicated prior to the exam so that the individual's affinity for technology does not influence the student's exam performance and hereby the willingness to take online exams. On the other hand, the absence of technical problems can also influence the results. It can be assumed that, in particular, students with less technical affinity could reject an online exam in the future if there are more extensive technical problems, which cannot be solved alone.

Mental Challenges

In hypothesis 2, we postulated that students with less mental challenges in online exams also have an increased intention to participate in them.

The results show a significant positive correlation between the mental challenges and the intention to participate in online exams. The mental challenge was described using the items concentration and stress during the exam. All in all, the students stated that online exams cause less stress compared to face-to-face exams. As described above, the process of the exam changes compared to the classic face-to-face exam. Additional steps are required before, during, and after the exam (e.g., the preparation of the exam environment and the system login). In particular, one’s own responsibility for an undisturbed exam environment can be an important point in explaining a higher level of stress for individual examinees. In a face-to-face exam, university employees are responsible for the smooth running of the exam, whereas students, more or less, only must be present and complete the exam tasks. They do not have to worry about the organization of the rooms, the functionality of the equipment used, or about handing in the exams at the end. In the case of online exams, however, the decision and responsibility of the exam environment is moved to the participants. Possible noise pollution, problems with the technical equipment or mistakes when saving and handing in exam answers are mostly in the private sphere of each student. Therefore, in addition to the normal stress before or during an exam, further organizational stress may arise so that the perceived stress level rises at an individual level. Especially in combination with a restricted exam time, problems during online exams can reduce the available time to solve the exam task (Warnecke / Burchard 2020) and can thus be an important stressor. Research has shown that this stress is mainly due to a lack of experience and decreases with increasing practice and routine (Elsalem et al. 2020). As we have shown for the technology affinity, it seems that the handling of the system is not a relevant stressor. Furthermore, students had bigger problems of concentration in online exams compared to face-to-face exams. A possible explanation for this could be external factors beyond the student’s sphere of influence that disturb the exam environment (e.g., the loud music of a neighbor). In addition, there might be distractions that come from a private atmosphere of one’s own living environment. Taking a written exam at the university is a clear separation from the private home, both spatially and mentally. Thus, the furnishings of a private living space are designed to be comfortable and often pictures or other objects are placed with which positive memories are associated. All these factors can lead to distraction during the exam.

Cheating

In the third hypothesis, we hypothesized that an increase in the intention to participate in online exams occurs with a lower or equal expected level of cheating in them.

For hypothesis 3, a significant positive correlation to the intention to participate could be demonstrated. The factor of cheating describes the participant’s personal expectation of fellow students cheating during an online exam. It is not surprising that students who do not rate cheating during an online exam as significantly higher than during a face-to-face exam have a higher intention to participate in online exams. Overall, the students indicated that they expect a higher level of cheating in online exams. When looking at the reasons given for the higher rate of cheating, it becomes apparent that, in particular, the lack of supervision is seen as the main reason for extensive cheating. Cheating in order not to have a disadvantage against other cheating students was classified as less relevant. As already mentioned above, video surveillance in Germany, both in public places and in private homes, is perceived as a serious encroachment on personal rights. Therefore, compulsory and continuous surveillance by video or keylogging is not a solution to the problem. Hereby, different strategies are implemented at universities in Germany (Warnecke / Burchard 2020). One measure are online exams with continuous video surveillance, whereby students also have the opportunity to write the exam at the same point in time in presence at the university. Therefore, video surveillance is only voluntary. Furthermore, the conception of the exams can make cheating more difficult. Reduced exam times, the use of randomized exams (e.g., pools of questions and mixed order) and an increase in transfer tasks can help to reduce cheating due to a higher coordination effort (Cluskey Jr et al. 2011). Since the implementation of the online exams took place under time pressure, the recommended countermeasures to avoid cheating could not be implemented in all exams on time. In addition, students expect that examinees who have cheated for the first time in online exams will not repeat this behavior in future face-to-face exams (MD = 6.42; SD = 1.00). This contradicts previous research that found a higher likelihood of cheating again among students who have already cheated once (Owunwanne et al. 2010).

Perceived Suitability of Online Exams for Fair Grading

In hypothesis 4, we hypothesized that students who evaluate online exams as suitable as face-to-face exams to assess the knowledge and skills also show a higher intention to participate in online exams.

The results show a significant positive correlation between the perceived suitability of online exams for fair grading to assess the knowledge and skills on the intention to participate. Overall, the students indicated that online exams are conditionally suitable to assess their own knowledge and skills fairly. An almost identical result can be observed for the fair evaluation of one’s own performance in comparison to fellow students. As we have shown before, the decentralized organization of the exams means that there are different exam environments depending on the number of participants. Both the technical equipment of the individual participants and the external environment can have an impact on the individual performance and thus influence the exam result. Looking at the degree of difficulty of the online exam, the students did not find it more difficult than a typical face-to-face exam.

1.6 Limitation

As with any similar quantitative questionnaire studies, we are aware that this work is also subject to various limitations. First of all, it must be taken into account that the study was conducted under the influence of the COVID-19 pandemic. The resulting digital transformation of teaching and exams took place under increased time pressure and had special requirements for rapid implementation. This is reflected above all in the implementation of the online exam system and the training of students and university staff in the use of the system. Second, our model to explain the intention to use online exams is based on the factors mental challenges, cheating, the suitability of online exams for fair grading, and technology affinity, which had been identified within the literature. By using factor analysis, we combined multiple items we used in our survey. Since the research project was carried out for an initial determination of the status quo, we cannot assure that our results are complete. Looking at the factor for mental challenges, there are multiple further psychological effects that could have been used to define this factor (Reddy et al. 2018). Additionally, no other external factors were considered that influence stress levels and concentration independently of online exams. Personal feelings, e.g., a general lack of information about the course or personal situations during the COVID-19 pandemic, can be other stressors that were not asked for in the questionnaire. Furthermore, our survey does not consider the influence of digital teaching during the semester or the content of the online exams. If the provision of the teaching is not exclusively done with video recordings but also includes self-study modules on learning platforms, exam-related exercises can already be provided to train the examinee in the use of the exam system. In addition, different approaches to the design of teaching could be observed during the semester. While individual professors provided a 90-minute lecture as a 90-minute video recording, some professors changed their didactic concept for the digital semester. In some cases, a shortened video of the lecture was provided, and the proportion of self-study with textbooks was increased. Third, the results of a survey study are dependent on the selection of participants. In our study, only students who took online exams were addressed. However, a distinction must be made between mandatory and voluntary courses. Mandatory courses must be passed by all students of a study program up to a certain point in time. This can result in a compulsory attendance especially for students in higher semesters. The participants therefore also include students who would have preferred to write a classic face-to-face exam. In the case of voluntary courses, students can usually choose from many modules. Since there were also courses that offered face-to-face exams, students had the possibility to avoid online exams. Students who, for example, have been exposed to too many mental challenges, cheating, or who expect an unfair assessment could thus be underrepresented. Fourth, as mentioned before, the survey was only conducted at one university in Germany. Although the teaching at German universities is quite similar, there may occur regional differences. The same applies to the international comparison with other universities. The results of this study can therefore only represent a starting point for further research and still need to be verified regarding their generalizability.

1.7 Conclusion

In the past, online exams at universities in Germany were mostly an exception. During the first wave of the COVID-19 pandemic in the spring and summer of 2020, online exams were used to implement contact restrictions and thus to protect the students and employees of the university. What some lecturers regard a one-time or temporary tool is seen as a further step in the digitization of university education by others. However, as practice shows, online exams are not automatically considered desirable by all students. Here, it could be shown that many students in times of COVID-19 are grateful for the alternative form of exams and have used them extensively. In the long run, and especially for the time after the pandemic, students prefer the return to traditional face-to-face exams according to our survey. As relevant factors for the participation in online exams, the mental challenges could be identified, whereas a higher mental challenge is associated with a lower intention to participate in online exams. Overall, students find online exams less stressful and report a lack of concentration at home compared to the traditional face-to-face exams. A possible reason for individual stress can be the transfer of responsibilities for the execution of the exam, for which the students themselves are responsible. In addition, the students are exposed to a higher level of distraction at home. The extensive possibility of cheating during exams was another important point, which had been identified. Students, who expect a higher level of cheating in online exams, show a lower intention to participate in online exams. The main factor for cheating in online exams is the limited supervision. Especially in Germany, where there are many legal and societal concerns regarding data protection, full monitoring during an online exam is not an alternative to prevent individual students from cheating. Therefore, lecturers should at least try to make misconduct more difficult when designing the exam. This can be achieved, for example, by using a question pool and reduced processing time. The third identified factor is the expected suitability of online exams for fair grading. Our research shows a positive correlation between this factor and the intention to participate in online exams. A possible reason for this can be the acceptance of using a specific exam type as long as the grading is perceived as fair. No significant correlation could be shown between the technology affinity and the intention to participate. One reason could be the ease of the execution of the online exams, which allows students, independent of their technology affinity, to participate in online exams without disadvantages.

2 Lessons Learned from Emergency Remote Exams

FROM EMERGENCY REMOTE ASSESSMENT TO A NEW STATUS QUO? – LESSONS LEARNED FROM ONLINE ASSESSMENTS DURING THE COVID-19 PANDEMIC

FROM EMERGENCY REMOTE ASSESSMENT TO A NEW STATUS QUO? – LESSONS LEARNED FROM ONLINE ASSESSMENTS DURING THE COVID-19 PANDEMIC

Research Paper

Philipp Hartmann, University of Goettingen, Germany, philipp.hartmann@uni-goettingen.de
Sebastian Hobert, University of Goettingen, Germany, shobert@uni-goettingen.de

Abstract

Before the COVID-19 pandemic, studying at German universities was traditionally dominated by in-class activities. During the pandemic, technical and organizational foundations were laid not only for teaching but also for exams. With the resumption of in-class events, the question is to what extent online exams will also be used in the future. Here, examiners and educational institutions can learn from the experience during the pandemic to increase the students' intention to participate in online exams in the future. To identify the problems encountered and derive lessons from them, a mixed-method approach was followed. For this purpose, 478 examinees were surveyed over 3 semesters using a quantitative questionnaire, followed by in-depth semi-structured follow-up interviews with 11 students. As a result, 5 recommendations for action could be derived, covering the topics of exam routine and familiarity, problem-solving, cheating, and teaching-examination arrangements.

Keywords: Emergency Remote Assessment, Online Exams, Assessment, Education.

1 Introduction

Digital education has been gaining popularity for the last decades and has once again been massively reinforced by the COVID-19 pandemic (Crawford et al., 2020; Kelly and Columbus, 2020). Two major advantages are attributed to digital education. First, it can be used regardless of location and time, increasing flexibility and, thus, attractiveness for participants. Second, educational institutions can save costs through the reuse and reduction of space and staff capacities (Yusuf and Al-Banawi, 2013). While e-learning courses have become widespread as a result, the associated exams have nevertheless continued to be carried out in-class although the advantages mentioned above may also apply to these.

Due to the forced suspension of in-class courses during the COVID-19 pandemic, institutions around the world had to switch to online exams, so-called emergency remote assessments (UNESCO, 2020). In particular, countries like Germany, where schools and universities are dominated by in-class activities, have been strongly affected by this (Marinoni et al., 2020). Since online exams were a novelty for most universities (at least in Germany), the technical and organizational foundations had to be laid. Despite extensive efforts, the number of participants in the first semester dropped by up to 50% during the pandemic (Warnecke and Burchard, 2020). This shows that the examinees' acceptance of online exams is of great importance. Indeed, different challenges influence the actual use of online exams. These challenges include, for example, the risk of problems, cheating, the fairness of grading, and equal exam conditions (Al-Maqbali and Raja Hussain, 2022; Hartmann et al., 2021; OECD, 2020; Tuah and Naing, 2021). While studies conducted at the onset of the pandemic reflect the short-term effects of the unfamiliar situation, some influences lose or gain importance with growing familiarization over time (Zhang et al., 2021). In addition, proctoring tools used at short notice in online examinations were examined with regard to their conformity with the General Data Protection Regulation (GDPR) and

Thirty-first European Conference on Information Systems (ECIS 2023), Kristiansand, Norway *1*

Abstract: *Before the COVID-19 pandemic, studying at German universities was traditionally dominated by in-class activities. During the pandemic, technical and organizational foundations were laid not only for teaching but also for exams. With the resumption of in-class events, the question is to what extent online exams will also be used in the future. Here, examiners and educational institutions can learn from the experience during the pandemic to increase the students' intention to participate in online exams in the future. To identify the problems encountered and derive lessons from them, a mixed-method approach was followed. For this purpose, 478 examinees were surveyed over 3 semesters using a quantitative questionnaire, followed by in-depth semi-structured follow-up interviews with 11 students. As a result, 5 recommendations for action could be derived, covering the topics of exam routine and familiarity, problem-solving, cheating, and teaching-examination arrangements.*

Keywords: *Emergency Remote Assessment, Online Exams, Assessment, Education.*

Citation: (Hartmann / Hobert 2023b) Hartmann, P.; Hobert, S.: *FROM EMERGENCY REMOTE ASSESSMENT TO A NEW STATUS QUO? – LESSONS LEARNED FROM ONLINE ASSESSMENTS DURING THE COVID-19 PANDEMIC*. In: *Proceedings of the 31st European Conference on Information Systems.* Kristiansand, Norway. 2023. pp. 1-15.

2.1 Introduction

Digital education has been gaining popularity for the last decades and has once again been massively reinforced by the COVID-19 pandemic (Crawford et al. 2020; Kelly / Columbus 2020). Two major advantages are attributed to digital education. First, it can be used regardless of location and time, increasing flexibility and, thus, attractiveness for participants. Second, educational institutions can save costs through the reuse and reduction of space and staff capacities (Yusuf / Al-Banawi 2013). While e-learning courses have become widespread as a result, the associated exams have nevertheless continued to be carried out in-class although the advantages mentioned above may also apply to these.

Due to the forced suspension of in-class courses during the COVID-19 pandemic, institutions around the world had to switch to online exams, so-called emergency remote assessments (UNESCO 2020). In particular, countries like Germany, where schools and universities are dominated by in-class activities, have been strongly affected by this (Marinoni et al. 2020). Since online exams were a novelty for most universities (at least in Germany), the technical and organizational foundations had to be laid. Despite extensive efforts, the number of participants in the first semester dropped by up to 50 % during the pandemic (Warnecke / Burchard 2020). This shows that the examinees' acceptance of online exams is of great importance. Indeed, different challenges influence the actual use of online exams. These challenges include, for example, the risk of problems, cheating, the fairness of grading, and equal exam conditions (Al-Maqbali / Raja Hussain 2022; Hartmann et al. 2021; OECD 2020; Tuah / Naing 2021). While studies conducted at the onset of the pandemic reflect the short-term effects of the unfamiliar situation, some influences lose or gain importance with growing familiarization over time (Zhang et al. 2021). In addition, proctoring tools used at short notice in online examinations were examined with regard to their conformity with the General Data Protection Regulation (GDPR) and found to be suitable only under certain conditions. With the return to at least partially in-class teaching, the question now is to what extent online exams will continue to be used and which experiences will help educational institutions to adapt their online exam concepts and to promote acceptance in the future. The goal of this study is, to identify positive and negative experiences made in online exams from the examinees' perspective. Moreover, it will also show examiners what lessons could be drawn from these experiences.

For this purpose, a mixed-methods approach is chosen. A quantitative questionnaire study on online exams among 478 examinees at a German university was conducted during the COVID-19 pandemic (three semesters; summer semester of 2020 to summer semester of 2021). The factors of mental challenges, cheating, and fairness of grading derived from the literature were addressed (Al-Maqbali / Raja Hussain 2022; Hartmann et al. 2021; OECD 2020; Tuah / Naing 2021). In addition, the exam type, the question and knowledge types, the technology affinity and occurring problems during the exam are taken into account to enhance the understanding of the results. To better interpret these data, a semi-structured interview study was carried out with additional 11 examinees.

2.2 Related Research

In the following, the mentioned challenges during the implementation of online exams will be considered in more detail.

Mental Challenges

Students are generally exposed to different individual mental challenges in exams, e.g., technical problems, noise pollution, or content-related problems (Elsalem et al. 2020). If these mental challenges become too disturbing, this leads to a negative attitude towards exams and, in extreme cases, to exam anxiety (Cassady / Johnson 2002). The factors that can influence exam anxiety are very different. In addition to a variety of factors that apply to online exams in general, resulting from a lack of exam perception or potential problems (Ilgaz / Afacan Adanır 2020; Ocak / Karakus 2021; Rytkönen / Myyry 2014), there are also pandemic-related influences like effects on daily-life, academic suspension and so on (Alsaady et al. 2020). Since we want to focus on online exams, we will concentrate on online-specific mental challenges and their causes. The differences between in-class and online exams result from two factors. First, the overall exam process changes, deviating from the previous routine. Second, the exams are decentralized. Therefore, the aspects of stress and concentration in online exams will be considered in the following. Thus, we assume that the concentration in the new exam environment and the additional stress caused by the changed exam process and technical environment affect the participants in addition to the actual exam performance. Here, it was shown that a lack of habituation slows down the response to the task and puts the examinees under additional time pressure when working on the exam tasks (Thomas et al. 2002). In this context, students need time to familiarize themselves with the new situation and to get used to the procedure at home. It is therefore assumed that students experience increased mental stress, especially at the beginning of online exams, but that this should decrease over time (Hillier 2014; Zhang et al. 2021). The decentralized implementation creates further problems in this regard. For example, students primarily take exams on their own devices. In addition to the need to own a suitable device, examinees must be familiar with how to use it. Although research shows that the use of one's device does bring advantages due to familiarity (Küppers / Schroeder 2018), there are also approaches that point to possible problems that can lead to stress (Elsalem et al. 2020). For instance, even though many people have some general knowledge when it comes to the use of an electronic device, very few people can solve technical problems on their own. Therefore, examinees may fear a significant disadvantage compared to other participants when technical problems arise (Küppers / Schroeder 2018; OECD 2020). In addition, due to the decentralized nature of the exam, only limited support can be provided when problems arise, which can exacerbate this effect (Gamage et al. 2020). However, this aspect does not only apply to technical problems but also to problems with the organizational process and content. These problems may result from the added responsibility, additional procedures for registration and completion, and the technical problems that may be involved (Stowell / Bennett 2010). Another important factor is the availability of an undisturbed exam environment. Not every student has access to a quiet exam environment at home, which affects concentration during the exam (OECD 2020). Therefore, one's home is sometimes considered inappropriate as an exam environment (Elsalem et al. 2020). It can be assumed that if the mental challenges for examinees are reduced, a higher acceptance of the intention to participate will be achieved in the long run.

Cheating

There are many reasons for cheating. Often, the pressure to perform in competition with other students plays an important role (Gamage et al. 2020; Noorbehbahani et al. 2022; Owunwanne et al. 2010). Thus, good grades are needed for achieving future professional goals. In addition, cheating can take place in response to the cheating of others and can thus also influence the behavior of other students during exams (Noorbehbahani et al. 2022). Examinees may feel compelled to cheat in order not to be disadvantaged (Owunwanne et al. 2010). The problem of cheating in exams is a major challenge for all universities (McCabe 2005). Even though supervision is difficult in centralized in-class exams, it is easier to control than in online exams. The use of modern technology is not the only reason why decentralized online exams pose a major challenge. Even simple attempts at communication are often difficult to monitor (OECD 2020; McCabe 2005). Especially since private devices are already being used for online exams, an increased risk of cheating among examinees can be expected (Küppers / Schroeder 2018). Furthermore, the potential for cheating increases for online exams without a proctoring process if these online exams are also written decentral (Harmon / Lambrinos 2008). For example, unproctored online exams allow examinees to use unauthorized aids as well as to communicate with third parties (Gamage et al. 2020; King et al. 2009). Especially in Germany, where there are tight legal limits to monitoring and data collection on private computers, online proctoring is not an option for many universities (Dieckmann 2021). In addition, proctoring can create further requirements for the technical suitability of the devices used (Zhang et al. 2021) and thus increase mental stress (Conijn et al. 2022). It can be assumed that by reducing cheating, the intention to participate in online exams can be increased.

Fairness of Grading

The fairness of assessment is a key element of examination, which is influenced by a variety of factors. The general assumption of fair assessment is the diagnostic quality criteria of validity, reliability, and objectivity (Joint Committee on Standards for Educational and Psychological Testing 2014; Hewlett / Kahl-Andresen 2014). Here we look at the design of the exam and the fit between the teaching and the exam, which influences the difficulty of the exam. Furthermore, the fair assessment of knowledge and skills can become a challenge in online examinations. The grading of examinations can be viewed from two perspectives. While the standards-based system aims to determine a certain level of knowledge and skills, the norm-referenced system is concerned with the evaluation of the individual's performance in relation to others (Tierney et al. 2011). Here we expect a correlation to the previously mentioned mental challenges and cheating. While all students face roughly similar conditions in in-class exams, the exam environment may differ in decentralized online exams. Studies show that students without experience with computer-based exams believe that they favor individual participants, although these biases become less significant as experience increases (Hillier 2014). In addition, it can be assumed that perceived cheating in online exams is related to perceived fairness of grading, as individual students use it to influence the classification of the individual knowledge. It can be assumed that a higher perceived fairness of grading online exams is associated with a higher intention to participate in them.

2.3 Research Design

Status Quo and Emergency Remote Assessment

Since the use and scope of examinations can differ significantly between institutions, a classification of the present scenario is given in the following. In this study, we address the case of a German university whose teaching and exams are heavily in-class oriented. The primary exam procedure is a written exam at the end of the semester in a lecture hall of the university. The time and location of the exam are predetermined. At the beginning of the exam, the examinees identify themselves with their ID card, receive the printed exam and have a specified exam time. The examiner decides whether any and if so which aids are permitted. The exam takes place under the supervision of the university staff.

Following the onset of the pandemic and the suspension of in-class teaching in spring 2020, courses were offered completely digitally in the summer semester of 2020. For the conduction of the examinations, many lecturers decided to replace the previously described written exam with other sorts of examinations. The lecturers who stuck to a written exam had the choice between in-class exams under increased hygiene requirements or online exams. Due to legal uncertainties and the fact that the necessary organizational and technical foundations were only made available at short notice, very few exams were conducted as online exams. In the winter semester of 2020/21, the majority of examinations was carried out digitally due to an increasing number of COVID-19 infections. Students avoiding online exams in the summer semester of 2020 were, therefore, also obliged to participate in online exams. A decline in the number of infections and an increase in the vaccination rate ensured that online exams were only optional in the summer semester of 2021.

The process for online exams was as follows. The examinees first had to identify themselves using a photo-ident check and were then given access to the exam. The two primary exam types used included download-upload exams and exams in an online examination system. For the download-upload exams, the exam task was downloaded and answered locally. In the end, the answered exam was uploaded to a server of the university. For the system exams, the exam tasks were answered directly in an examination system based on the ILIAS learning management system (ILIAS 2022). Supervision using proctoring tools during the exam did generally not take place due to legal concerns arising from the GDPR.

Questionnaire & Interview Study

To research the intention to participate in online exams and the related factors, a four-part questionnaire was developed for the quantitative survey. The individual parts cover the topics of demographics, the exam process, the challenges in the form of mental challenges, cheating, and fairness of grading, as well as a standardized questionnaire on technology affinity.

The demographics include the participants' age, gender, and faculty. The questionnaire on the exam process covered the exam type (download-upload exam and system exam), the question types used (e.g., single/multiple choice or essay), the tested knowledge types (e.g., replication or transfer of knowledge), and the aids permitted. In addition, the questionnaire recorded the extent to which technical, organizational, and content-related problems occurred during the examination. The challenges-related items comprise statements on the three identified areas from Section B.2.2. The pre-formulated statements had to be answered on a scale from 1 (disagree) to 7 (agree). The 7-point Likert scale offers the possibility of making a neutral statement by selecting the mean value. In addition, the finer subdivision enables the evaluation of tendencies between the mean value and the respective extreme point. A detailed list and assignment of the considered items to the factors are presented in Table 10 in Section B.2.4. The challenge-related statements were piloted with students before implementation to check their comprehensibility in the target group as well as their validity. The technology affinity was assessed using the ATI scale (Franke et al. 2019). This consists of a standardized 9-item questionnaire that captures the tendency to engage or avoid technology interaction. The assessment is based on agreement with pre-formulated statements using a scale from 1 (avoid) to 6 (engage). In the following, the respective sample is discussed. Participation in the quantitative survey study was voluntary and anonymous. No incentives were offered.

The data was collected during a three-semester period from summer 2020 to summer 2021 at a German university. Only participants of online exams got access to the questionnaire. Due to the pandemic development described in Section B.2.3, a higher number of examinees was reached in the winter semester of 2020/2021. Since the survey was conducted anonymously, students who have participated in online exams in several semesters may have participated in the questionnaire more than once. A total of 478 responses were surveyed, of whom 66.11 % were made by female and 30.13 % by male students. 2.51 % of the responses were made by diverse students or students who did not name their gender. The age varied from 18 to 43 years (MD = 22.64; SD = 3.61). A detailed overview is shown in Table 9.

Faculty	Summer Semester 2020	Winter Semester 2020/21	Summer Semester 2021
Law	37.42 %	42.15 %	20.00 %
Business & Economics	21.29 %	21.52 %	27.00 %
Social Sciences	10.97 %	17.04 %	19.00 %
Philosophy	30.32 %	19.29 %	34.00 %
Total (N = 478)	100 % (n = 155)	100 % (n = 223)	100 % (n = 100)

Table 9. Summary of Participants

The guideline of the semi-structured interview was based on the previously described questionnaire. It focused on the conception of the exam (question types and knowledge types) as well as the three challenges from Section B.2.2. In addition, the participants were asked specifically about explanations for the answers given as well as possible improvements. The semi-structured interviews lasted between 25 and 30 minutes per participant. Participation was voluntary, and no incentives were offered. In the following, the respective sample is discussed. Randomly eligible students from the faculties under consideration were requested for the qualitative interviews. A student was considered eligible if he or she participated in in-class as well as online exams at the university. Since qualitative surveys can be expected to have at least partially overlapping responses in a very narrow scenario, the principle of diminishing marginal utility was applied to the number of respondents. Here, a near saturation was found after 11 respondents. The data was collected in summer 2022.

2.4 Results

Descriptive Results of the Questionnaire Study

The results of this section are derived from the results of the quantitative survey. A detailed overview of the following statement-based results can be found in Table 10. Overall, 27.82 % of the examinees participated in system exams and 72.18 % in download-upload exams. In semester 1 and semester 2, an almost constant percentage of examinees participated in system exams (26.45 % and 24.22 %) and download-upload exams (73.88 % and 75.78 %). In semester 3, the share of system exams increased to 38.00 % (download-upload exams decreased to 62.00 %).

Question & Knowledge Types

As shown in Figure 6, fundamental differences regarding the question types used and the knowledge types tested can be identified for the two exam types and three semesters.

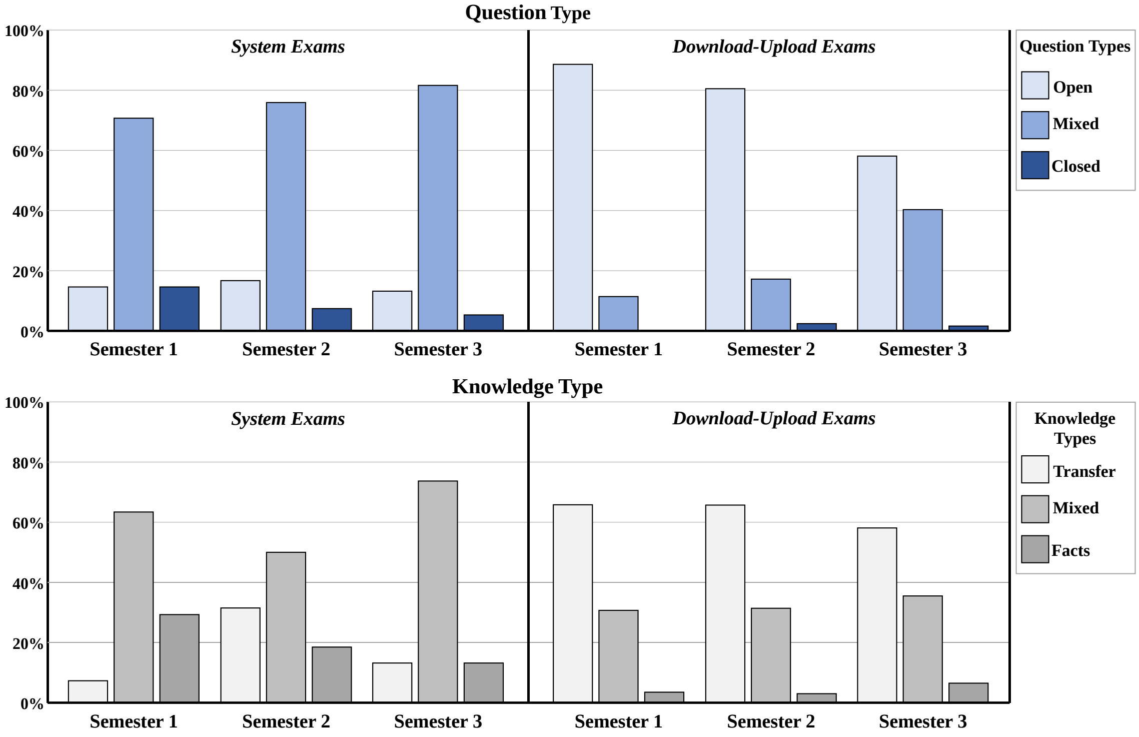


Figure 6. Knowledge and Question Types

In the system exams, the proportion of exams that were set entirely with open question types was 15.04 % overall, although this proportion fluctuated only slightly over the semesters. The proportion of exams that used only closed question types was 9.02 % overall, with the proportion decreasing continuously from semesters 1 to 3. The proportion of exams using both question types was 75.94 % across all semesters and increased over the semesters. For the knowledge types, the proportion of purely transfer-oriented exams was 18.80 % across all semesters, with the value rising sharply from semester 1 to semester 2 and falling again in semester 3. An opposite trend can be seen for exams with mixed knowledge query, which was used in 60.90 % of the system exams, but compared to semesters 1 and 3 drops sharply in semester 2. A steady decline can be observed for exams with pure factual knowledge queries from semesters 1 to 3, which were used in 20.30 % of cases overall. For the download-upload exams, the proportion of exams that used only open question types is 79.13 % across all semesters, with the value decreasing over the semesters (especially between semesters 2 and 3). In contrast, the proportion of exams that use a mixture of question types (19.42 %)

increases over the semesters. With a share of 1.45 %, examinations with exclusively closed question types are almost negligible here. There are only a few changes in the knowledge types during the semesters. The proportion of exams with only transfer knowledge (64.35 %), mixed knowledge (31.88 %), and pure factual knowledge (3.77 %) remains almost constant across the semesters.

ATI

For general technology affinity, no significant differences can be observed between semesters or exam types. The respective semester average lies in the range between 3.30 and 3.43. Since the ATI score is determined on a scale of 1 to 6, this value is close to the theoretical mean, expressing on average a neutral affinity towards interaction with technology.

Problems

Overall, there is a slight tendency for more problems occurring with download-upload exams than with system exams. For system exams, examinees reported only minor problems, with average scores between 1.83 and 1.98 in semester 1. For semester 2, an increase in all problem types can be observed. For semester 3, a weak decrease can be observed in the technical and organizational problems. The content problems remain at 2.95, almost unchanged from semester 2. This expresses a medium level of problems in this area. For the download-upload exams, no significant change in problems can be observed between the semesters. Thus, all scores range from 2.00 to 2.48 across the semesters, expressing only minor problems.

Mental Challenges

In the area of mental challenges, the study examined the extent to which online exams are less stressful than in-class exams and the extent to which it is easier to concentrate at home. For the system exams, no significant differences were observed for the stress and concentration level during the online exams between the semesters, with all values lying in the range of 3.92 to 4.89. Thus, examinees report a lower stress level and better concentration during online exams compared to in-class exams. In the download-upload exams, no significant changes were observed between semester 1 and semester 2, with values between 3.63 and 4.47. In semester 3, a significant increase in both values can be observed. The value for stress increases to 5.47, and the value for concentration to 4.89. Therefore, after a comparable stress and concentration level at the beginning of the online exams, a much lower stress level and a higher degree of concentration in online exams was reported in semester 3. This might be a sign of familiarization with online exams.

Cheating

In the area of cheating, the examinees were asked about the overall extent of and the reason for potential cheating in online exams. For the system exams, a constant level was indicated for all items over the semesters. Regarding the agreement that cheating occurs in online exams to a comparable extent as in in-class exams, values from 2.27 to 2.61 were observed on average, expressing a higher level of cheating in online exams. As an important reason for cheating, the lack of supervision was named with values ranging from 2.84 to 3.31 over the semesters. Increasing importance is attributed to the disadvantage compared to cheating examinees as a reason for cheating. The agreement with the statement that examinees' cheating does not encourage other examinees' cheating was rated with a value of 4.27 in semester 1. This agreement drops gradually to 3.53 by semester 3, expressing increasing importance. For the download-upload exams, an overall decreasing importance of cheating can be observed. In the first semesters, overall cheating is rated at a comparable level to system exams with values of 2.52 and 2.96. Compared to semester 1 an increase in the agreement was shown in semester 3 with an increase to 3.61. Thus, examinees see download-upload exams as significantly less likely to attract cheating. The relevance of the lack of supervision as a cause of cheating showed a weak positive trend but remained almost constant with values between 3.34 and 4.10. A gradual increase was also observed in the disadvantage towards cheating students as a reason for cheating. While in the first two semesters the level remained almost constant with values of 3.77 and 3.92, a significant increase to 4.68 was found in semester 3. This shows a decreasing importance from the examinees' point of view, which might be related to the question and knowledge types used.

Fairness of Grading

The fairness of grading deals with the perceived difficulty of the online exams as well as the fair assessment of individual skills. For system exams, students rated the statement that online exams are no more difficult than in-class exams at a mean of 4.95 in semester 1, with a significant decrease to 3.63 in semester 2. In semester 3, a comparable value of 3.61 was observed. The fairness for the individual assessment of skills was also found to decrease over the semesters. Thus, in semester 1 the examinees still agreed with the fair assessment with values of 4.32 and 4.51, which then decreased to values between 3.58 and 3.95 in the following semesters. This shows an increasing perceived difficulty of online exams.At the same time, the examinees assume that the fairness of the individual assessment decreased. In download-upload exams, the agreement regarding the comparable difficulty to in-class exams ranged from 3.54 to 4.27. For the fairness of the individual assessment, a constant level can be observed across all semesters, too. The average scores range from 3.93 to 4.37. Contrary to the trend in system exams, the perceived difficulty of the exams is decreasing and the suitability for a fair assessment of individual knowledge is increasing.

Intention to Participate

Overall, a comparable trend can be observed for the two types of exams over the semesters. In the system exams, a constant intention to participate was observed across all three semesters. The average agreement with the statement that online exams were preferred to in-class exams ranged from 4.24 to 4.82. This shows a neutral to slightly positive preference for online exams during the pandemic. A similar trend was observed for the intention not to return to in-class exams after the pandemic, with a gradual increase from 3.02 to 4.05 for semester 1 to semester 3. However, here a previously negative attitude changed to a neutral stance toward the future introduction of online exams. For the download-upload exams, a comparable development can be seen. While a constant level was observed in the first semesters with values of 3.99 and 4.14, a significant increase to 5.27 was identified in semester 3. This expresses a clear preference for conducting online exams during the pandemic, and could be a sign of an increasing familiarization. Regarding the continuance of online exams after the pandemic, a constant level was also observed for the first two semesters with values of 3.50 and 3.19. Likewise, semester 3 shows a significant increase to 4.48.

Based on the results, a factor analysis was conducted for the two types of exams. The same factor compositions resulted for both types of exams, so the corresponding factors Problems, Cheating, Mental Challenges, Fairness of Grading, and Intention to Participate were formed for dimension reduction. The sample has a KMO-value of 0.814 and can be considered suitable for factor analysis (Hair et al. 2018; Kaiser 1974). Table 10 includes the respective Crombach's alpha (CA) value.

Factors / Items	Exam Type	Semester 1 (MD / SD)	Semester 2 (MD / SD)	Semester 3 (MD / SD)
ATI (CA = 0.922)				
ATI-Scale based on 9 items (Franke et al. 2019)	SE DUE	3.30 / 1.28 3.32 / 1.11	3.36 / 1.06 3.33 / 1.14	3.24 / 1.21 3.43 / 1.32
Problems (CA = 0.690)				
To what extent did you experience technical problems during the online exam?	SE DUE	1.98 / 1.35 2.00 / 1.64	2.35 / 2.00 2.33 / 1.70	1.63 / 1.34 2.35 / 1.83
To what extent did you experience organizational problems during the online exam?	SE DUE	1.83 / 1.48 2.29 / 1.68	2.43 / 1.89 2.21 / 1.58	1.97 / 1.57 2.15 / 1.60
To what extent did you experience content-related problems during the online exam?	SE DUE	1.93 / 1.37 2.33 / 1.66	2.98 / 2.13 * 2.48 / 1.65	2.95 / 2.00 † 2.31 / 1.74
Mental Challenges (CA = 0.857)				
Taking exams online causes less stress than taking them at the university.	SE DUE	4.56 / 2.26 4.43 / 2.22	4.06 / 2.33 4.47 / 2.37	4.89 / 2.32 5.47 / 2.25 * †
I find it easier to concentrate in online exams than in in-class exams.	SE DUE	4.00 / 2.37 3.74 / 2.29	4.04 / 2.41 3.63 / 2.22	3.92 / 2.39 4.89 / 2.33 ** ††
Cheating (CA = 0.861)				
The percentage of cheating examinees is lower for online exams than for in-class exams.	SE DUE	2.27 / 1.47 2.52 / 1.86	2.61 / 1.76 2.96 / 2.02	2.37 / 1.85 3.61 / 2.37 ††
Students do not cheat more in online exams due to a lack of supervision.	SE DUE	3.24 / 1.96 3.34 / 2.10	3.31 / 1.92 3.41 / 2.03	2.84 / 1.90 4.10 / 2.25
Students do not cheat more in online exams due to the risk of disadvantages in online exams.	SE DUE	4.27 / 2.19 3.77 / 2.20	3.94 / 2.08 3.92 / 2.17	3.53 / 2.22 4.68 / 2.27 ††
Fairness of Grading (CA = 0.735)				
Compared to in-class exams, I found the online exam easier.	SE DUE	4.95 / 1.97 3.90 / 2.20	3.63 / 2.44 * 3.54 / 2.15	3.61 / 2.30 † 4.27 / 2.26
The grading of the online exam will reflect my actual knowledge and skills on the topic.	SE DUE	4.51 / 1.89 4.05 / 1.86	3.81 / 1.95 3.93 / 1.87	3.58 / 1.80 4.37 / 1.87
The grading of the online exam will reflect my actual knowledge & skills in comparison to my fellow students.	SE DUE	4.32 / 1.81 4.11 / 1.75	3.80 / 1.86 3.95 / 1.83	3.95 / 1.49 4.24 / 1.63
Intention to Participate (CA = 0.780)				
I wanted to write the exam as an online exam rather than an in-class exam.	SE DUE	4.27 / 2.19 3.99 / 2.47	4.24 / 2.51 4.14 / 2.50	4.82 / 2.32 5.27 / 2.31 ** ††
After the end of the COVID-19 pandemic, online exams should not be replaced again by in-class exams.	SE DUE	3.02 / 2.06 3.50 / 2.38	3.48 / 2.36 3.19 / 2.25	4.05 / 2.49 4.48 / 2.57 ** †

DUE - Download-Upload Exams; **SE** - System Exams

p-value: Compared to… …the **pre-semester** ** < 0.001; * < 0.01; …**semester 1** †† < 0.001; † < 0.01

Table 10. Descriptive Results

Statistical Analysis of the Questionnaire Study

Since the items were selected by talking to students and without an underlying model, the statistical analysis is based on the correlation of the factors. The correlations are examined for each exam type. The values above the diagonal in Table 11 represent the correlation coefficients for the system exams, while the values below the diagonal represent the values for the download-upload exams. The correlation was determined using the Kendall rank correlation coefficient (Kendall tau-b) since the factor values were determined as an equally-weighted average over the respective items and the item number for every factor varies.

DUE▼ SE►	ATI	Problems	Mental	Cheating	Fairness	Participate
ATI		0.051	0.005	-0.012	0.004	-0.017
Problems	0.036		-0.203**	0.027	-0.361**	-0.131
Mental	-0.041	-0.269**		0.164**	0.350**	0.578**
Cheating	0.061	-0.092	0.271**		0.024	0.221**
Fairness	0.009	-0.322**	0.443**	0.190**		0.186**
Participate	0.015	-0.208**	0.585**	0.299**	0.360**	

DUE - Download-Upload Exams; **SE** - System Exams

p-value: ** < 0.001; * < 0.01

Table 11. Correlation Matrix

Since the intention to participate in online exams is in the foreground, we will focus on the correlation with it. The correlation between the other factors may provide further explanations in the discussion. The results show that there is no correlation between technology affinity and participation in online exams regardless of the exam type. For system exams, furthermore, no correlation between the intention to participate and problems encountered could be proven. Thus, there is a significant correlation only between the intention to participate and the mental challenges (Tau-b = 0.578), cheating (Tau-b = 0.221), and fairness of grading (Tau-b = 0.186). In the case of download-upload exams, a significant correlation was found between the intention to participate, and the problems encountered (Tau-b = -0.208), the mental challenges (Tau-b = 0.585), cheating (Tau-b = 0.299), and fairness of grading (Tau-b = 0.360).

Results of the Interview Study

All respondents reported participation in both exam types. The view of which type of exam is most suitable is very different, as both types were attributed advantages and disadvantages.

Question and Knowledge Types

Especially in system exams, increased use of closed question types was observed. In addition, lower taxonomy levels (Bloom et al. 1956) were examined using closed question types (e.g., multiple-choice questions) more frequently than in in-class exams. One possible reason from the perspective of individual students is that examiners can have the system automatically evaluate these question types. In this context, closed-question types were rated more unsuitable than open-question types, since one often cannot justify the answers here. Although different knowledge types were used, the respondents reported an increased use of transfer tasks compared to in-class exams. The questioning of pure factual knowledge, on the other hand, declined. Thus, almost all respondents reported that fewer pure reproductions of definitions were asked. Here two respondents expressed the advantages and disadvantages of the approach. On the one hand, easy to answer questions (e.g., the mere naming of a definition) are missing as an introduction to the exam. On the other hand, the purpose of university education should not be to learn by heart and then merely reproduce word-for-word definitions.

All respondents indicated that more open-book exams were offered than before in in-class exams. According to the students, one reason for this is presumably the lack of cheating prevention in online exams. The usefulness of the lecture material in open-book exams was rated as limited as time constraints often resulted in refraining from looking up individual content.

Cheating

All respondents said that they had the feeling that more cheating occurred in online exams. Due to the change in exam assignments and allowed aids, the primary issue cited was communication between examinees. Reasons for this are the lack of supervision and the ease with which communication can take place. All respondents indicated that they only had to go through the photo-ident process for identification before taking the exam. However, since this was a snapshot, this process was deemed to be of limited use. In terms of cheating prevention, three respondents indicated that they only had to click on a sworn statement at the beginning and confirm that they were not cheating. One student also indicated that she was consistently monitored by video in a video conference during an exam. Other methods of cheating prevention were not reported. While students were able to name possible monitoring methods (audio, video, tracking, etc.), these were also reported as not being fully adequate. On the one hand, it was said that complete monitoring is not possible at the exam site and if one wants to cheat, one can cheat. On the other hand, surveillance is too much of an invasion of privacy. Therefore, cheating prevention with external tools was said to be inappropriate.

Mental Challenges

The mental challenges during an exam are very individual and, therefore, difficult to compare. Two respondents stated that they had no problems with concentration at home, as they had exam environments that were suitable for them. The remaining respondents indicated concentration problems at home due to a lack of suitable exam environments. Regarding stress, almost all examinees stated that they had experienced more stress during online exams, which arose primarily from problems that could possibly occur during the exam. Whether or not these problems occurred was presented as secondary. However, all respondents said that mental challenges decreased with the habit of online exams.

Fairness of Grading

Regarding the fairness of grading, two respondents stated that they found the exam comparably difficult to in-class exams. The rest of the examinees stated that they found the exams more demanding due to the time pressure resulting from more exam tasks and changed knowledge types. Most examinees stated that they adapted their exam preparation to the online exams. For example, factual knowledge was learned more superficially since it can be looked up again during open-book exams. Nevertheless, online exams were mentioned as a suitable means for performance assessment.

Intention to Participate

The need to conduct the exams digitally during the pandemic depended on the pandemic's development. Especially in semester 1, digital delivery was preferred due to uncertainty. For the future, all respondents stated that they would prefer to return to in-class exams.

2.5 Discussion, Implication & Limitation

In the following, the results from Section B.2.4 are discussed and the limitations of our study and results are highlighted. Since our goal is to formulate generalizable recommendations for examiners and educational institutions, it is important to clearly document the generated knowledge to make it accessible to these stakeholders. Since we see in our specific case the parallel to the formulation of design principles on how the exam process should be prepared and conducted, the documentation and communication of the lessons learned follow the components of the design principles schema, according to Gregor et al. (2020). Table 12 shows the derived recommendations for action.

Mental Challenges

One of the most important aspects of conducting online exams is mental challenges. The individual differences between examinees were shown in the results of the quantitative study. On average, comparable challenges to presentation examinations could be observed, although the results show high variances. In the qualitative study, most respondents stated less concentration and more stress at home. Regarding concentration, two factors could be identified. First, external influences on the decentralized exam environment were mentioned. Especially construction work, doorbells, and noisy neighbors were exemplary reasons for the lack of concentration. Second, the private living environment was named as a problem. Especially in student apartments or shared apartments, it is often not possible to find a suitable exam space, so exams had to take place in the comfortable living room or bedroom. For example, one student said that she was "provided with a cup of tea and something to eat" and that this had "more of a living room feeling". Another participant reported that "you can also get out of bed 5 minutes before the exam". One possible approach to improving concentration is to introduce routines that many students had already developed during in-class exams. For example, two interviewees stated that they dressed and made up as they would for in-class exams. In this regard, one respondent referred to a checklist to maintain a routine: "I always had a list where I checked off everything so that I then knew 'Okay, you've done the most important things now.' After that, I sat down for the last 5 minutes and waited quietly, as I would before an in-class exam. I always found that very relaxing in an in-class exam because you just had to sit there for 5 minutes until it started, and you couldn’t do anything". This was to provide a deliberate contrast to more relaxed dress during digital teaching. In addition, the deliberate introduction of a separate exam space was mentioned several times as a solution. As the first recommendation for action (RA1), examiners and educational institutions should enable examinees to develop individual exam routines for online exams, thereby increasing concentration and performance readiness. Similar results can be observed for stress. Here, especially the fear of problems during the conduct of the exam, regardless of their occurrence, was mentioned. These findings confirm previous research results (Ilgaz / Afacan Adanır 2020; Ocak / Karakus 2021; Rytkönen / Myyry 2014). Thus, almost all interviewees emphasized that regardless of the type of problem, the responsibility during implementation lay solely with the examinees. Although contact persons were available for different problems, the process of making contact was described as time-consuming and too complicated. Furthermore, the examinees stated that while the lecturer is responsible for the execution of in-class exams, this responsibility is transferred to the examinees. The stress of increasing responsibility is also reflected in the evaluation of the two exam types. For example, about half of the examinees explicitly referred to the reduced workload in system exams. Examples include easy navigation between and review of questions, as well as displaying the remaining exam time. In addition, the exam system automatically saves the results, eliminating the need to manually upload exam results. This represents another step in download-upload exams. In contrast, the positive, activating stress is reduced by the decentralized execution. Half of the participants stated that the gathering of all examinees in front of the exam room is perceived as activating and performance-enhancing. Thus, one sees "that one is not alone". In online exams, examinees sit alone in front of the computer and wait for the exam to begin. However, the interviewees reported that most of the stress is reduced with repeated exam participation. One interviewee said: "You have found a way for yourself how to deal with it organizationally". However, stress remains more pronounced due to fear of technical problems, as this is an existing problem. This results in two recommendations for action. First, train examinees in the exam process and system before the online examination to increase familiarity and decrease the probability of problems (RA2). Second, for problems during the exam, implement processes to resolve examinees’ problems quickly and easily (RA3).

Cheating

Both the quantitative and qualitative results show that a strong increase in cheating was expected or observed especially at the beginning of the online exams in semester 1. Although a photo-ident check had to be carried out at the beginning of the exam, this was not considered to be very effective. In particular, the lack of supervision during the exam was a major reason for this. A video recording was mentioned as a possible way to prevent cheating, but the examinees expressed concerns about privacy protection. This is an important factor that we have already highlighted in related research (Harmon / Lambrinos 2008). In addition, two respondents feared the distraction of feeling like they were being watched. Furthermore, all interviewees felt that complete cheating control is nearly impossible. For example, one student said that "camera surveillance can be helpful, but is relatively difficult to implement for large exams". In addition, a common statement was that "if someone wants to cheat, they can do it". Overall, the design of the exam was considered to be better suited. To complicate communication between examinees, the aspects of time pressure due to a higher scope of tasks as well as versioning and randomizing were presented. In addition, conducting open-book exams would "eliminate the need for supervision during the exam". The usefulness of open-book exams has already been mentioned as one solution in previous research (Zhang et al. 2021). Therefore, in our fourth recommendation for action (RA4), we recommend that exams should be designed accordingly to avoid cheating. For example, randomization, versioning, exam time, and open book are better suited to reduce cheating since control by proctoring tools can increase students' mental stress during the exam.

Fairness of Grading

For both exam types, a significant positive correlation between the fairness of grading and the intention to participate was shown. Overall, system exams were perceived as more comparable or easier than in-class exams, especially at the beginning. In the following semesters, the agreement with this statement decreased. The same applies to the perception of the suitability of system exams for a fair assessment of individual knowledge. For both exam types, a correlation with the problems encountered was also observed. Thus, the perceived fairness of grading decreases with an increasing number of problems. It can be assumed that the external problems in the decentralized implementation put individual examinees at a disadvantage. Furthermore, a correlation between mental challenges and fairness of grading could be shown. Here, too, it can be assumed that the individual disadvantages of decentralized implementation, in particular, influence the perceived fairness of grading. Surprisingly, no correlation with perceived cheating could be demonstrated for system exams. However, this can be attributed to an overall relatively high-rated cheating. The interviewees attributed an important argument for higher cheating in system exams to the design of these. System exams are characterized by more extensive use of factual knowledge and closed question types. Thus, a large part of the system exams is at least characterized by mixed question types and knowledge queries. This is partly because exam systems can automatically evaluate closed questions. This allows factual knowledge to be checked with a reduced workload for examiners. However, these are said to have a higher susceptibility to cheating as well as poorer traceability. Download-upload exams were perceived as more difficult than in-class exams in the first semester. A trend that continued in semester 2, before in semester 3 an increase to a level with semester 1 could be observed. In addition, there is a weak correlation between the fairness of grading and cheating. Compared to in-class exams, the interviewed students reported that more transfer knowledge was asked. Since it was not possible to check for cheating during the exam, more open-book exams were written in this context, in which all documents were allowed. For this purpose, the exam tasks were partly adapted to avoid the search for individual keywords in the script. This was mentioned as critical since students "prepare and orient themselves within the exam based on these terms". This is also reflected in the statements of many other interviewees reporting problems with orientation in the exam. Thus, there is a danger that students cannot answer the question precisely. One examinee said in this regard, "You also don't want to write anything irrelevant during an exam because you are already pressed for time anyway". Another student said that she leaves online exams with a bad feeling because she cannot orient herself based on the terms from the lecture and therefore does not know whether the task was solved correctly. In addition, in some cases, the feeling arose that the required knowledge goes "beyond normal student knowledge". As a result, about half of the respondents reported that they felt that this required more knowledge than in in-class exams. As the fifth recommendation for action (RA5), educational institutions should train examiners in the design of online exams. This should create appropriate teaching-examination arrangements.

To encourage students to participate in online exams with educational context …		
Principle of …	**Mechanism**	**Rationale**
… promoted exam routine (RA1)	… encourage the development of individual exam routines …	… because this may increase concentration and promotes performance readiness.
… increased familiarity (RA2)	… train these in dealing with the respective exam process and system …	… because this can already increase familiarity before the execution and thus reduce stress.
… problem-solving (RA3)	… implement easy-to-perform problem-solving processes during the exam …	… because this reduces the examinees' mental load, regardless of the occurrence of the problems.
… reduced cheating (RA4)	… reduce cheating by deliberate design of online exams (e.g., randomization, versioning, exam time, open book) …	… because cheating control by proctoring tools can increase students' mental stress in the exam.
… teaching-examination arrangements (RA5)	… train examiners to design appropriate teaching-examination arrangements …	… because a lack of fit between teaching and examination makes valid knowledge assessment difficult.

Table 12. Recommendations for Action

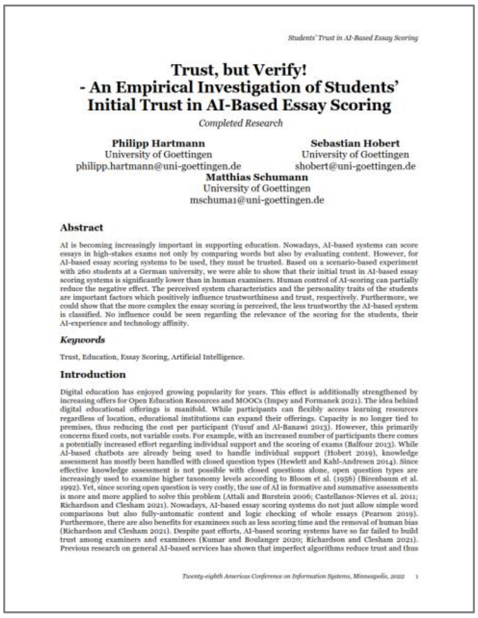
Although we added a qualitative survey to the quantitative questionnaire studies, we are aware of the various limitations of such a study. The quantitative study was conducted under the influence of the COVID-19 pandemic. Especially in the first semester, there was an adjustment in teaching and exams under increased time pressure. This effect influences possible differences in the results of the second and third semesters. In particular, the fit between teaching and exams should be higher in later semesters. Furthermore, the three semesters were strongly influenced by the development of the pandemic. Teaching and exams (optional) had to be conducted digitally without preparation in semester 1, and there was greater uncertainty at the beginning of the pandemic. In semester 2, a wave of COVID-19 infections during autumn/winter was observed in Germany, which was already expected beforehand. For this reason, teaching was offered in isolated cases as a hybrid at the beginning of the semester and was carried out digitally from Christmas onwards, as were the exams. In this case, there was a compulsion for online exams. In the summer semester of 2021, a large proportion of students and society were vaccinated, and decreasing numbers of cases were observed in the spring/summer. Here, online exams were again optional. Because we were guided by the factors identified from the literature, other effects may have occurred in practice that were not measured. Therefore, we cannot guarantee the completeness of our results. In addition, our quantitative study only addressed students who participated in online exams, regardless of whether they were required to participate in or had previous experience with the exams. Although it can be assumed that students have had experience with online exams over the semesters, it is also possible that students with no or negative experience were surveyed. In addition, as already mentioned, the survey was only conducted at one university in Germany, which means that there may be regional or national differences. The results of this study, therefore, only serve as a recommendation for action, the individual implementation of which should be discussed in consultation with the respective examinees.

2.6 Conclusion

Until the beginning of the COVID-19 pandemic, despite the increasing importance of e-learning, related exams were still often conducted in-class. The pandemic forced educational institutions to create the necessary conditions for conducting online exams, like technical or organizational infrastructure. Now that in-class exams can be held again, the question is whether the investments made can be used in the future or must be written off. For this, it is important to identify and learn from the problems identified during the pandemic. The possible lessons learned can prevent possible obstacles before they arise or reduce possible negative effects. For this purpose, we surveyed different students in the three semesters from summer 2020 to summer 2021 about their experiences during online exams. We supplemented these results with semi-structured interviews with students to be able to interpret possible causes for the observed aspects. Our results show that students consider the routine of online exams to be important for their suitability, even if they currently still prefer in-class exams. There are many reasons for this. In particular, the mental challenges as well as the problems encountered play an overriding role in the perception of online exams. These aspects can be addressed in two steps by educational institutions and examiners. One aspect is the structure of the exam process. It can be assumed that a structured exam process, the appropriate selection of exam types, and easy access to helpers will reduce stress. In addition, special importance must be attached to the design of the exams. This includes the coordinated selection of question and knowledge types with potential aids as well as the correction of the exam performance. In this way, an attempt can be made to reduce possible negative effects on the validity of the exam performance.

3 Examinee's Trust in AI-based Essay Scoring

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



Students' Trust in AI-Based Essay Scoring

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

Completed Research

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

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

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

Abstract

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

Keywords

Trust, Education, Essay Scoring, Artificial Intelligence.

Introduction

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

Twenty-eighth Americas Conference on Information Systems, Minneapolis, 2022 1

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

Keywords: *Trust, Education, Essay Scoring, Artificial Intelligence*

Citation: (Hartmann et al. 2022b) Hartmann, P.; Hobert, S.; Schumann, M.: *Trust, but Verify! - An Empirical Investigation of Students' Initial Trust in AI-Based Essay Scoring*. In: *Proceedings of the 28th Americas Conference on Information Systems.* Minneapolis, USA. 2022. pp. 1-10.

3.1 Introduction

Digital education has enjoyed growing popularity for years. This effect is additionally strengthened by increasing offers for Open Education Resources and MOOCs (Impey / Formanek 2021). The idea behind digital educational offerings is manifold. While participants can flexibly access learning resources regardless of location, educational institutions can expand their offerings. Capacity is no longer tied to premises, thus reducing the cost per participant (Yusuf / Al-Banawi 2013). However, this primarily concerns fixed costs, not variable costs. For example, with an increased number of participants there comes a potentially increased effort regarding individual support and the scoring of exams (Balfour 2013). While AI-based chatbots are already being used to handle individual support (Hobert 2019), knowledge assessment has mostly been handled with closed question types (Hewlett / Kahl-Andresen 2014). Since effective knowledge assessment is not possible with closed questions alone, open question types are increasingly used to examine higher taxonomy levels according to Bloom et al. (1956) (Birenbaum et al. 1992). Yet, since scoring open question is very costly, the use of AI in formative and summative assessments is more and more applied to solve this problem (Attali / Burstein 2006; Castellanos-Nieves et al. 2011; Richardson / Clesham 2021). Nowadays, AI-based essay scoring systems do not just allow simple word comparisons but also fully-automatic content and logic checking of whole essays (Pearson Education Ltd 2019). Furthermore, there are also benefits for examinees such as less scoring time and the removal of human bias (Richardson / Clesham 2021). Despite past efforts, AI-based scoring systems have so far failed to build trust among examiners and examinees (Kumar / Boulanger 2020; Richardson / Clesham 2021). Previous research on general AI-based services has shown that imperfect algorithms reduce trust and thus acceptance (Kocielnik et al. 2019). Hence, when it comes to educational issues, students have more trust in people they know in the field than in the technologies being used (Richardson / Clesham 2021). This may be because AI-based essay scoring has its limitations such as the dependence on training data (Kumar / Boulanger 2020). Especially when examinees have to give their own opinion or a freely chosen example, AI reaches its limits. User trust is a particularly important but multifaceted construct here, influencing acceptance and thus usage (Wu et al. 2011). In the following, we will therefore investigate which factors influence an examinee’s trust in AI-based scoring systems. In this context, trust in a relationship depends on three dimensions, namely the trustor (examinee), the trustee (AI-based essay scoring system), and the environment or situation (high-stakes exams), which are determined by different factors (Siau / Wang 2018; Mayer et al. 1995). While previous research has often focused on the trust of active users, we will look at the trust of passive users, who do not use the system themselves but are affected by its decisions. Thereby, trust is considered a dynamic system that consists of an individual basic trust (initial trust) as well as a trust that develops during the interaction (continuous trust)

(Siau / Wang 2018). Since the use of AI-based essay scoring in high-stakes exams is still in its infancy, we will focus on initial trust. Initial trust describes the first contact between the two parties and is crucial for supporting the adoption of new technology. It is based on pre-implementation expectations (Li et al. 2008).

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

3.2 Related Research and Hypotheses Development

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

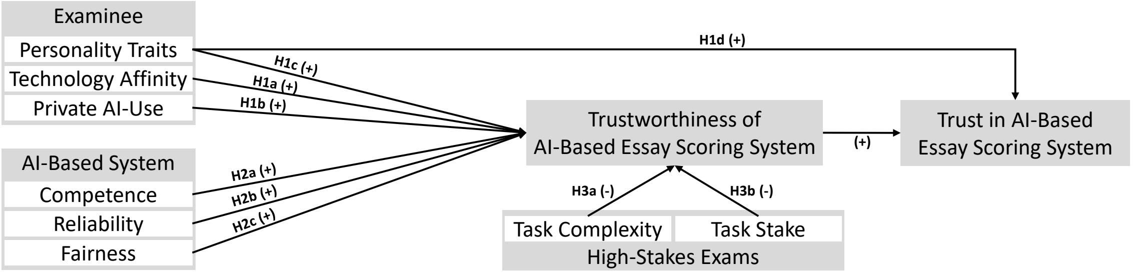


Figure 7. Trust Model Used for this Research Study

Examinee (The Trustor)

In our model, the examinees take on the role of human trustors. Each trustor has an individual propensity, i.e., willingness, to trust (Mayer et al. 1995). It is based on a generalization of various unique experiences (Lee / See 2004). The propensity to trust can be subdivided into ability- and personality-based factors (Siau / Wang 2018). Ability-based factors are grounded on information and knowledge about the trustee as well as on prior experiences and help to form predictions about the system's behavior. Since there is no comparable system in the context under investigation, the trustors do not have any information or knowledge from prior use of AI-based essay scoring systems. Therefore, this aspect is examined using the students' overall technology affinity and experience with other AI-based services (e.g., virtual assistants like Amazon Alexa or Apple's Siri). Former research showed that a high technology affinity promotes an increased tendency to actively approach and thus trust new technologies (Franke et al. 2019). We follow this argumentation and expect that a similar impact exists through the use of other AI-based services because experience with AI-based services in a private environment promotes understanding / reputation and hereby trust in other areas of use (Bao et al. 2021). Personality-based factors reflect the trustor's personality traits (Oleson et al. 2011). Prior research describes trust-related personality traits as the basis for general trust before having information on a particular trustee (Siau / Wang 2018; Mayer et al. 1995). Especially in case of initial use, without sufficient information for a cognitive evaluation of the system, different personality traits (e.g., agreeableness) influence the emotional response to the system (Madsen / Gregor 2000; Bao et al. 2021). We assume that a higher agreeable personality trait leads to a higher trustworthiness of the AI-based scoring system as well as a higher overall trust in the AI-based scoring.

H1a: *A higher technology affinity leads to a higher expected trustworthiness of the AI-based essay scoring system.*

H1b: *A higher experience in private use of AI leads to a higher expected trustworthiness of the AI-based essay scoring system.*

H1c: *A higher agreeable personality trait leads to a higher expected trustworthiness of the AI-based essay scoring system.*

H1d: *A higher agreeable personality trait leads to a higher trust in the AI-based essay scoring system.*

AI-Based Essay Scoring System (The Trustee)

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

H2a: *A higher expected competence of the AI-based system leads to a higher expected trustworthiness of the AI-based essay scoring system.*

H2b: *A higher expected reliability of the AI-based system leads to a higher expected trustworthiness of the AI-based essay scoring system.*

H2c: *A higher expected fairness of the AI-based system leads to a higher expected trustworthiness of the AI-based essay scoring system.*

High-Stakes Exams (The Environment)

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

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

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

3.3 Research Design

Scenarios and Questionnaire Introduction

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

In section 1, participants were asked about their demographic information, including age and gender. To ensure that all participants had a common knowledge concerning the scoring of essays at high-stakes exams, a short animated video about an exemplary exam situation and the associated scoring process was shown. Since the type and length of exam assignments can vary between courses, it was stated that only essays of approximately half to three-quarters of a page in length are included in the exam. The tasks included the reproduction, explanation, and transfer of the learned contents. The described scoring process represents the common procedure at German universities, which is carried out completely manually. Here, a four-eye principle was presented, which consists of a pre-scoring by a qualified employee and a final scoring by the professor in charge. In addition, the students were informed that this procedure entails longer scoring times, especially for larger courses. Based on this scenario, the

AI-independent items were collected first. These included the trust in the described manual scoring process as well as an estimation of the expected scoring accuracy. For the trustor characteristics, the personality trait was queried using the German adaption of the Big Five Personality Traits Taxonomy (John et al. 2008), focusing on the trust-facet (dimension agreeableness). The ability factors were measured by using the students' technology affinity and individual experience in using AI-based services. The technology affinity was examined by employing the ATI-scale, consisting of a standardized questionnaire covering 9 items about engaging or avoiding technology interaction (Franke et al. 2019). The individual experience was assessed based on the frequency of use of voice assistants, facial recognition, and individual recommendation systems. Section 1 closed with the environmental factors, using the task complexity and task stakes.

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

Data Collection and Pre-Processing

Factor	Items
Competence (CA = 0.765)	The AI-based system has in-depth knowledge of scoring exams.
	The scoring results of the AI-based system are as good as those of a highly competent person.
	The AI-based system correctly scores the exam answers I submit.
	The AI-based system uses all the knowledge and information at its disposal to score an exam.
Reliability (CA = 0.704)	The AI-based system works reliably.
	The AI-based system scores comparable exam answers of different exam participants equally.
	I can rely on the AI-based system to work flawlessly.
	The AI-based system scores the exam answers without contradictions.
Fairness (CA = 0.625)	I believe that an AI-based system would be used in my best interest.
	The AI-based system looks after my interests, not just those of the professor.
	During AI-based scoring, preference is given to individual examinees.
	The AI-based system ensures a fair scoring of the individual performance of examinees.
Task Complexity (CA = 0.563)	The scoring of exams is demanding.
	For the scoring of an exam task, one needs a specialized knowledge that exceeds the knowledge for the answering of the task.
	The optimal answer to an exam task is always unique.
	Errors rarely occur in the scoring of exams.
Task Stake (CA = 0.615)	The correct scoring of an exam is very important.
	The grade in an exam has a long-term impact on the student's life.
	I care about good grades.
	If I get a lower grade than expected, I don't think about it for very long.
Trustworthiness of AI-Based Essay Scoring System (CA = 0.800)	The AI-based scoring process is trustworthy.
	I would change one or more aspects of the scoring process to make AI-use trustworthy.
	The AI-based scoring process will result in a fair outcome for the examinees.
	Examinees need more information about how the AI-based system scores in order to trust the scoring process.
Trust in AI-Based Essay Sc0ring System (CA = 0.632)	I trust the AI-based scoring process of exams.
	I would like to keep an eye on the AI-based system during scoring.
	The exam reviews of the final scoring by examinees are needed to control the AI-based scoring.
	The AI-based system should be more controlled.
	For the AI-based system, its own interests (e.g., the lowest possible scoring effort) are paramount in the scoring process.

Table 13. Reliability Coefficients of the Factors and Items Used

The questionnaire was forwarded to students at a German university via multiple channels (e.g., e-mail, forum, personal addresses in classroom lectures). Participation was anonymous and voluntary. Vouchers were raffled among all participants who completed the questionnaire. A weighting of the participants according to gender, age, or other criteria was not carried out. A total of 330 students took part in the survey. Due to the use of incentives, it can be assumed that some participants did not show the required seriousness. To reduce disruptive effects, we tried to remove these participants by identifying outliers in the processing time. This leaves a data sample of

260 participants, of whom 51.92 % were male and 48.08 % female. Their age varied between 18 and 35 years (MD = 21.85; SD = 2.69). 51.54 % of the participants were shown scenario 1. Scenario 2, on the other hand, was seen by 48.46 %.

In the selection of items used, we drew on existing and scientifically tested items, which were adapted to the subject of manual and (semi)-automatic AI-based essay scoring. To assess the fit of the model with our collected data, a confirmatory factor analysis was conducted for the existing scales, while an exploratory factor analysis was used for the others. Since we used the already validated ATI-score (Franke et al. 2019) to measure the technology affinity and the Big Five Personality Traits Taxonomy (John et al. 2008) to measure the personality trait, we excluded these items from the factor analysis. The sample has a KMO-value of 0.840 and can be considered suitable for factor analysis (Kaiser 1974; Hair et al. 2018). Due to cross-loadings and poor factor loadings, we removed certain items to ensure construct reliability (highlighted in gray in Table 13). For the remaining items, we conducted tests for convergent validity by determing the composite reliability (CR) and the average variance extracted (AVE). The values for both indicators are above the critical values and therefore at an acceptable level (Hair et al. 2018). The items used and the associated Cronbach's Alpha (CA) values for the identified factors are listed in Table 13.

3.4 Results

Descriptive Analysis

Regarding the scenario-independent factors, the following values were obtained. For trust in manual scoring, the participants stated that they trusted the scoring in principle (MD = 3.86; SD = 0.83). Among the personal trait factors, above-average values can be observed for the trust-facet (MD = 3.76; SD = 0.67). When dealing with technologies, the ATI shows a mean average technology affinity (MD = 3.65; SD = 1.08). Greater differences are evident in the use of AI-based services. For example, when using voice assistants, 60.8 % said that they use them only once a month or fewer, whereas only 20.4 % use them (almost) daily. Regarding the use of facial recognition, 35.0 % indicated infrequent use, while 62.3 % use it (almost) daily. For the use of individual recommendations, the proportions are 32.3 % and 28.1 %. High values can be observed for the environmental factors. The participants rated the scoring of exams as complex (MD = 4.94; SD = 0.84) and important (MD = 4.89; SD = 0.86). T-tests show no significant difference between the participants of the scenarios. The results for the scenario-dependent factors are shown in Table 14.

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

Table 14. Descriptive Results of AI-Related Factors

Furthermore, in the semi-automatic scenario, 59.7 % of the respondents indicated that they would be more likely to attend an exam review if AI was used. In the automatic scenario, the proportion was 71.4 %. The statistical analysis showed that there is a significantly higher percentage of students in scenario 2 who would participate in the review (p<0.001). The expected accuracy of the manual scoring was reported by the participants with a mean value of 83.71 % (SD = 9.63 %). Surprisingly, no significant difference to the AI-based scoring can be observed. Thus, the participants indicated comparable accuracies for the semi-automatic scoring from scenario 1 (MD = 83.99 %; SD = 12.21 %) and for the automatic scoring from scenario 2 (MD = 82.61 %; SD = 13.00 %). T-tests show no significant difference between the scenarios and in comparison to the human scoring.

Statistical Analysis

To test the postulated hypotheses, the statistical software Stata SE was used. Before we conducted the structural equation model, the assumptions were checked. The multivariate normality was checked using the Mahalanobis distance. No further outliers were observed. For multicollinearity, all VIF-values and tolerances were at an acceptable level (Hair et al. 2018). The structural equation model was estimated using the maximum likelihood estimation and model fit indices were determined. The coefficients and the corresponding significance levels can be seen in Figure 8. Overall, the model has an acceptable to good fit for different quality indices. So the values for RMSEA (0.080), CFI (0.971), TLI (0.930), and SRMR (0.023) are all at an acceptable level (Hair et al. 2018).

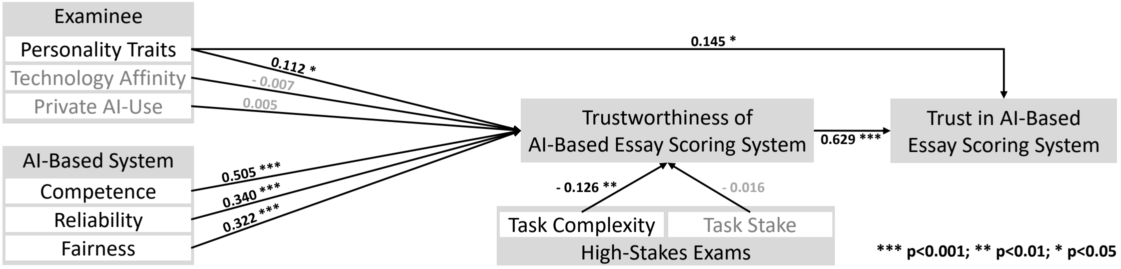


Figure 8. Results of the Structural Equation Model

3.5 Discussion and Implication

Examinee (The Trustor)

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

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

AI-Based Essay Scoring System (The Trustee)

The trustee characteristics influence how the trustee is perceived by the trustor and the amount of trustworthiness assigned to him. We tested the factors of competence, reliability, and fairness. The hypotheses H2a to H2c were all confirmed. Fairness was rated as equally high in both scenarios. Thus, the additional spot-checks by the professor did not lead to any changes. For reliability, a significant difference was observed between the scenarios: the value of the semi-automatic is higher than that of the automatic scoring. The system seems to have a lower overall reliability, which can be partly compensated by the control of the professor. We observed a significant difference regarding competence. Here, too, the semi-automatic process is perceived as significantly more competent. So, the role of AI in the scoring process has an important influence. Indeed, previous research has shown higher trust in human deciders than in automatic AI-based systems, especially for important decisions (Ashoori / Weisz 2019). In our initial scenario, we described the task as the reproduction, explanation, and transfer of learned content. The system does not seem to be trusted to possess a competence equal to that of humans. One reason for this may be the task of explanation and transfer, whose answers cannot be classified into right or wrong in the level of expectations and are thus difficult to teach to the system. The closer the answer to a firmly defined level of expectations, the higher the quality of the scoring. Here, a lack of transfer to individual examples could be a possible cause for the lower perceived competence. Previous research has also shown that students still trust the people they associate with the activity more than systems (Elson et al. 2021; Richardson / Clesham 2021). This may show a negative image of AI since these results are in contradiction to the expected accuracy, where we could not identify any differences between the manual scoring and the scenarios. Overall, system-related factors represent the most important influence on trustworthiness and thus trust over AI-based essay scoring. As a result, an attempt could be made to increase fairness and reliability through the transparent implementation of protocols for proper essay scoring.

High-Stakes Exams (The Environment)

For the environment characteristics, we focused on the task-related factors. Hypothesis 3a, in which we stated a negative relationship between the perceived task complexity and the trustworthiness of the AI-based essay scoring system, was confirmed. Thus, the task was perceived as very complex, with a significant negative influence on trustworthiness confirming the results of previous research (Ashoori / Weisz 2019, 5). Hypothesis 3b, assuming that a high perceived relevance of the scoring also influences trustworthiness, could not be confirmed. Although the correct scoring of exam tasks was also assigned as important, this did not have any significant influence on trustworthiness in our case. One reason for this could be a good task-AI fit. For the trustor, the appropriate completion of the task is of primary importance. If a trustee, in our case the AI-based scoring system, is in sum rated as competent to perform the assigned task, it may not matter how relevant the task is to the examinee. An indication of a good task-AI fit may be that in both scenarios the scoring was perceived as fair and the system as competent. Depending on the task complexity, we recommend to design the use of the AI-based system appropriately. We therefore suggest, that for complex tasks, the semi-automatic use of AI as a decision support system should be considered. Thus, human control can increase perceived competence and ensure a better task-AI fit.

3.6 Limitation

As with any similar quantitative questionnaire study, we are aware of various limitations. First, our model attempts to explain trust in AI-based (semi-)automatic essay scoring in high-stakes exams through the trustworthiness of the AI-based system and personal characteristics. By conducting a factor analysis, we combined multiple existing and newly created items into the postulated factors. Since the subject of trust in AI-based essay scoring is quite new, we cannot assure that our results are complete. Thus, many assumptions of the model under consideration are based on the trustor as an active user. In our case, however, the examinees represent passive users who are just confronted with the results. In this area, prior research is still in its infancy. Second, we primarily tried to use existing, valid items, which were translated and adapted to our context and target group. As a result, important linguistic facets may have been lost. Additionally, a narrow set of factors (competence, reliability and fairness) was selected for the AI-based system characteristics, so that possible dimensions may not have been considered. Furthermore, the personal traits were measured using the Big Five Personality Traits Taxonomy (John et al. 2008). The determination of a complete personality profile can comprise up to 240 items and is therefore difficult to implement in the context mentioned (Costa; Jr. / McCrae 2000). Here, the focus was placed only on the trust-facet as part of the dimension agreeableness, which is measured by 4 items. Overall, other personality traits could also influence trust. Future research should therefore focus on additional personality traits to provide further insights into the influence on the trustworthiness of and trust in AI-based services. Third, new items were developed for individual constructs. In this respect, the factor analysis revealed possibilities for improvement. The difficulty in operationalizing trust is that different items are reliable for measuring trust in AI and trust in humans. It is therefore difficult to formulate a uniform set of items that allows direct comparisons of humans and AI. Furthermore, there is room for improvement in the scales for the environmental factors. To be able to validly assess trustworthiness and thus trust in different situations, it must first be possible to clearly define the situational context. Here, our Cronbach's Alpha-values for task complexity and task stake still show potential for improvement. Fourth, as mentioned before, the survey was only conducted at one university in Germany. Although the culture and scoring process among German universities is quite similar, there may occur regional as well as national differences. The results of this study can therefore only represent a starting point for further research and still needs to be verified regarding its generalizability.

3.7 Conclusion

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

4 Design of AI-based Essay Scoring Systems

(AI)N'T NOBODY HELPING ME? – DESIGN AND EVALUATION OF A MACHINE-LEARNING-BASED SEMI-AUTOMATIC ESSAY SCORING SYSTEM

(AI)N'T NOBODY HELPING ME? – DESIGN AND EVALUATION OF A MACHINE-LEARNING-BASED SEMI-AUTOMATIC ESSAY SCORING SYSTEM

Research Paper

Philipp Hartmann, University of Goettingen, Germany, philipp.hartmann@uni-goettingen.de
Nils Holthoff, University of Goettingen, Germany, nils.holthoff@stud.uni-goettingen.de
Sebastian Hobert, University of Goettingen, Germany, shobert@uni-goettingen.de
Matthias Schumann, University of Goettingen, Germany, mschuma1@uni-goettingen.de

Abstract

Education is increasingly being delivered digitally these days, whether due to the COVID-19 pandemic or the growing popularity of MOOCs. The increasing number of participants poses a challenge for institutions to balance didactical and financial demands, especially for exams. The overall goal of this design science research project is to design and evaluate a semi-automatic machine-learning-based scoring system for essays. We focus on the design of a functional software artifact including the required design principles and an exemplary implementation of an algorithm, the optimization of which is, however, not part of this project and subject for future research. Our results show that such a system is suitable for both scoring essay assignments and documenting these scorings. In addition to the software artifact, we document our results using the work of Gregor et al. (2020). This provides a first step towards a design theory for semi-automatic, machine-learning-based scoring systems for essays.

Keywords: automated essay scoring, machine learning, design principles, education, learning

1 Introduction

The process of digitization has had a strong influence on education over the last years which has been reflected in the increasing number of online courses (Impey and Formanek, 2021). This trend has further been reinforced by the COVID-19 pandemic, as in-class lectures and exams have been suspended and conducted digitally to fight the pandemic (Crawford et al., 2020; Kelly and Columbus, 2020; UNESCO, 2020). In addition to the didactic benefits of multimedia content, e-learning is primarily used to reduce costs (Lai and Liou, 2010; Yusuf and Al-Banawi, 2013). For example, digital learning modules can be reused and allow a larger number of participants due to location-independent use, such as in MOOCs. As a consequence of the growing number of participants, there is also an increasing effort in assessing the taught content at final exams (Balfour, 2013). Whereas in the past, mainly the fully automatically assessable, closed question types (e.g., multiple-choice questions) were used for the examination, the use of essay tasks in e-assessments is also becoming increasingly relevant. From a didactic perspective, essays are more suitable for evaluating competences because through them the understanding of complex relationships, for example the higher taxonomy levels of Bloom et al. (1956), can be tested (Birenbaum et al., 1992; Castellanos-Nieves et al., 2011). While open-ended questions are not difficult to conceptualize, manual scoring proves complex because each response is nearly unique (Attali and Burstein, 2006; Richardson and Clesham, 2021). In particular, when based on superficial features of the answer, such as the use of important terms, students writing generalities and nonsensical content,

Thirtieth European Conference on Information Systems (ECIS 2022), Timisoara, Romania *1*

Abstract: *Education is increasingly being delivered digitally these days, whether due to the COVID-19 pandemic or the growing popularity of MOOCs. The increasing number of participants poses a challenge for institutions to balance didactical and financial demands, especially for exams. The overall goal of this design science research project is to design and evaluate a semi-automatic machine-learning-based scoring system for essays. We focus on the design of a functional software artifact including the required design principles and an exemplary implementation of an algorithm, the optimization of which is, however, not part of this project and subject for future research. Our results show that such a system is suitable for both scoring essay assignments and documenting these scorings. In addition to the software artifact, we document our results using the work of Gregor et al. (2020). This provides a first step towards a design theory for semi-automatic, machine-learning-based scoring systems for essays.*

Keywords: *automated essay scoring, machine learning, design principles, education, learning*

Citation: (Hartmann et al. 2022a) Hartmann, P.; Holthoff, N.; Hobert, S.; Schumann, M.: *(AI)N'T NOBODY HELPING ME? – DESIGN AND EVALUATION OF A MACHINE-LEARNING-BASED SEMI-AUTOMATIC ESSAY SCORING SYSTEM*. In: *Proceedings of the 30th European Conference on Information Systems*. Timisoara, Romania. 2022. pp. 1-16.

4.1 Introduction

The process of digitization has had a strong influence on education over the last years which has been reflected in the increasing number of online courses (Impey / Formanek 2021). This trend has further been reinforced by the COVID-19 pandemic, as in-class lectures and exams have been suspended and conducted digitally to fight the pandemic (Crawford et al. 2020; Kelly / Columbus 2020; UNESCO 2020). In addition to the didactic benefits of multimedia content, e-learning is primarily used to reduce costs (Lai / Liou 2010; Yusuf / Al-Banawi 2013). For example, digital learning modules can be reused and allow a larger number of participants due to location-independent use, such as in MOOCs. As a consequence of the growing number of participants, there is also an increasing effort in assessing the taught content at final exams (Balfour 2013). Whereas in the past, mainly the fully automatically assessable, closed question types (e.g., multiple-choice questions) were used for the examination, the use of essay tasks in e-assessments is also becoming increasingly relevant. From a didactic perspective, essays are more suitable for evaluating competences because through them the understanding of complex relationships, for example the higher taxonomy levels of Bloom et al. (1956), can be tested (Birenbaum et al. 1992; Castellanos-Nieves et al. 2011). While open-ended questions are not difficult to conceptualize, manual scoring proves complex because each response is nearly unique (Attali / Burstein 2006; Richardson / Clesham 2021). In particular, when based on superficial features of the answer, such as the use of important terms, students writing generalities and nonsensical content, including the terms sought by the examiners, can cause concentration problems when there are a large number of exams to be scored (Castellanos-Nieves et al. 2011). Therefore, automating the process of scoring essay tasks could significantly reduce the workload for examiners. In addition, the scoring process could be accelerated and the standardization of grading that accompanies automated scoring could lead to greater consistency and fairness in the grades awarded (Richardson / Clesham 2021; Hung et al. 1993). There are already approaches that deal with automated scoring of essay tasks (Ramesh / Sanampudi 2021). However, these works mostly address the technical perspective rather than the underlying process of scoring and the examiners' needs. Previous research has shown that trust towards AI-based services, and thus their acceptance or use, is influenced by the relevance of the decision (Ashoori / Weisz 2019; Lee / See 2004). Decisions to which users assign a high relevance, such as high stake exams, are often met with reluctance (Ashoori / Weisz 2019). In addition, human influence on the final decision is often considered important. Thus, users often feel more personally attached if the final decision is made by a human being and the AI-based service merely serves to support the decision (Ashoori / Weisz 2019). Therefore, in order to gain the users' acceptance, their needs must be considered. In the following, a holistic system for semi-automated essay scoring is considered, taking into account these user requirements. Within the semi-automated system, tasks are accomplished through an appropriate mix of human labor and automated, computerized assistance (Frohm et al. 2008). The scoring system supports the evaluation of answers to essay tasks in the first step by an automated pre-scoring of the answers using an adaptive system and in a second step by helping examiners with the manual post-scoring.

In our research project, we focus on the design of a first functional software artifact that is able to support essay scoring using a semi-automated essay scoring system. To achieve this, we implement a fully-functional user interface artifact and include a first version of a machine-learning-based scoring algorithm. Thus, the aim of this project is to identify core design principles for semi-automated essay scoring systems. Within our research project, we follow a design science research (DSR) approach based on PEFFERS ET AL. (2007) and HEVNER ET AL. (2004) to answer the following research questions.

RQ1: *How to design a machine-learning-based, semi-automatic scoring system for essays?*

RQ2: *How do potential users assess the semi-automatic scoring system, supporting the essay-grading process?*

Within this research approach, we aim to contribute design knowledge on how to implement semi-automated scoring systems in educational contexts. In the remainder of this paper, we demonstrate and evaluate a first software artifact consisting of a fully-functional user interface and an exemplary machine-learning-based scoring algorithm. The developed software artifact is based on two design-build-evaluate iterations (March / Smith 1995). While deriving design principles and demonstrating a functional software artifact is the goal of this research project, improving and optimizing the scoring algorithm is subject to future research.

The remainder of this paper is structured as follow: Next, we present related research on semi-automatic scoring systems. Following this, our DSR approach used is first briefly outlined and then applied to the problem in detail. Subsequently, the results are discussed, and the derived design knowledge is summarized based on GREGOR ET AL. (2020).

4.2 Related Research on Semi-Automatic Essay Scoring

The automatic scoring of essays has been a topic of research for a long time. However, due to technological limitations, these approaches have usually been restricted to single, isolated factors of text composition (e.g., response and word length or grammatical correctness) and the identification of individual terms. For instance, MITCHELL ET AL. (2003) were able to observe an accuracy of semi-automatic scoring of almost 95 % for short free-text responses at an early stage. This accuracy was consistent with the accuracy of human examiners but required the creation of complex solution patterns for each task. With the help of an authoring tool, mark scheme templates were created in which syntactic-semantic structures were created by the examiners. The individual parts of the level of expectation were manually broken down into their individual components (nouns, verbs and prepositions). Subsequently, the identified terms were extended by synonyms, which were also to be scored as correct. Due to technical progress as well as the increasing availability of data, approaches for open questions of natural language are feasible nowadays. The use of machine learning enables accurate predictions to be made on the basis of existing empirical values (Mohri et al. 2018). The quality of the scoring depends on the quality and the extent of the empirical values, which, for example, consist of a previous scoring of comparable tasks by human examiners. The approaches that can be taken in this regard differ. TAGHIPOUR / NG (2016) use an approach that works with a long-short-term memory network for evaluation. A Kaggle dataset (Kaggle 2012) is used, with 60 % of the data as the training, 20 % as the development and 20 % as the test dataset. The evaluation takes place on a technical level, with the best model architecture for essay scoring being sought. In contrast, SHARMA / JAYAGOPI (2018) combine two neural networks for transcribing and scoring handwritten responses to essay tasks. The training dataset used 90 % of the data and the validation dataset 10 %. No differences were observed between AI-based and manual transcription. Another method is the memory network described by ZHAO ET AL. (2017), outperforming a comparable LSTM approach in 7 out of 8 sets. CHEN ET AL. (2010) used an unsupervised automated essay scoring system, which has an adjacent agreement rate of more than 90 % and an exact agreement rate of 50 %.

In addition, studies have shown that automated essay scoring is highly reliable (Foltz et al. 1999). Thus, individual scoring systems have shown a high correlation between the AI and the examiner scores (Attali / Burstein 2006; Pearson Education Ltd 2019). COHEN ET AL. (2018) also investigated the validity of automated essay scoring, where the "true scores" were determined as the mean across a group of examiners scoring the same task. Compared to a single examiner, the AI was able to achieve comparable results. In contrast, the validity of two or more examiners was significantly higher. From this, they derived that a system as support outperforms careless examiners in particular. Automatic scoring systems are also said to be highly objective. This results primarily from the elimination of the direct influence of human bias, whereby the dependence of the scoring is affected by the scope, quality, and objectivity of the underlying data (Kumar / Boulanger 2020).

The examples outlined above as well as other identified literature most often focus on the algorithm or model of assessment. However, little attention has been paid to other aspects, such as the process of scoring and the resulting requirements of the examiners. An evaluation of the whole system (instead of only focusing on algorithms) is rarely done. Since the automatic assessment of essays works well, though not perfectly, semi-automatic systems may be a possible solution. For the reasons stated in the introduction, a semi-automatic essay scoring system is developed in this study. The system is at the fourth level of automation, according to the classification of BILLINGS (1997). Here, the system supports the assigned activity, but the user remains in full control of the AI.

4.3 Research Design

To achieve our research goal of developing and implementing a semi-automatic scoring system for essay answers in exams, we apply a DSR method, as shown in Figure 9. Thereby we intend to (1) develop an artifact for solving the problems listed in the introduction and (2) derive generalizable design principles for the Information Systems (IS) research discipline, according to GREGOR ET AL. (2020). Our research design is based on the DSR process proposed by PEFFERS ET AL. (2007), which is suitable when the focus is on developing a practically applicable artifact using a flexible process. In addition, a practical and outcome-oriented evaluation of the artifact will be conducted. Since the present project aims at the design and implementation of an IT artifact in the sense of HEVNER ET AL. (2004), their guidelines are additionally applied. Figure 9 shows an overview of the applied process. In green, the process according to PEFFERS ET AL. (2007) with its six phases is shown. The guidelines described by HEVNER ET AL. (2004) are shown in yellow and assigned to the six phases. In blue, equivalent to the phases of the process, the concrete implications for this work are shown.

The first phase deals with identifying the problem by describing it in a comprehensible way based on the user stories and the challenges of the examiners. This is supposed to clarify the added value of our artifact. In the second phase, building on the previous phase, we establish the requirements before deriving the design principles for the planned artifact. Through this, we want to show the potential for improvement compared to manual scoring. The third phase comprises the design and implementation of the artifact. In the first iteration, we focus on the user interface and the presentation of the relevant requirements for the application. In the fourth phase of the DSR process, we demonstrate the application to examiners, and in the subsequent fifth phase, we evaluate the artifact, paying special attention to the user interface and its basic functionalities for the users. The evaluation results are compared with the goals defined in the second phase and, if suitable, implemented in the following. In the second iteration, the evaluation results of the user interface from iteration one are used for improving the system and the scoring mechanism is implemented. The subsequent evaluation of the second iteration focuses on the scoring mechanism. Here we examine the technical suitability of machine learning in scoring essay answers and also address the question of whether it is suitable for out-of-domain data. For the phase of the documentation and communication of the gained knowledge, we use the components of the design principles recommended by GREGOR ET AL. (2020).

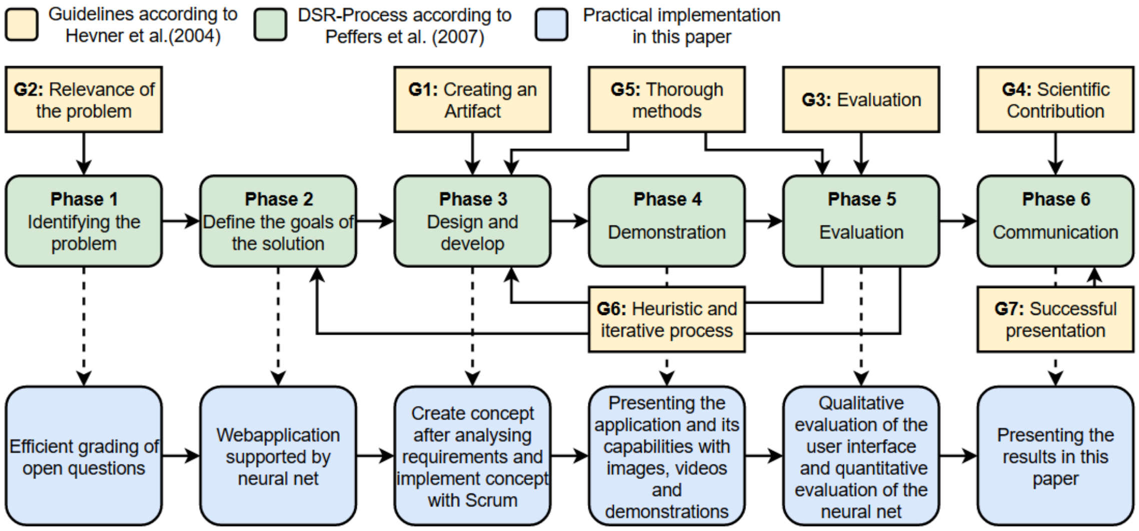

Guidelines according to Hevner et al.(2004)
DSR-Process according to Peffers et al. (2007)
Practical implementation in this paper
G2: Relevance of the problem
G1: Creating an Artifact
G5: Thorough methods
G3: Evaluation
G4: Scientific Contribution
Phase 1 Identifying the problem
Phase 2 Define the goals of the solution
Phase 3 Design and develop
Phase 4 Demonstration
Phase 5 Evaluation
Phase 6 Communication
G6: Heuristic and iterative process
G7: Successful presentation
Efficient grading of open questions
Webapplication supported by neural net
Create concept after analysing requirements and implement concept with Scrum
Presenting the application and its capabilities with images, videos and demonstrations
Qualitative evaluation of the user interface and quantitative evaluation of the neural net
Presenting the results in this paper

Figure 9. Adapted Research Approach

4.4 Design and Evaluation

Specifying the Problem Statement

As stated in the introduction, the challenges of scoring by examiners are addressed in this paper. The goal is to design and implement a machine-learning-based system that automatically performs a pre-scoring of essay tasks and supports examiners in conducting a post-scoring based on the pre-scoring. The main reason for the challenge is an increasing number of students in university teaching and an increasing importance of essay assignments. To overcome these challenges, the system to be designed must provide the examiners with the most accurate pre-scoring possible, which can be comprehended and adapted by the human examiners during the post-scoring process.

Based on the scoring process of essay assignments, we derived three user stories for examiners. First, examiners have to create a level of expectation for the respective essay task before the scoring (U1). This is usually developed as part of the task design and includes possible answers that will be assessed as correct. In addition, the level of expectation should include possible cut scores for individual parts of answers if not all aspects are presented completely and correctly (Joint Committee on Standards for Educational and Psychological Testing 2014). The next step in the scoring process is the actual scoring of the essays based on the level of expectation (U2). Finally, the scoring process must be documented for each essay task and each student (U3) to give the exam taker access to the essay score and the interpretation by the examiner (Joint Committee on Standards for Educational and Psychological Testing 2014). The documentation includes the identification of the correct components of the level of expectation as well as the respective point allocation. In addition, missing aspects are noted.

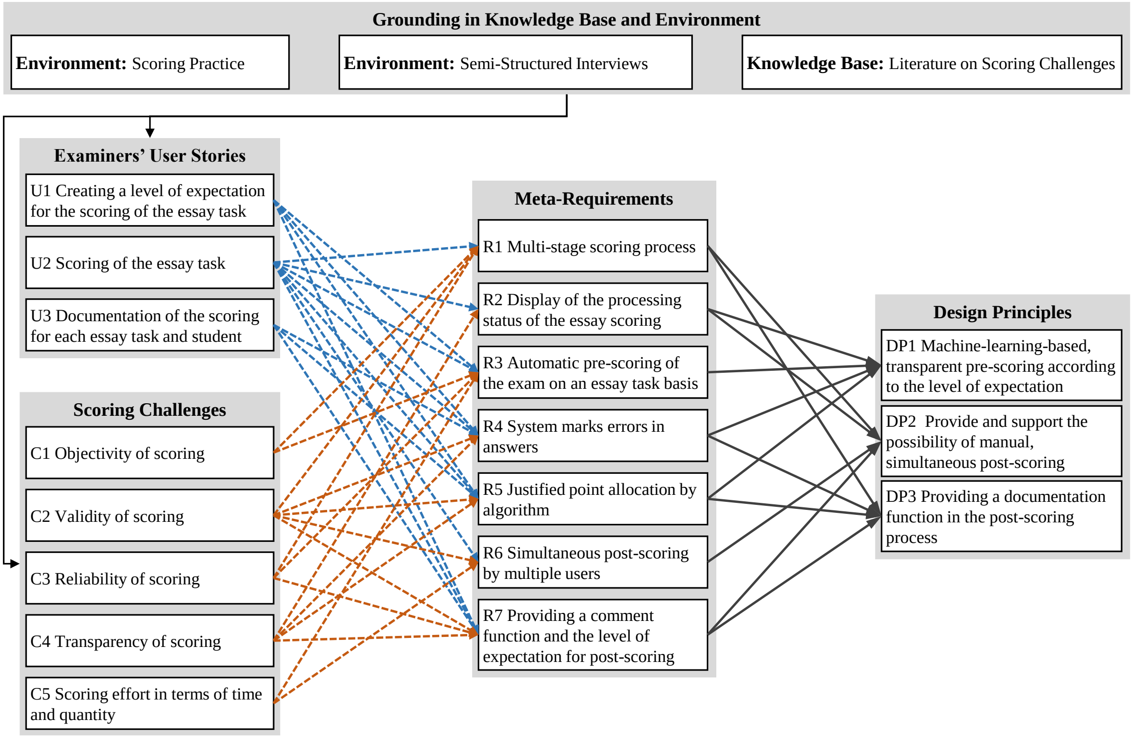

Grounding in Knowledge Base and Environment
Environment: Scoring Practice
Environment: Semi-Structured Interviews
Knowledge Base: Literature on Scoring Challenges
Examiners' User Stories
U1 Creating a level of expectation for the scoring of the essay task
U2 Scoring of the essay task
U3 Documentation of the scoring for each essay task and student
Scoring Challenges
C1 Objectivity of scoring
C2 Validity of scoring
C3 Reliability of scoring
C4 Transparency of scoring
C5 Scoring effort in terms of time and quantity
Meta-Requirements
R1 Multi-stage scoring process
R2 Display of the processing status of the essay scoring
R3 Automatic pre-scoring of the exam on an essay task basis
R4 System marks errors in answers
R5 Justified point allocation by algorithm
R6 Simultaneous post-scoring by multiple users
R7 Providing a comment function and the level of expectation for post-scoring
Design Principles
DP1 Machine-learning-based, transparent pre-scoring according to the level of expectation
DP2 Provide and support the possibility of manual, simultaneous post-scoring
DP3 Providing a documentation function in the post-scoring process

Figure 10. Overview of Requirements and Design Principles

The first three challenges arise from the implementation of the principles of good testing. These include the aspects of objectivity (C1), validity (C2), and reliability (C3) and must also be ensured in the context of scoring (Joint Committee on Standards for Educational and Psychological Testing 2014; Hewlett / Kahl-Andresen 2014). Objectivity (C1) expresses the extent to which a score is derived independently of an examiner's subjective evaluation (Hewlett / Kahl-Andresen 2014). The JOINT COMMITTEE ON STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING (2014) also lists the notion of "fairness in testing" in this context. Thus, certain test characteristics may not be comparably difficult for all subgroups (e.g., defined by disability or language) within an exam. However, within a subgroup and its particular test, there should be a fair, objective assessment. Validity (C2) describes the suitability of the measurement for a given goal. A task is intended to test for the presence of a predetermined knowledge or skill (Hewlett / Kahl-Andresen 2014; Joint Committee on Standards for Educational and Psychological Testing 2014). Accordingly, the scoring is intended to provide an accurate assessment of the knowledge and skill being examined. Reliability (C3) describes the consistent replication of scoring. Under comparable conditions, the scoring of examination results should not be random but should show comparable results (Hewlett / Kahl-Andresen 2014; Joint Committee on Standards for Educational and Psychological Testing 2014). In this context, we also talk about the precision of scoring. For the previously mentioned reasons, it is important that the assessment is presented in a transparent and comprehensible way (C4). The last challenge deals with test economy (Hewlett / Kahl-Andresen 2014). Especially in the case of large numbers of participants, the effort and benefit of the written exam must be taken into account in its design. As the scoring of essays is considered costly, with the costs depending very much on the time needed for correction per essay and the number of examinees, it is rarely used for cost-efficiency reasons (C5).

Deriving Requirements from Environment and Knowledge Base

In the following, the requirements for the scoring system resulting from the user stories and the scoring challenges are derived. A detailed overview of the relationship between the user stories (U1-U3), the scoring challenges (C1-C5) and the requirements (R1-R7) can be found in Figure 10. This figure also contains the relationship between the requirements and the design principals (DP1-DP3) described in Section B.4.4.

The first two requirements address the overall structure of the scoring process. A multi-stage scoring process should be implemented, which includes a machine-learning-based pre- as well as a human post-scoring (R1 based on U1, C1-C3). This is intended to make the scoring process objective and, at the same time, to increase validity through human review. The user should be informed about the current processing status of the scoring at any time (R2 based on U2 and C4) to increase the transparency of the procedure. The following three requirements address the machine-learning-based essay scoring. The system should perform the scoring process automatically and on a task basis (R3 based on U1, U2, C1, C3 and C5). During the scoring, the relevant aspects of the scoring should also be made visible in the answer (R4 based on U1-U3, C2 and C4) and the score assignment should be presented (R5 based on U1-U3, C2 and C4). This should improve the transparency of the automation and thus also increase the accuracy in the post-scoring. The possibility of assigning the results of the machine-learning algorithm to the individual text components allows for better human post-scoring. The last two requirements concern human scoring based on the automatic pre-scoring. A major challenge for examiners is the workload of large amounts of essays that need to be scored. Therefore, in addition to the AI-based support, it should also be possible for several examiners to carry out the post-scoring of the essays at the same time (R6 based on U2, C2 and C5). Furthermore, the level of expectation and a comment function should be implemented (R7 based on U1-U3, C1-C4) in order to improve the transparency and accuracy of the scoring as well as ensure the documentation.

Deriving Design Principles

Based on the seven requirements, we derived three design principles. While the first two design principles deal with automatic pre-scoring and human post-scoring, the last design principle describes the documentation of the scoring results.

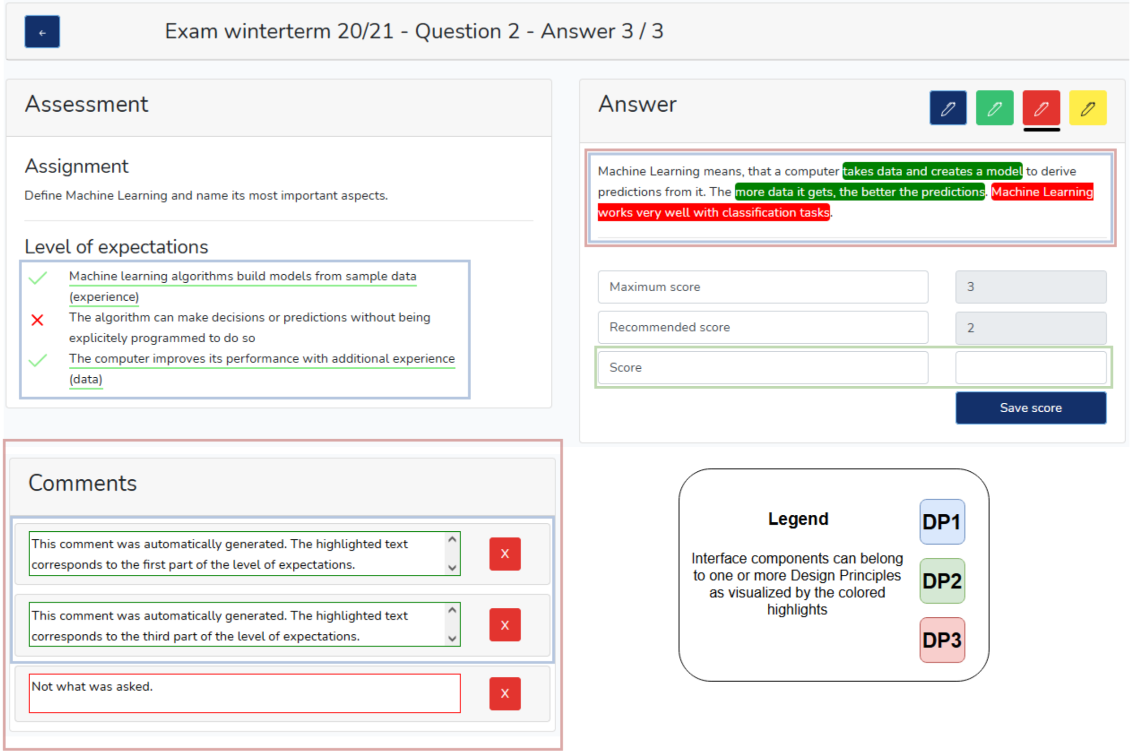
The first design principle defines that the system should perform an automatic machine-learning-based scoring. This scoring should be comprehensible for the examiners and based on the level of expectation assigned to the respective task (DP1 based on R2-R5).

For this purpose, the examiners are given the opportunity to create the essay task and the associated level of expectation in the system. In a next step, the essays of the individual participants are added to the tasks in the system. The essays are then scored using machine learning. The items that are rated as partially or completely correct are highlighted in color. As a second design principle, the system should enable and support the possibility of manual, simultaneous post-scoring of the pre-scored essays (DP2 based on R1, R2, R6 and R7). For this purpose, the human examiners are presented with the level of expectation right next to the essay answer. In addition, it is possible to adjust the scoring from the machine-learning-based pre-scoring. In order to make the results comprehensible, functions for documenting the result of the post-scoring are needed (DP3 based on R1, R4, R5 and R7). Although these also include documentation functions from the AI-based scoring, the final evaluation is carried out by the human examiner. Therefore, the human examiner should have all possibilities to adapt and document the pre-scoring as well.

First Iteration: Designing the User Interface

To visualize the user interface, first, we implemented mockups as a web-based front-end using HTML, JavaScript and CSS. The user interface is responsible for displaying the application's data and for receiving and forwarding user input to the server. Due to the nature of the task, the application is primarily designed for use on desktop PCs. However, the use of the Bootstrap framework allows for basic compatibility with mobile devices.

As shown in Figure 11, the front-end is divided into three sections. The *assessment section* includes the assignment and the associated level of expectation. Components of the level of expectation that have been identified by the scoring mechanism or the human scorer in the respective response are marked with a green tick. Components that were not identified are shown with a red cross. The *answer section* includes the individual answers of the participants. Individual text passages can be highlighted in different colors using a highlighter function. Below the answer, the maximum score, the recommended score by the AI, and the final score are displayed. In the *comments section*, comments can be added to the colored highlights.



Exam winterterm 20/21 - Question 2 - Answer 3 / 3

Assessment

Assignment

Define Machine Learning and name its most important aspects.

Level of expectations

- ✓ Machine learning algorithms build models from sample data (experience)
- ✗ The algorithm can make decisions or predictions without being explicitely programmed to do so
- ✓ The computer improves its performance with additional experience (data)

Answer

Machine Learning means, that a computer takes data and creates a model to derive predictions from it. The more data it gets, the better the predictions. Machine Learning works very well with classification tasks.

Maximum score	3
Recommended score	2
Score	

Comments

- This comment was automatically generated. The highlighted text corresponds to the first part of the level of expectations.
- This comment was automatically generated. The highlighted text corresponds to the third part of the level of expectations.
- Not what was asked.

Legend

Interface components can belong to one or more Design Principles as visualized by the colored highlights

Figure 11. Screenshot of the Manual Scoring Front-end

Evaluating the User Interface

The user interface was evaluated using a survey to determine the user experience and completeness of the application versus user expectation. Respondents with experience in the process of scoring essay questions were surveyed. As an introduction, the exemplary use was shown in a four-minute video to represent the dynamics in the process. The individual steps were explained via audio commentary. During the subsequent questioning, screenshots of the application were shown as a reminder so that the respondents could once again intensively deal with the application.

In the first part of the questionnaire, four questions were asked about each of the sub-areas, namely dashboard, course administration, exam administration, scoring interface, and the analysis of the exam results. The questions asked whether the respondents were satisfied with the respective functionality, whether the elements of the user interface were comprehensible and whether the user guidance was satisfactory. Finally, there was the opportunity to formulate further comments and improvement requests. A total of 25 questionnaires were sent out to people who regularly correct essays, of which 13 fully completed responses were received. The adjustments that were made in the second iteration are listed in Section B.4.4.

The User Experience Questionnaire (UEQ) by LAUGWITZ ET AL. (2008) was used to evaluate the user experience. It enables a general assessment of the user experience of the application at a superordinate level and is suitable as a good supplement to the concrete questions of the first part of the evaluation (Schrepp et al. 2014). In comparison to the

UEQ benchmark (Schrepp et al. 2014; Schrepp 2021), the results show that the application is altogether perceived as good to very good by the users. Figure 12 shows the evaluation divided into the six categories of the UEQ. The results for the user interface in four of the six categories are in the top ten percent of the benchmark. The results in the efficiency and stimulation categories are in the top 25 % of the benchmark data and can thus be rated as good. Due to the small number of participants in the evaluation, the confidence intervals were also considered. These are at least in the range of above-average results for all categories considered. Only in the evaluation of stimulation a large variance and a lower expression can be observed. We assume that this is due to the nature of the activity under consideration.

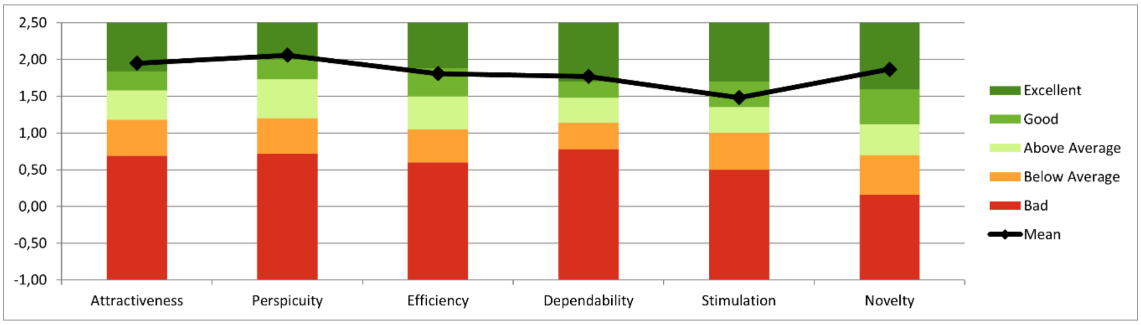


Figure 12. Results of the UEQ (n = 13)

The questionnaire closed with a question about the overall impression of the users, which was predominantly described as positive. Thus, the application was mostly described as helpful. One participant called the application "definitely an improvement because the system does pre-scoring, and you can also quickly and easily approve examiners who are subject to a secure rights concept."

In particular, potential increases in efficiency were attributed to the overview of the scoring process and the provision of the information needed in each case. One participant mentioned that pre-scoring "is definitely an improvement to the current situation [as] many exams currently have to be written digitally or online. This makes it easier to import students' solutions into the system without having to digitize them first." Another participant added that the additional statistical assessment supports the scoring process by saving time otherwise needed for looking at the level of expectation.

Second Iteration: Revising the User Interface and Implementing the Scoring Mechanism

In the second implementation phase, in line with the DSR process, feedback from the evaluation is used to improve the user interface. The potential improvements identified in Section B.4.4, which primarily regard the score assignment in the post-scoring assessment overview, as well as several minor improvements that display additional requested information, have been implemented. Most of the requests addressed the scoring itself.

Regarding DP 2, it was implemented that the scores given, when creating the level of expectation, are also displayed in the scoring interface next to the level of expectation. This should make it easier for examiners to assign (partial) points during post-scoring. In addition, the adjustment of the AI-based pre-scoring has been simplified. Thus, in the post-scoring assessment overview, a part of the level of expectation can now be switched between fulfilled and not fulfilled by clicking on the icon. Hereby, in terms of semi-automatic scoring, the decision of the prototype can be overridden by the human examiner in a quick and uncomplicated way. After switching, the recommendation is adjusted, and the recommended score is automatically corrected by the corresponding amount. To further support the documentation (DP3), it was also implemented that the inserted comments of the multi-level, simultaneous post-scoring are now directly assigned to the authors and that these are identified. In addition, a search function for students and examinations was created. The search is based on the user interface of the exam viewer and serves to make the documentation accessible to the students. Thus, all results of the student with the same student ID number for the selected exam are displayed and the comments can be viewed and changed during and after the scoring process. Additionally, the transfer of a score into a grade by manually specifying a grade delta was implemented. To increase the flexibility of the assignment of grades, the grade deltas can now be assigned manually in addition to the automatic mode. The technical design of the prototype can be divided into three components, the neural network, the application logic, and the user interface.

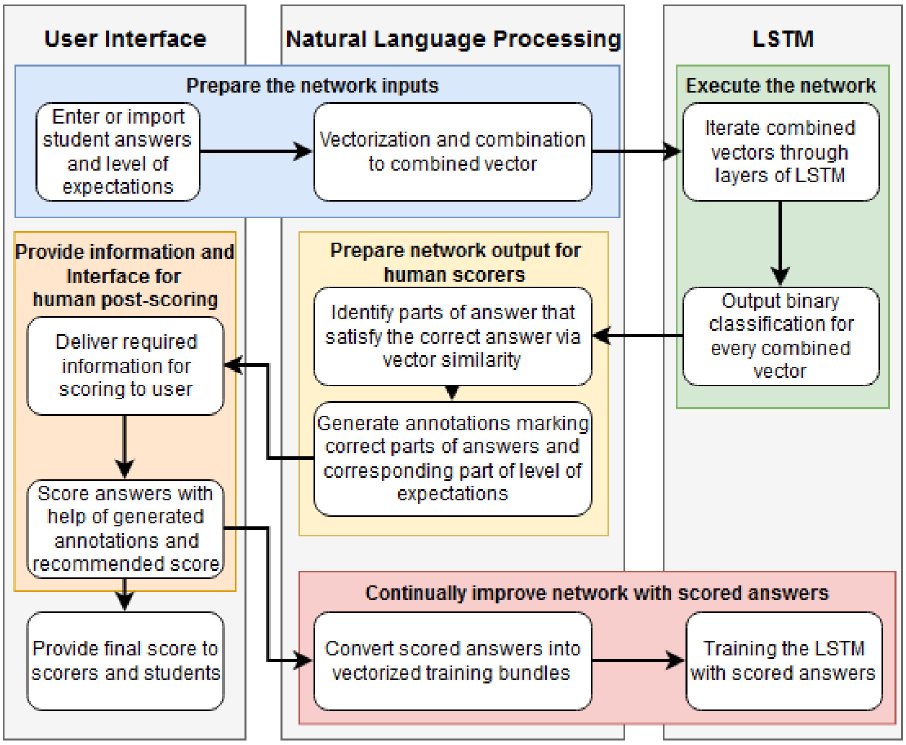

User Interface
Natural Language Processing
LSTM
Prepare the network inputs
Enter or import student answers and level of expectations
Vectorization and combination to combined vector
Execute the network
Iterate combined vectors through layers of LSTM
Output binary classification for every combined vector
Prepare network output for human scorers
Identify parts of answer that satisfy the correct answer via vector similarity
Generate annotations marking correct parts of answers and corresponding part of level of expectations
Provide information and Interface for human post-scoring
Deliver required information for scoring to user
Score answers with help of generated annotations and recommended score
Provide final score to scorers and students
Continually improve network with scored answers
Convert scored answers into vectorized training bundles
Training the LSTM with scored answers

Figure 13. Implementation of the Scoring Mechanism

Furthermore, we implemented the AI-based scoring mechanism as shown in Figure 13. The machine-learning component is responsible for generating the score suggestions of the pre-scoring (DP1) and is separated from the rest of the application logic. A Python server makes the network available to the application via an API. POST requests can be used in the local network to address the neural network and transmit the required data. The PyTorch framework and the Python programming language were used for the implementation. The vectors required for the network were calculated using the spaCy framework, which analyzes the answers to the essay questions and makes the linguistic context and other properties of natural language understandable and usable for the machine-learning component. The evaluation of exam questions in terms of points on a fixed scale poses a classification problem. Therefore, a neural network (Recurrent Neural Network) is used for the machine learning component. The neural network is trained using past exam scorings consisting of an answer and a score (supervised learning). In addition, the algorithm should be improved by the answers scored in the application (reinforcement learning) and be able to transfer the knowledge of previous scorings to new unknown tasks (transfer learning). The specific neural network becomes a kind of Recurrent Neural Network that maintains the order of information and thus understands contextual information better (Huang / Feng 2019). Moreover, due to their recursive nature, they can work well with inputs of different sizes and lengths (Chung et al. 2014). Since essay responses are of variable length, an LSTM is chosen as the network for this application. The implementation is done with the open-source framework PyTorch, as it offers a faster implementation as well as shorter training times than comparable frameworks (Cohen et al. 2018; Simmons / Holliday 2019; Heghedus et al. 2019). A custom word embedding was not used since such an embedding would only represent a known set of words and the generalization to unknown words and topics would be limited.

On the server-side, the application logic is implemented using Laravel. User input is implemented and stored, and database content is retrieved. With the help of the latter, views are created for users.

Evaluating the scoring mechanism

Since the software artifact is only an improvement on the status quo if the semi-automatic scoring is of sufficient quality, several evaluations were conducted. To train the network, annotated data were needed in which the fulfillment of individual parts of the level of expectation was recognizable in an answer. Since no suitable dataset was found, we built on the Hewlett Foundation's Kaggle dataset (Kaggle 2012), which partially satisfies the requirements. The dataset contains ten questions with approximately 17,000 responses. For three of these questions, an assignment of awarded points to individual parts of the level of expectation is possible. Annotation was done retrospectively and by hand. Scoring criteria were given for each question, and two expert point ratings were given for each answer. A total of 500 answers for each of the three questions were annotated and used for training and testing.

10-fold Cross-Validation

We performed a 10-fold cross-validation, using one-tenth of each of the 1,500 responses as test dataset while training the network with the remaining data. A separate network was trained and evaluated for each of the ten combinations. Figure 14 shows the quotas of the training and test data of the networks as well as the mean values of the training and test datasets.

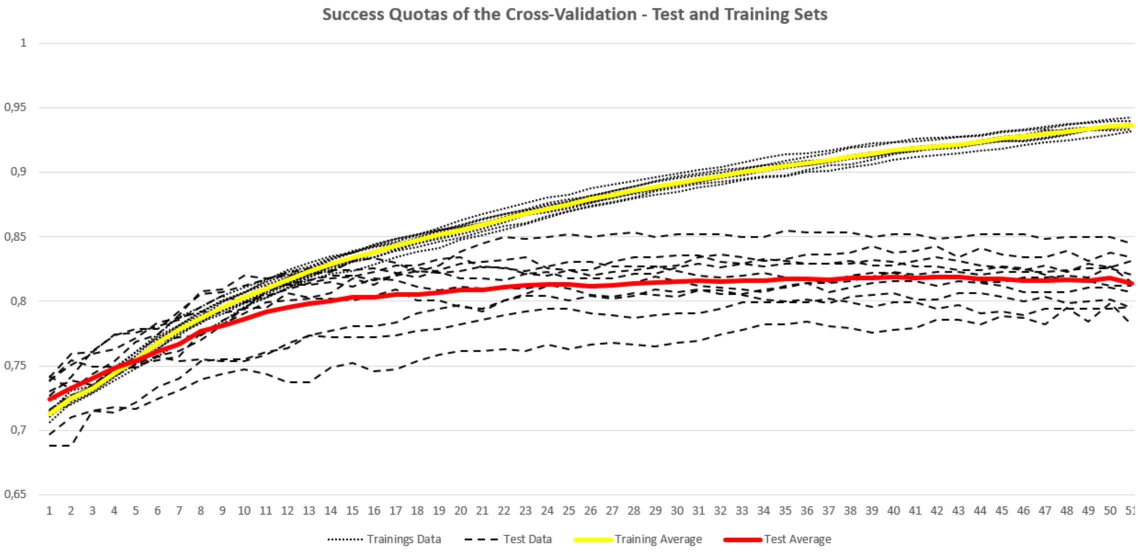


Figure 14. Success Quotas of the Cross-Validation

The results of the training data show that all training quotas are within a very narrow range of maximum two percentage points. For the quotas of the test data, a larger range was shown, with a maximum range of nine percentage points between 76 % and 85 %. Two outliers were observed here. The observation of the remaining test series showed a range of only five percentage points. For nine out of ten test series, a rate of over 80% was observed. As can be seen in the figure, there is a convergence of the mean value at 81 %, which was achieved in most test series between epoch 20 and 23.

Suitability Out-of-Domain Data

Since the trained network will also be used for future unknown essays, its suitability is also evaluated using out-of-domain data. These present a new challenge for the network since both the responses and the level of expectation are unknown. For this purpose, in addition to the existing data, another manually created and comparable dataset was used. This consists of 100 annotated responses to a question with a three-point level of expectation. In this validation, the network achieved a rate of correct assessments of 48 %. The result is due to the fact that 48 % of the answers in the dataset were incorrect, and all answers were rated to be incorrect by the network. Therefore, the generalization capability of the network can be considered to be strongly limited.

Learning Samples Needed to Adapt to Out-of-Domain Data

In order to increase generalizability, the network already trained with the original data was trained several times with the new dataset of 100 annotated responses. In each case, a portion of the data was used as training dataset and the proportion was incrementally increased. Similarly, the network was trained with this training dataset in an incremental number of epochs. Before calculating the quotas of correct estimates of a combination, the network was reset to its original state. For this rate, the network had to estimate the remaining responses not used as training data. The calculations were performed using a five-fold cross-validation. The results show that a minimum of about 30 learning phases, arbitrarily combined of answers and epochs, are necessary before improvement occurs. For a level of 70 %, a minimum of 22 new responses and 19 epochs were required. For a level of 75 %, at least 37 answers and 14 epochs were necessary. The number of available answers was more important than the number of epochs. While the quota rose steadily with a constant number of epochs but more answers, it stagnated conversely from a number of 15 to 20 epochs.

Summarizing and Documenting the Design Knowledge

In the first iteration, the focus was on the front-end and the basic functionalities for examiners. The evaluation results show that the user experience was rated as good to excellent according to the UEQ dimensions. Only minor changes were required, which were implemented in the second iteration. Since these changes only addressed the implementation of DP2 and DP3, but not the design principles themselves, we end the DSR process for these two design principles. The second iteration primarily focused on the implementation and subsequent evaluation of the scoring mechanism. According to DP1, the assessment should be based on a level of expectation and serve as a pre-scoring. Satisfactory results were achieved, particularly in the 10-fold cross-validation. For the use with out-of-domain data only worse results could be observed, so that a training with learning samples was necessary. We were able to show that the machine-learning approach is suitable for scoring essays. The accuracy of the algorithm can be significantly improved by prior training. Since the design of the algorithm was not the first priority in our project, the knowledge gained is sufficient for us to also end the DSR process for the first design principle. The following documentation and communication of the results of the design process is based on the components of the design principles schema according to GREGOR ET AL. (2020). Table 15 shows the derived design principles.

Design Principle Title	Principle of…		
	…machine-learning-based pre-scoring (DP 1)	… manual, simultaneous post-scoring (DP 2)	…documenting the post-scoring process (DP 3)
Aim, and user	To support examiners in the scoring of essays …		
Context	… in exams with a large-scale educational context …		
Mechanism	… provide a machine-learning-based scoring mechanism that is able to generate automated pre-scoring drafts based on transparent evaluation criteria…	… provide an easy-to-use user interface for manual post-scoring that uses the pre-scoring (DP1) as basis to reduce the workload …	… provide a traceable and transparent scoring process by documenting the examiners decisions and the underlying evaluation criteria …
Rationale	… because this can help examiners to reduce the scoring workload while increasing the objectivity of essay scoring.	… because this can help to improve the overall scoring quality, required workload, students' acceptability of exam results, and acts as a manual verification step of the diagnostic quality of the machine-learning-based pre-scoring mechanism.	… because this can increase the transparency of the scoring process, might be mandatory for providing an explainable scoring process, and can be helpful in communicating the scoring results to the students.

Table 15. Documentation of the Design Principles based on Gregor et al. (2020)

In DP1, a machine-learning-based, transparent pre-scoring of essays according to the level of expectation was formulated. Hereby, a more efficient and objective scoring process should be ensured. DP 2 addressed the provision and support of manual, simultaneous post-scoring. The human post-scoring is intended to ensure the diagnostic quality of the whole scoring process. DP3 dealt with documentation in the post-scoring process. This was intended to promote transparency in scoring and is particularly relevant for the communication with students.

4.5 Discussion and conclusion

The goal of the research project was to derive design principles for a semi-automatic scoring system for essay tasks. In particular, the examiners should be supported in mastering the challenges in the execution of the defined user stories described by the scoring process of the essay tasks. The challenges cover the demands placed on the scoring of an essay by different stakeholders. To achieve this goal, the first step was to derive requirements and design principles based thereon. The implementation in our software artifact was done in an iterative process according to the DSR approach of Hevner et al. (2004) and Peffers et al. (2007). The user interface and the functions were implemented and then evaluated. We showed that potential users were largely satisfied with the functionalities and the user interface and that the artifact can provide additional value for the scoring. Additionally, a technical evaluation of the machine-learning algorithm was carried out, since the added value of the artifact only arises if there is an improvement on the manual scoring process. For the technical evaluation, a modified Kaggle dataset was used. In a first 10-fold cross-validation, it could be shown that on average a convergence of 81 % (20 epochs) takes place. For the use with out-of-domain data, no satisfactory results could be achieved at first. A further training was therefore carried out, achieving a success quota of roughly 75 % (30 epochs). In addition to the software artifact, we also contributed design knowledge to the scientific knowledge base. The systematic documentation of this knowledge was done in our last step of the DSR process using the structure of Gregor et al. (2020).

The derived design knowledge can not only be applied to the specific use case, but can also be seen as a generalizable basis for comparable software artifacts with different levels of automation and other (semi-) open task types. Thus, for a deviating level of automation, only the share of decisions to be made manually has to be adapted (Frohm et al. 2008). Furthermore, it could be shown that with the help of the machine-learning component, freely formulated answers can be evaluated correctly to a high degree. This can also be transferred to semi-open answers by adapting the respective design of the level of expectation. Furthermore, the design knowledge about the user case can be used in teaching-learning arrangements where essay tasks are used in the context of diagnostic or formative assessments. In this way, feedback can be given to participants immediately after answering an essay task, thus improving error reflection and the learning process. Due to the possibility to transfer our generated design knowledge, our research does not only provide level 1 DSR contribution but also level 2 DSR contribution (Gregor / Hevner 2013). In addition, we could show that even with a manageable training effort, especially for out-of-domain data, a good result in the pre-scoring could be achieved and examiners can be supported in the scoring process in practice.

The quality of the scoring depends on the volume and quality of the available data. Especially in the search for suitable training data, we have shown that this cannot be taken for granted, even in a data-driven era. Although extensive data are available, there are requirements concerning different key attributes. Since our neural network evaluates the fulfillment of parts of a level of expectations, data must be available that allow a direct correlation between points and parts of the level of expectation. To learn this connection, the network needs data in which it is annotated which part of the level of expectation is fulfilled by the respective answer. In addition, we were able to show in an out-of-domain context that good but only limited use can be made of existing training data for interdisciplinary use. Here, the use of subject-specific networks could be a solution. In addition, the essay task in our scenario had exactly one correct answer. Thus, the prototype is primarily suitable for questions that serve knowledge assessment. Tasks in which a scenario-based subjective evaluation must be carried out also require consideration of the argumentation structures within the decision-making process of the examinees. The added value increases with the number of participants and is probably not suitable for smaller courses. Typical scenarios can be courses in which several hundred students participate or which are repeated identically on a regular basis. Thus, a semi-automatic scoring system can support the execution of the aforementioned user stories especially for large events by reducing the time needed for the scoring and documentation of essay exams. At the same time, it facilitates the fulfillment of the requirements for the scoring of exams by transparently standardizing the scoring and implementing a reduction of human bias. However, the extent to which a system described is actually used depends on other additional factors besides the design principles. For example, potential efficiency benefits can only be realized if both examinees and examiners trust the system and thus use it (Wu et al. 2011).

5 Design of AI-based Scoring Explanation

Explain AI-Based Essay Scorings without XAI - Empirical Investigation of an User-Centered UI Design for AI-Based AES Systems

Explain AI-Decisions in AI-based AES-Systems

Explain AI-Based Essay Scorings without XAI - Empirical Investigation of an User-Centered UI Design for AI-Based AES Systems

Completed Research Full Paper

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

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

Abstract

The lack of understandability of AI-based decisions is increasingly posing trust-related and regulatory problems. This also applies to the educational sector, where AI is a central element of modern automated essay scoring (AES) systems. However, current research on explainable AI primarily focuses on complex technical approaches. These explanations usually show a lack of understandability by the actual users, who often have no knowledge of AI. Based on an experiment with 245 students at a German university, we were able to show that even the basic principles of user interface design can improve understandability and hereby trustworthiness. Thus, the use of visual elements promotes understandability even when only little information is provided. Especially when providing further AI-specific information on the scoring of AES systems, however, it must be considered that in combination with visual elements an information congruency can be observed, leading to a cognitive overload in the worst case.

Keywords

Automated essay scoring, AES, explainable AI, XAI, user interface design, assessment, education.

Introduction

AI is increasingly being used in digital education, improving the individual support of students, and reducing the effort to score essays in exams (Chen et al. 2020). Hereby, essay answers from examinees can be automatically scored based on different criteria using AES systems. In these systems, AI is said to have advantages such as greater objectivity and reliability. However, it also causes new disadvantages. Probably the biggest disadvantage is the lack of transparency. Thus, the decision making of an AI-based system is often compared to a black box (Adadi and Berrada 2018). Particularly in the case of high-stakes exams, which must be justiciable, conflicts of interest often arise between the stakeholders, the examinees, and the examiners. For this reason, the principles of FATE (Fairness, Accountability, Transparency, and Ethics in education) have been proclaimed for the use of AI in education (Khosravi et al. 2022). These also coincide with the EU guidelines on the traceability of AI and, as explainable AI (XAI), represent an increasingly addressed area in AI research (Sartor 2020). Here, users are provided with additional information on how the AI achieved its result. Through this, decision makers should be able to intervene in the decision if something goes wrong (Adadi and Berrada 2018). Another important aspect of explainability is the impact on perceived trustworthiness, which influences trust and thus significantly the acceptance and use of such systems (Mayer et al. 1995; Siau and Wang 2018; Wu et al. 2011). However, whether an explanation is perceived as helpful differs between stakeholders. For example, a technical staff member tasked with programming the AI has different prerequisites than an examinee and pursues different goals through explainability (Barredo Arrieta et al. 2020; Gunning et al. 2019). While XAI consideration from the perspective of algorithms has been the subject of extensive research, there has been little actual user-centric consideration in the education domain. The following research question will be addressed: *How do explanations in an AI-based AES system have to be presented to be evaluated as understandable by examinees from the perspective of trustworthiness?* Here, we focus on the local interpretability of individual scoring results and how these must be designed in an AES system in a target group-centric way, independent of the algorithms used. For this purpose, different user interface design elements were

Twenty-ninth Americas Conference on Information Systems, Panama, 2023 1

Abstract: *The lack of understandability of AI-based decisions is increasingly posing trust-related and regulatory problems. This also applies to the educational sector, where AI is a central element of modern automated essay scoring (AES) systems. However, current research on explainable AI primarily focuses on complex technical approaches. These explanations usually show a lack of understandability by the actual users, who often have no knowledge of AI. Based on an experiment with 245 students at a German university, we were able to show that even the basic principles of user interface design can improve understandability and hereby trustworthiness. Thus, the use of visual elements promotes understandability even when only little information is provided. Especially when providing further AI-specific information on the scoring of AES systems, however, it must be considered that in combination with visual elements an information congruency can be observed, leading to a cognitive overload in the worst case.*

Keywords: *Automated essay scoring, AES, explainable AI, XAI, user interface design, assessment, education*

Citation: (Hartmann / Hobert 2023a) Hartmann, P.; Hobert, S.: *Explain AI-Based Essay Scorings without XAI - Empirical Investigation of an User-Centered UI Design for AI-Based AES Systems*. In: *Proceedings of the 29th Americas Conference on Information Systems.* Panama City, Panama. 2023. pp. 1-10.

5.1 Introduction

AI is increasingly being used in digital education, improving the individual support of students, and reducing the effort to score essays in exams (Chen et al. 2020). Hereby, essay answers from examinees can be automatically scored based on different criteria using AES systems. In these systems, AI is said to have advantages such as greater objectivity and reliability. However, it also causes new disadvantages. Probably the biggest disadvantage is the lack of transparency. Thus, the decision making of an AI-based system is often compared to a black box (Adadi / Berrada 2018). Particularly in the case of high stakes exams, which must be justiciable, conflicts of interest often arise between the stakeholders, the examinees, and the examiners. For this reason, the principles of FATE (Fairness, Accountability, Transparency, and Ethics in education) have been proclaimed for the use of AI in education (Khosravi et al. 2022). These also coincide with the EU guidelines on the traceability of AI and, as explainable AI (XAI), represent an increasingly addressed area in AI research (Sartor 2020). Here, users are provided with additional information on how the AI achieved its result. Through this, decision makers should be able to intervene in the decision if something goes wrong (Adadi / Berrada 2018). Another important aspect of explainability is the impact on perceived trustworthiness, which influences trust and thus significantly the acceptance and use of such systems (Mayer et al. 1995; Siau / Wang 2018; Wu et al. 2011). However, whether an explanation is perceived as helpful differs between stakeholders. For example, a technical staff member tasked with programming the AI has different prerequisites than an examinee and pursues different goals through explainability (Barredo Arrieta et al. 2020; Gunning et al. 2019). While XAI consideration from the perspective of algorithms has been the subject of extensive research, there has been little actual user-centric consideration in the education domain. The following research question will be addressed: *How do explanations in an AI-based AES system have to be presented to be evaluated as understandable by examinees from the perspective of trustworthiness?* Here, we focus on the local interpretability of individual scoring results and how these must be designed in an AES system in a target group-centric way, independent of the algorithms used. For this purpose, different user interface design elements were evaluated concerning the three aspects of trustworthiness, namely performance, purpose, and process, using a questionnaire study.

5.2 Related Research and Hypotheses Development

Explainable AI (in AES)

Explainable AI is a broad term that is difficult to define. Overall, the focus is not on the AI, but on the effort to make AI-based decisions transparent and trustworthy. It is hence rather the need for explaining or justifying a certain result, instead of an exact description of the inner decision-making process of the AI (Adadi / Berrada 2018). Therefore, explainability is strongly related to understandability (Adadi / Berrada 2018;

Haque et al. 2023). While explainability is concerned with what and how the AI-based system explains something, understandability is concerned with how the user processes the information. The challenge is to communicate a complex AI-based decision to a human as simply as possible to achieve a high explainability value. Research has shown that not all decisions need to be fully explained, but that usability and practical interpretability, for example through a mix of textual and visual explanations, are relevant factors (Abdul, Ashraf et al. 2018; Adadi / Berrada 2018). A distinction can be made between different types of AI users. In particular, end users who have no in-depth knowledge of how an AI works but have to work with AI-based systems have a special role to play here. So, user-centered explanations must be adapted to the corresponding user types. This can improve the understandability and thus the trustworthiness towards the AI-based system (Barredo Arrieta et al. 2020; Borrego-Díaz / Galán-Páez 2022; Ribera / Lapedriza 2019). In the field of AI-based AES systems, there are only a few considerations as to how these aspects can be addressed. Most of the studies conducted in this area only consider global explanations at the model level and not at the user interface level. FERRARA / QUNBAR (2022) address the limited evidence and validity arguments for AES scores in their study. Many AES systems still fail because of the trade-off between high scoring accuracy and high explainability, respectively understandability. This limitation makes it necessary to reflect on the currently used validity arguments and fairness to examinees. According to the authors, one possible solution is to make the scores at least partially interpretable by the examinees. Therefore, scores should consist of an easily explainable component to create transparency and a less explainable and opaque component to increase scoring accuracy. (Kumar / Boulanger 2020) focus on the prediction of the quality of the writing style of essays in their study. In this context, the factors that influenced the scoring are also evaluated. The results show that SHAP and model-agnostic implementations may have comparable accuracy. According to the authors, these results show the suitability of XAI implementation in AI-based AES systems, opening further application possibilities in the field of learning analytics. A more user-centered approach was taken by SCHLIPPE ET AL. (2023) who evaluated specific user interface design elements of an automatic short answer grading system concerning aspects such as trust, informative content, or comprehensibility. The participants were a small group of potential examiners. The presentation of points and matching positions was identified as the best approach.

Trustworthiness of AI-Based AES Systems

Trustworthiness is an important factor in trust research, describing the perception of the trustee’s suitability to perform an assigned task to the satisfaction of the trustor, even without monitoring (Mayer et al. 1995; Siau / Wang 2018). The influence of the factors ability (performance), benevolence (purpose), and integrity (process) on trustworthiness determined by MAYER ET AL. (1995) in interpersonal trust, was adapted for trustworthiness towards automatization and AI (Lee / See 2004; Siau / Wang 2018). In the following, an AI-based AES system is considered that is intended to score essays in high-stakes exams. For examinees, the scoring should be easily comprehensible. Performance describes the expected skills and competencies of the trustee within a specific domain (Mayer et al. 1995). According to Lee / See (2004), performance is always tied to a specific task and situation and includes characteristics such as reliability, predictability, and capability. Thus, it provides information about what the trustee does. Higher perceived performance leads to higher trustworthiness. In our case, examinees expect the AI-based system to score reliably, non-randomly, and correctly. It can be assumed that a higher understandability by providing additional explanation increases the perceived performance of the AI-based AES system and thereby the trustworthiness. The factor process describes the adherence to predefined common principles that are intended to promote the trustee's actions with integrity. The principles include aspects such as information about the trustee from others or the belief in the trustee's sense of justice. If common principles are insufficiently present from the trustor's point of view, the trustee is perceived as lacking integrity (Mayer et al. 1995). The process thus provides information on how the trustee operates. Higher evaluation of integrity leads to higher trustworthiness. In our case, we assume the AI-based AES system to have the same interests as the examinees, which is the correct essay scoring in high-stakes exams. Here, (partial) points are to be awarded for (partially) correct answers. Incorrect answers should be scored as incorrect. Therefore, we assume that a higher understandability by providing additional information about the scoring process increases the perceived process and thereby leads to higher trustworthiness. In the field of automation and AI, purpose refers to the extent to which the automation is used in the interest of the designer (Lee / See 2004). The purpose describes why the automation was developed. In the case of automation and IS, the positions of the designer and the users mostly diverge. The higher the rating of the purpose of the deployment, the higher the trustworthiness towards the system. In our case, we assume that there is no contact between the AI developer and the examinees. However, it can be assumed that neither developer nor examinees have a personal advantage from false scoring. The basic principle of assessment is a correct and fair classification of individual knowledge and competencies on an abstract level and in relation to other examinees (Tierney et al. 2011). Thus, it can be assumed that a higher understandability by providing additional information increases the perceived purpose and thereby trustworthiness.

To answer our research question stated in the introduction, the following hypotheses were derived from the expected relationships formulated in the previous sections. We assume that more comprehensive textual information positively affects the trustworthiness-related variables and thus expresses increasing understandability. We also expect a positive moderating effect of visual information as this emphasizes existing textual information. However, we expect an exception to this moderation for a simultaneously high level of textual and visual information. Thus, we expect a negative effect due to cognitive overload if both types of information are provided too extensively (Sweller 2003).

H1: *More extensive textual information on the scoring by an AI-based AES system increases trustworthiness via increased (a) perceived performance / (b) perceived process / (c) perceived purpose.*

H2: *The effects described in H1 (a / b / c) are (a) reinforced by additional visual information for a medium level of textual information / (b) weakened for a high level of textual information.*

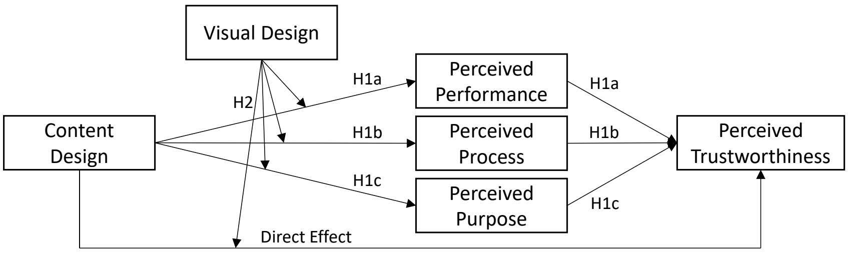


Figure 15. Conceptual Research Model

5.3 Research Design

First, a pre-study of an exemplary research approach of XAI in education was conducted to evaluate the understandability of the scoring for examinees. Based on these results, design aspects were derived and implemented in different user interface designs to evaluate their suitability regarding the trustworthiness-related factors as well as trustworthiness and understandability themselves. In this context, a moderated parallel mediation analysis was carried out for the derived design components and their characteristics.

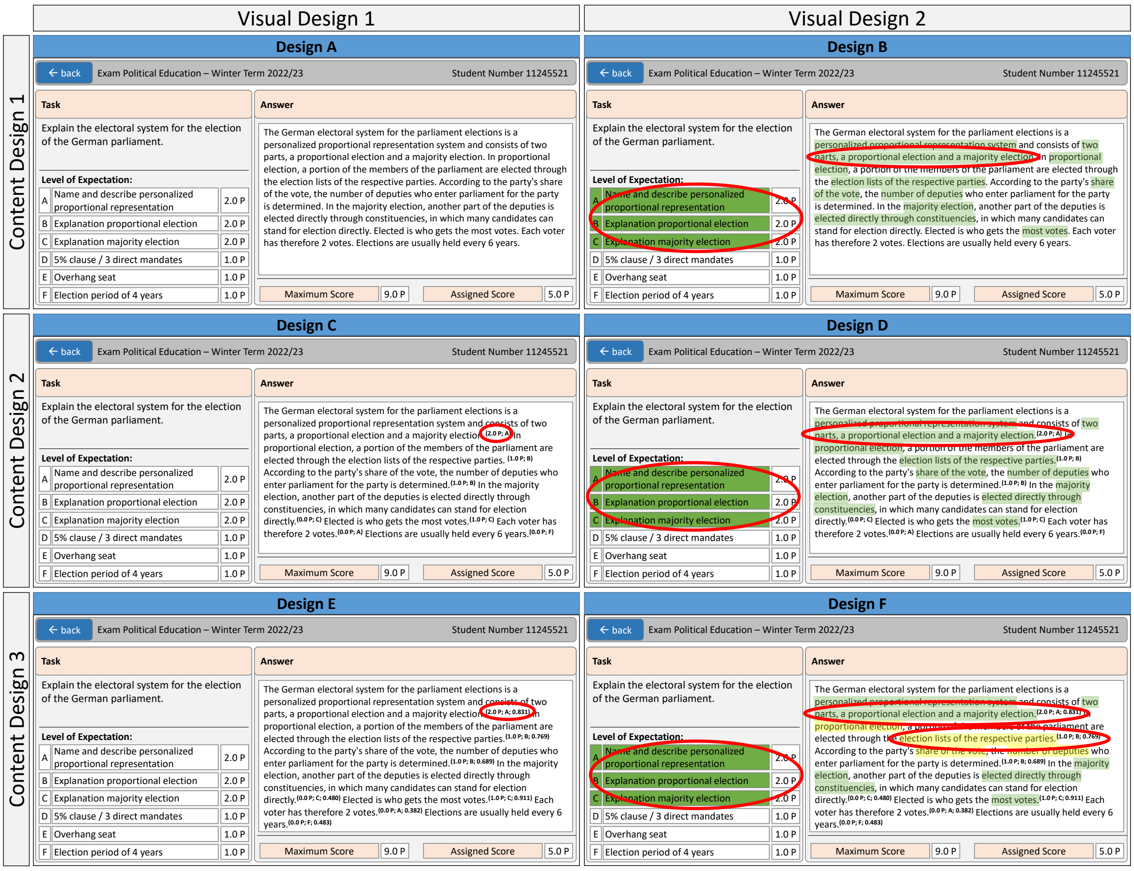
Pre-Study

To estimate the expected level of explanation required from examinees, a pre-study with 15 students was conducted. In this pre-study, the initial user interface Design A as shown in Figure 16, which serves as the starting point for the main study, was supplemented by a figure of the SHAP values since this kind of presentation is currently receiving a lot of attention in research (Adnan et al. 2022; Jang et al. 2022). This modified user interface design was used with the questionnaire of the main study. The participants rated the explanation provided by the SHAP values as not understandable (MD = 2.69; SD = 0.99) on a scale from 1 (low) to 6 (high). The same applies to the trustworthiness towards the AI-based AES system, which was rated as low (MD = 3.23; SD = 1.31). Possible improvements named by the participants included highlighting correct and incorrect statements or making them recognizable and assigning partial points. Therefore, it can be assumed that the XAI-based explanations are not suitable for examinees. Based on these results, an attempt was made to create an easier-to-understand explanation for the scorings using simple, already used design approaches from user interface design.

User Interface Design

There are many different design recommendations for user interface design. In the following, we will use the 10 principles for interaction design by NIELSEN (1994), which address the interaction between humans and computers and are therefore considered as suitable when focusing on XAI. We focus on the 3 principles (Principle 2, 4, and 8) dealing with the presentation and understandability of information. Principle 2 ("Match between system and real world") states that the system should speak the language of the user and orient itself to real conventions known to the user. This should promote familiarity and hereby increase the understandability of the information. Principle 4 ("Consistency and standards") states that the same content should be presented in the same way and should mean the same thing. Thus, related information should be presented in the same place in the user interface to reduce the cognitive load and improve information processing as well as understandability. Principle 8 ("Aesthetics and minimalist design") states that user interfaces should not display irrelevant information. Hence, all presented information competes during perception, reducing the visibility of relevant information and increasing the cognitive load. This disturbs information processing and understandability. These design principles are implemented in Figure 16 via a content and a visual design component, whose specific modifications are circled in red as an example.

The content design comprises three levels of descriptive information about the AI based score. The first level (Content Design 1; Designs A and B) includes the total score assigned by the AI and the level of expectation. In the following, Content Design 1 represents the initial content design. On the second level (Content Design 2; Designs C and D), the respective score per sentence and the assignment to the respective part of the level of expectation are additionally shown directly within the evaluated answer. The third level (Content Design 3; Designs E and F) includes the second level as well as an additional indication of the similarities to the respective level of expectation. The chosen implementation of the content design addresses principle 2, the match between the system and the real world. Thus, in human scoring, points can be assigned aggregated for a task, but also within a task. Here, correct statements are scored with partial points. Content Design 3 additionally includes a central numerical value in the AI scoring since it has a direct influence on the allocation of points. Principles 4 and 8 are mapped in the system using the same representation. The information is placed close to the respective statement and kept minimalistic in order not to additionally increase the cognitive load.



Visual Design 1 | Visual Design 2

Content Design 1 | Content Design 2 | Content Design 3

Design A

← back Exam Political Education – Winter Term 2022/23 Student Number 11245521

Task

Explain the electoral system for the election of the German parliament.

Level of Expectation:

A	Name and describe personalized proportional representation	2.0 P
B	Explanation proportional election	2.0 P
C	Explanation majority election	2.0 P
D	5% clause / 3 direct mandates	1.0 P
E	Overhang seat	1.0 P
F	Election period of 4 years	1.0 P

Answer

The German electoral system for the parliament elections is a personalized proportional representation system and consists of two parts, a proportional election and a majority election. In proportional election, a portion of the members of the parliament are elected through the election lists of the respective parties. According to the party's share of the vote, the number of deputies who enter parliament for the party is determined. In the majority election, another part of the deputies is elected directly through constituencies, in which many candidates can stand for election directly. Elected is who gets the most votes. Each voter has therefore 2 votes. Elections are usually held every 6 years.

Maximum Score 9.0 P Assigned Score 5.0 P

Design B

← back Exam Political Education – Winter Term 2022/23 Student Number 11245521

Task

Explain the electoral system for the election of the German parliament.

Level of Expectation:

A	Name and describe personalized proportional representation	2.0 P
B	Explanation proportional election	2.0 P
C	Explanation majority election	2.0 P
D	5% clause / 3 direct mandates	1.0 P
E	Overhang seat	1.0 P
F	Election period of 4 years	1.0 P

Answer

The German electoral system for the parliament elections is a personalized proportional representation system and consists of two parts, a proportional election and a majority election. In proportional election, a portion of the members of the parliament are elected through the election lists of the respective parties. According to the party's share of the vote, the number of deputies who enter parliament for the party is determined. In the majority election, another part of the deputies is elected directly through constituencies, in which many candidates can stand for election directly. Elected is who gets the most votes. Each voter has therefore 2 votes. Elections are usually held every 6 years.

Maximum Score 9.0 P Assigned Score 5.0 P

Design C

← back Exam Political Education – Winter Term 2022/23 Student Number 11245521

Task

Explain the electoral system for the election of the German parliament.

Level of Expectation:

A	Name and describe personalized proportional representation	2.0 P
B	Explanation proportional election	2.0 P
C	Explanation majority election	2.0 P
D	5% clause / 3 direct mandates	1.0 P
E	Overhang seat	1.0 P
F	Election period of 4 years	1.0 P

Answer

The German electoral system for the parliament elections is a personalized proportional representation system and consists of two parts, a proportional election and a majority election.(2.0 P; A) In proportional election, a portion of the members of the parliament are elected through the election lists of the respective parties. (1.0 P; B) According to the party's share of the vote, the number of deputies who enter parliament for the party is determined.(1.0 P; B) In the majority election, another part of the deputies is elected directly through constituencies, in which many candidates can stand for election directly.(0.0 P; C) Elected is who gets the most votes.(1.0 P; C) Each voter has therefore 2 votes.(0.0 P; A) Elections are usually held every 6 years.(0.0 P; F)

Maximum Score 9.0 P Assigned Score 5.0 P

Design D

← back Exam Political Education – Winter Term 2022/23 Student Number 11245521

Task

Explain the electoral system for the election of the German parliament.

Level of Expectation:

A	Name and describe personalized proportional representation	2.0 P
B	Explanation proportional election	2.0 P
C	Explanation majority election	2.0 P
D	5% clause / 3 direct mandates	1.0 P
E	Overhang seat	1.0 P
F	Election period of 4 years	1.0 P

Answer

The German electoral system for the parliament elections is a personalized proportional representation system and consists of two parts, a proportional election and a majority election.(2.0 P; A) In proportional election, a portion of the members of the parliament are elected through the election lists of the respective parties.(1.0 P; B) According to the party's share of the vote, the number of deputies who enter parliament for the party is determined.(1.0 P; B) In the majority election, another part of the deputies is elected directly through constituencies, in which many candidates can stand for election directly.(0.0 P; C) Elected is who gets the most votes.(1.0 P; C) Each voter has therefore 2 votes.(0.0 P; A) Elections are usually held every 6 years.(0.0 P; F)

Maximum Score 9.0 P Assigned Score 5.0 P

Design E

← back Exam Political Education – Winter Term 2022/23 Student Number 11245521

Task

Explain the electoral system for the election of the German parliament.

Level of Expectation:

A	Name and describe personalized proportional representation	2.0 P
B	Explanation proportional election	2.0 P
C	Explanation majority election	2.0 P
D	5% clause / 3 direct mandates	1.0 P
E	Overhang seat	1.0 P
F	Election period of 4 years	1.0 P

Answer

The German electoral system for the parliament elections is a personalized proportional representation system and consists of two parts, a proportional election and a majority election.(2.0 P; A; 0.831) In proportional election, a portion of the members of the parliament are elected through the election lists of the respective parties.(1.0 P; B; 0.769) According to the party's share of the vote, the number of deputies who enter parliament for the party is determined.(1.0 P; B; 0.689) In the majority election, another part of the deputies is elected directly through constituencies, in which many candidates can stand for election directly.(0.0 P; C; 0.480) Elected is who gets the most votes.(1.0 P; C; 0.911) Each voter has therefore 2 votes.(0.0 P; A; 0.382) Elections are usually held every 6 years.(0.0 P; F; 0.483)

Maximum Score 9.0 P Assigned Score 5.0 P

Design F

← back Exam Political Education – Winter Term 2022/23 Student Number 11245521

Task

Explain the electoral system for the election of the German parliament.

Level of Expectation:

A	Name and describe personalized proportional representation	2.0 P
B	Explanation proportional election	2.0 P
C	Explanation majority election	2.0 P
D	5% clause / 3 direct mandates	1.0 P
E	Overhang seat	1.0 P
F	Election period of 4 years	1.0 P

Answer

The German electoral system for the parliament elections is a personalized proportional representation system and consists of two parts, a proportional election and a majority election.(2.0 P; A; 0.831) In proportional election, a portion of the members of the parliament are elected through the election lists of the respective parties.(1.0 P; B; 0.769) According to the party's share of the vote, the number of deputies who enter parliament for the party is determined.(1.0 P; B; 0.689) In the majority election, another part of the deputies is elected directly through constituencies, in which many candidates can stand for election directly.(0.0 P; C; 0.480) Elected is who gets the most votes.(1.0 P; C; 0.911) Each voter has therefore 2 votes.(0.0 P; A; 0.382) Elections are usually held every 6 years.(0.0 P; F; 0.483)

Maximum Score 9.0 P Assigned Score 5.0 P

Figure 16. User Interface Designs

The visual design includes two levels of visual highlighting of the given information. The first level (Visual Design 1; Designs A, C, and E) does not include any color highlighting. In the following, Visual Design 1 represents the initial visual design. The second level (Visual Design 2; Designs B, D; and F) highlights the (partially) correctly answered parts of the level of expectations in green as well as the parts of the answer directly in the text. In Designs B and D, (partially) correct parts of the examinee's answer are highlighted in green. In Design F, the highlighting mentioned above is done in green for high similarities and in yellow for medium similarities. The highlighting corresponds to principle 2 since this type of supporting the understandability of the scoring of essays is frequently used in practice. Thus, correct or incorrect statements are highlighted to make the evaluation more accessible. It can be assumed that highlighting the (partially) correct parts of the answer and level of expectation is automatically accompanied by a (partially) missing match with components of the level of expectation for unmarked parts of the answers. By using a maximum of 2 colors for highlighting, we also aim to address principles 4 and 8, which require consistent use and simple presentation of information to reduce cognitive load (Sweller 2003).

Questionnaire

The questionnaire was divided into three parts. Parts 1 and 2 covered the representations and questions about the user interfaces. The participants were asked about their level of agreement with pre-formulated statements using a 6-point Likert scale (completely disagree (1) to completely agree (6)). Part 3 comprised the demographic questions. In total, the questionnaire included 34 statements and questions. In part 1, the initial user interface design (Design A) was presented and its components were explained using a guided tour. The design was shown to all participants of all groups and served as a reference point for the survey to create a uniform understanding of the AI-based AES system. Afterward, the participants were asked about the previously mentioned factors perceived performance, perceived process, and perceived purpose. In addition, the perceived trustworthiness towards as well as the understandability of the AI-based AES system was surveyed. The items used are based on existing items, which have been closely adapted to the current context (Li et al. 2008; Madsen / Gregor 2000) and are shown in Table 16. In part 2, each participant was randomly shown one of the modified user interface designs (Designs B to F) presented in Figure 16. The respective adjustments according to the assigned design were explained. Subsequently, the same questions as in part 1 were asked regarding the modified user interface. Furthermore, we asked for possible user interface adjustments to improve understandability. In part 3, demographic information, e.g., age, gender, faculty, and technology affinity, was collected from all participants. The technology affinity was examined by the ATI Short Scale (Wessel et al. 2019).

Data Collection and Pre-Processing

The questionnaire was sent to students at a German university. No weighting according to demographic characteristics was applied. Participation was anonymous and voluntary. Since vouchers were raffled among all participants, control questions were used. The final sample size was 245 participants, of which 46.12 % were female and 52.24 % were male. About 1.00 % stated their gender as diverse. Age ranged from 18 to 34 years (MD = 22.31; SD = 2.81).

The participants showed a slightly positive technology affinity according to the ATI score (MD = 3.78; SD = 1.09) and no significant differences could be observed regarding the subgroups. Since we drew on existing and scientifically tested items, the fit of the model with our collected data was assessed using a confirmatory factor analysis. The items used and the associated Cronbach’s Alpha (CA) values for the identified factors are listed in Table 16. The sample has a KMO-value of 0.915 and can be considered suitable for factor analysis (Hair et al. 2018; Kaiser 1974).

		Through the display of the scoring results provided by the AI-based AES, I know that the system...
Factor	**Perceived Performance** (CA = 0.824)	... has a sound knowledge of exam scoring.
		... scores just as well as a highly competent person.
		... correctly evaluates the exam answers I submitted.
	Perceived Process (CA = 0.891)	... reliably scores the exam answers.
		... evaluates the exam answers without error.
		... evaluates the exam answers without contradiction.
	Perceived Purpose (CA = 0.830)	...will be used in my best interest.
		... looks after my interests and not only those of the examiner.
		... ensures a fair scoring of the performance of the examinee.

Table 16. Reliability Coefficients of the Factors and Items Used

5.4 Results

Descriptive Analysis

	Design A	Design B	Design C	Design D	Design E	Design F
	Content Design 1		*Content Design 2*		*Content Design 3*	
	Visual Design 1	*Visual Design 2*	*Visual Design 1*	*Visual Design 2*	*Visual Design 1*	*Visual Design 2*
Perceived Performance	MD = 3.36 SD = 1.24	MD = 3.84 SD = 1.31	MD = 3.04 SD = 0.96	MD = 3.80 SD = 1.19	MD = 3.73 SD = 1.17	MD = 3.33 SD = 1.23
Perceived Process	MD = 3.31 SD =1.33	MD = 4.03 SD = 1.35	MD = 3.28 SD = 1.06	MD = 3.89 SD = 1.21	MD = 3.61 SD = 1.28	MD = 3.41 SD = 1.25
Perceived Purpose	MD = 3.65 SD = 1.17	MD = 3.97 SD = 1.23	MD = 3.60 SD = 0.93	MD = 3.86 SD = 1.02	MD = 3.69 SD = 1.04	MD = 3.83 SD = 1.04
Perceived Trustworthiness	MD = 3.53 SD = 1.51	MD = 4.02 SD = 1.41	MD = 3.43 SD = 1.47	MD = 3.91 SD = 1.20	MD = 3.69 SD = 1.28	MD = 3.58 SD = 1.49
Perceived Understandability	MD = 3.37 SD = 1.48	MD = 3.94 SD = 1.30	MD = 3.61 SD = 1.19	MD = 4.19 SD = 1.56	MD = 4.00 SD = 1.40	MD = 3.56 SD = 1.33

Table 17. Descriptive Results of Trustworthiness-Related Factors

Regarding the 6 designs considered, the following values were obtained. For the initial design (Design A), values between 3.31 and 3.65 were observed for all variables considered. Understandability was rated lowest with 3.37 compared to the other designs. For Design B, values between 3.84 and 4.03 were observed, with understandability being rated at 3.94. Compared to Design A, perceived performance, perceived process as well as perceived trustworthiness and understandability were rated significantly higher ($p < 0.05$). For Design C the variables were rated between 3.04 and 3.61 with understandability having the highest value. Compared to Design A, no significant differences between all variables were observed. The evaluation of Design D showed values between 3.80 and 3.89 for the three trustworthiness-related factors, with perceived trustworthiness itself being rated at 3.91. Understandability was rated as high with a value of 4.19, which is the highest mean value across all designs. Compared to Design A, significantly higher ($p < 0.05$) values were observed for the perceived performance, the perceived process, and the understandability. Trustworthiness was also stated as higher than in Design A ($p < 0.10$). For Design E, all trustworthiness-related values were rated between 3.61 and 3.73, while understandability was rated at 4.00. Significantly higher values ($p < 0.05$) compared to

Design A were observed for perceived performance and understandability. The results for Design F were at the same level as for Design A. The trustworthiness-related factors were rated between 3.33 and 3.83, with understandability being rated at 3.56. The detailed results are shown in Table 17.

Statistical Analysis

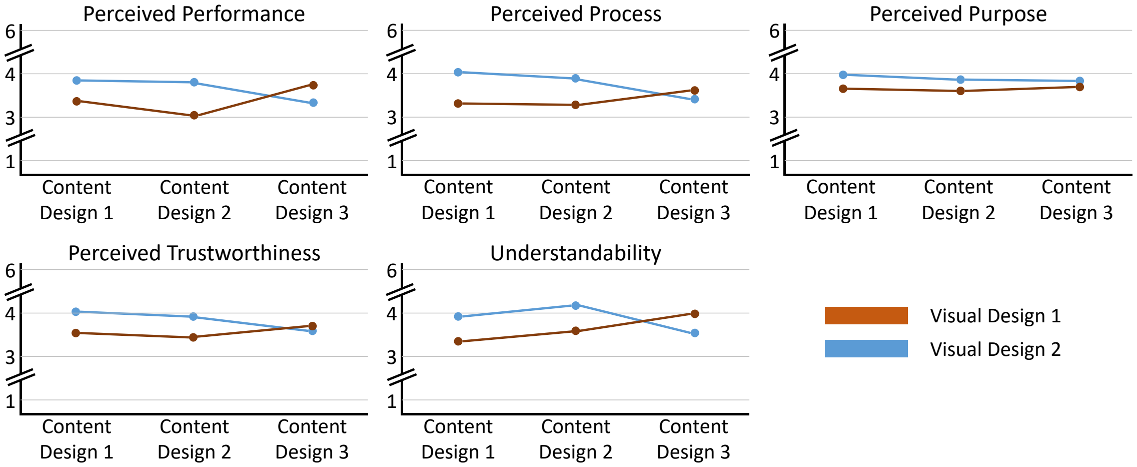


Figure 17. Mean Plot

To test the moderation effect of the Visual Design, we first performed a univariate ANOVA for the three mediators perceived performance, perceived process, and perceived purpose as well as the dependent variable perceived trustworthiness. As seen in Table 18 and Figure 17, no significant main effect of the Content Design was observed, meaning that increasing additional textual information has no influence on the factors considered. No significant differences in the evaluation of the mediators and dependent variable were found between Content Designs 1 and 2, with the Visual Design not having a moderating role. Between Content Design 3 and Content Design 1 (respectively 2), a significant interaction effect with the Visual Design was observed for perceived performance, perceived process, and understandability. This means that the evaluation of the Content Design depends on the Visual Design. A more detailed explanation is given in the following mediation analysis.

F (2, 239)	Perceived Performance	Perceived Process	Perceived Purpose	Perceived Trustworthiness	Understandability
Content (C)	0.656	0.526	0.164	0.319	0.950
Visual (V)	4.552***	7.319***	3.893 **	3.383*	2.365
C x V	6.356***	4.275**	0.220	1.574	4.585 **
p-value					***<0.01; **<0.05; *<0.10

Table 18. ANOVA Results

We tested our conceptual model (see Figure 15) using a moderated parallel mediation analysis with 10.000 bootstrap samples. The following results were interpreted in relation to the initial Design A (Content Design 1; Visual Design 1). The detailed results are shown in Table 4. Overall, we were unable to demonstrate any direct effect of the Content Design on the perceived trustworthiness. Three mediation effects were observed. First, the switch to Design C (Content Design 2; Visual Design 1) had a significant effect on the perceived trustworthiness via the perceived performance. Here, additional information known from paper-based scorings had an indirect negative effect on the perceived trustworthiness. Second, the switch to Design E (Content Design 3; Visual Design 1) had a significant positive effect on the perceived trustworthiness via the perceived performance. Thus, the use of AI-specific textual information, which goes beyond the known information of paper-based scorings, increased the perceived trustworthiness. Third, the switch to Design F (Content Design 3; Visual Design 2) had a significant negative effect on the perceived trustworthiness via the perceived process if further visual support is used. The use of extensive information and additional highlighting reduces the perceived trustworthiness. Furthermore, the positive effect of the switch from Content Design 1 to Content Design 3 of perceived performance as well as perceived process is weakened when there is an additional change in the visual support. Thus, the increase in these two factors due to the AI-specific additional information is reduced by the additional use of color.

	Direct Effect of Content *CD → TW*		Indirect Effect via Perc. Performance *CD → PER → TW*		Indirect Effect via Perc. Process *CD → PRO → TW*		Indirect Effect via Perc. Purpose *CD → PUR → TW*	
	EST	*SE*	*EST*	*SE*	*EST*	*SE*	*EST*	*SE*
CD2; VD1	0.024	0.154	-0.090**	0.052	-0.005	0.032	-0.027	0.08
CD2; VD2	-0.010	0.199	-0.011	0.075	-0.024	0.048	-0.060	0.122
CD3; VD1	-0.014	0.154	0.011**	0.061	0.045	0.042	0.023	0.092
CD3; VD2	-0.115	0.198	-0.143	0.086	-0.104**	0.067	-0.077	0.121
Index of moderated mediation VD		for CD2	0.079	0.090	-0.012	0.057	-0.033	0.145
		for CD3	-0.250**	0.117	-0.153**	0.091	-0.103	0.151

CD = Content Design; VD = Visual Design; PER = Perceived Performance; PRO = Perceived Process; PUR = Perceived Purpose; TW = Perceived Trustworthiness; EST = Mediation Effect; SE = Standard Error; p-value: ***<0.01; **<0.05; *<0.10

Table 19. Results of the Moderated Parallel Mediation Analysis

5.5 Discussion, Implications & Limitations

This study investigated the aspects of different content and visual design elements of the scores of an AI based AES system. It could be shown that especially for the Content Design no overall effect on the trustworthiness-related factors as well as the perceived trustworthiness towards and understandability of the AI-based AES system could be observed. However, the reason might be that no significant differences could be detected especially between the considered Content Designs 1 and 2. Content Design 1 represents just the assigned AI-based score, while Content Design 2 shows additional information known from paper-based scorings, i.e., the associated level of expectation as well as the assigned score per sentence. Content Design 3 consists of the design elements of Content Design 2 and includes the similarity of the sentences with the respective level of expectation. Regarding Content Design 3, significant differences to Content Design 1 for perceived performance and perceived understandability were identified. One reason for this observation might be that the additional scoring information contradicts the experiences with paper-based scorings. In our example, the AI-based AES system compares the answer with a part of the level of expectation. The partial points are awarded starting from a certain similarity. Compared to Content Design 2, misunderstandings can occur for incorrect answers and overall low similarities. For example, a part of the answer that refers to point A of the level of expectations (see Figure 16) can be given the highest similarity with point B by the AI. If the reference point of the similarity is not specified, even small values of the similarity can lead to the assumption of a misinterpretation by the AI. This misinterpretation is partially resolved by the value of the specific similarity. For the Visual Design, no overall moderation effects could be detected either. For Content Designs 1 and 2, significantly higher values for all variables were observed for Visual Design 2. Visual Design 2 highlights the partially correct parts of the level of expectation and the respective part of the text. This shows that graphic support can promote the understandability of the system, especially when the amount of descriptive information is low. For Content Design 3 no significant differences between Visual Designs 1 and 2 were observed. In our case, a potential information congruence can be observed. Here, interaction effects between Content Design and Visual Design occurred, so that not all information has to be processed to understand the scoring. Furthermore, based on the observed results, it can be assumed that an additional enrichment with content-related information or colors might lead to an information overload. An exception during the analysis has been the perceived purpose, which is the only trustworthiness-related variable that shows no significant differences for all designs. However, this observation is not entirely unexpected. The perceived performance and perceived process relate specifically to the scoring of the AI-based AES system and thus influence the trustworthiness towards and the understandability of the system. According to the previous definition, the perceived purpose includes the intention of the system development and, thus, does not address the actual understandability of the AI-based score. On a generalized level, the results of our pre-study show that current XAI approaches often are not user-centered for examinees. The actual users of the system often do not have enough knowledge about AI to understand more comprehensive global explanations. Furthermore, we show that visual support had the most influence on the participants. In particular, if it can be assumed that users have a basic knowledge of the task and answers, visual support can already promote understandability. However, this moderation effect was mainly achieved when only few information about the scoring were provided by the AI-based system. Providing both more extensive information as well as additional visual support, led to a negative interaction effect, reducing the trustworthiness towards and understandability of the AI-based AES system. This finding confirms the cognitive load theory mentioned earlier (Sweller 2003). As with comparable quantitative questionnaire studies, this work is subject to several limitations. First, our study is guided by a model to explain trustworthiness based on the factors widely used in research, namely performance, process, and purpose. While the factors are generally accepted as relevant, the interpretation of the factors differs based on the context considered. The same applies to the design criteria, where we implemented content and visual design elements. Since both concepts are multifaceted, we cannot guarantee the completeness of our results. Thus, other factors and design approaches that we have excluded may influence trustworthiness and understandability. Second, the results of a survey study are always dependent on the participants. In our case, students from a German university were surveyed. Therefore, before generalizing our results, international differences and differences between or within educational institutions, e.g., the design of the exam tasks, must be considered. Regarding the differences between or within educational institutions, the task-related knowledge of the examinees must be kept in mind, especially when evaluating the understandability of the explanations. Thus, a positive correlation between task-related knowledge and understandability can be assumed. Third, in the considered case a question with a specific level of expectation was used. It is questionable to what extent the results can be transferred to open essay tasks without a specific correct or incorrect answer. Here, there are other evaluation criteria (e.g., the structure and the level of argumentation) whose presentation requires a higher level of explanation.

5.6 Conclusion

The use of AI has been seen as a solution to a variety of problems in many areas, including education. However, while the use of AI is now possible from a technical perspective, the problems associated with its actual use are becoming increasingly apparent. The biggest problem is the understandability of the results since AI systems are often referred to as a black box. According to current research, the solution is the evolution of AI into XAI. However, how exactly AI results must be presented to be perceived as understandable is still a relatively new topic though there are already some approaches in research. In a pre-study, we were able to show that, e.g., the frequently discussed SHAP model is not understood, especially by actual users. Thus, the perspective of the AI administrators, who have in-depth knowledge of AI, is usually considered rather than the perspective of the users themselves. Therefore, we tried to implement real-world aspects of essay scoring using common user interface design principles. The goal was to provide a user-centered explanation of the results of an AI-based AES system to increase the understandability of the system and hence the trustworthiness towards it. Through our study, we could show that a stronger user-centered focus has to be chosen when it comes to the understandable design of (X)AI-based results. We were able to show that, when presenting the essay scores, even color highlighting can improve the comprehensibility among students. This effect was observed for familiar content information such as the points of individual sentences and the assignment to the level of expectation. Still, problems were observed with additional content information that was not known from paper-based exam scorings. For example, the inclusion of the similarity to the level of expectation and the respective color matching was perceived as disturbing. To avoid a cognitive overload, we recommend presenting known information at the content level, i.e., the score or assignment to the level of expectation, as well as highlighting it within the text. This significantly increases understandability and thus trustworthiness. The use of further AI-specific information such as similarity should be evaluated regarding its understandability before being used. This can also include training users to interpret the information provided. Future research should address the transfer of our results to comparable, non-AI-based systems as well as the user interface design from the perspective of the examiners.

C Contributions

The studies presented in Part B address the digital execution and scoring of high-stakes exams. Accordingly, this section will present an aggregated reflection of the research results from Part B. In Section C.1, the results are discussed in aggregate and placed in the current state of research. Based on the discussion, recommendations for the digital execution and scoring of high-stakes exams are derived according to GREGOR ET AL. (2020). Following the level of DSR contribution (Gregor / Hevner 2013), the focus is on generalizable level 2-equivalent recommendations for digital exams. In Section C.2, the results are presented based on the research questions from Section A.3. In Section C.3, the limitations of the aggregated results are presented.

1 Discussion and Implication

Digital Execution of High-Stakes Exams

As demonstrated in Section A.2.2, user-centric research on the digital execution of high-stakes exams is primarily concerned with the psychological aspects that influence examinees' perceptions of and performance on exams. Studies I and II echo this approach by addressing students' intentions to participate in online exams. Study I investigated emergency remote exams as a special case of online exams under the influence of the COVID-19 pandemic. The factors mental challenges, cheating, and fairness were identified as relevant factors. No influence on the intention to participate was found for the technology affinity. In Study II, a long-term observation with supplementary follow-up interviews was conducted to derive implications for online exams in general. Since the examinees in Study I had little to no experience with online exams, implications for introducing online exams can also be derived based on the results. Thus, the recommendations can be partially attributed to the phase of organizational preparation for as well as the execution of digital exams, in this specific case for online assessments.

In general, CBA (regardless of whether they are conducted as online exams) can differ greatly in terms of organization and content, despite a narrowing of the term. In Study II, a distinction was made between two types of online exams. In one exam type, questions were provided and answered via an exam system. In the other exam type, the questions were provided as well as answered in a separate file and then sent to the examiner. Both exam types have advantages and disadvantages for examinees and examiners so that no type can be called more or less suitable in general. The perception of the exam types depended on the examiners' appropriate conception of the exam tasks. Increased use of closed question types and lower taxonomy levels (Bloom et al. 1956) were observed when using an exam system. The increased use of closed question types can be attributed to the possibility of automatic scoring. This aspect was sought out by examiners in order to increase test efficiency. It was shown that closed question types were rated as less suitable by examinees if they did not provide the opportunity to justify their answers. Examiners should therefore ensure that besides the aspect of test efficiency, appropriate question types are used regarding the content tested and that the possibility of justification also exists for closed question types. Due to the decentralized execution of the online exams, an increase in transfer tasks was observed compared to PBA. However, the increase in this kind of task was also not content-driven but an attempt to make it more difficult for task solutions to be passed on to other examinees. In addition, the exams were increasingly implemented as open-book exams to reduce the monitoring effort against the unauthorized use of notes. Therefore, it can be stated that the respective conception of the exams must be considered when evaluating examinees' perceptions of CBA. Non-content-related factors can influence the design of the exams and thus introduce bias into the results regarding the use of and performance on digital exams.

The first factor identified in Studies I and II involved the mental challenges of examinees during online exams. The mental challenges were measured by differences in the perceived ability to concentrate and the perceived stress between the two modes of exam. The results showed that a higher level of mental challenges led to a lower intention to participate. The extent of mental challenges during the exams resulted from the decentralized conduction and varied between examinees. For online exams, a lower ability to concentrate was attributed to the lack of a suitable exam environment at home. Noise pollution and the lack of spatial separation between work and living environment were cited as relevant external influences. Examinees lacked practice in putting themselves into an exam mindset in their private living environments. Over time, an increasing familiarity with the exam process was reported. This was expressed through the formation of routines that improved the ability to concentrate. It was shown that routines were considered important by examinees for both online and in-class exams. A lower level of stress was observed in online exams immediately before and during the exam. This can be explained by the conduct of the exam in a familiar environment and that organizational aspects (e.g., travel to the exam location) were reduced. However, it is not clear whether this stress reduction can be evaluated as positive. Examinees reported activating and performance-enhancing stress before the beginning of PBA as they became aware of the exam situation. Furthermore, additional stress was reported due to the (potential) occurrence of problems. Due to the decentralized execution, the responsibility for the exam process was shifted from the institution and the examiner to the examinee. One proposed solution to this is standardized predefined processes for examinees in case of a problem. Since the examinees were concerned that the use of support increases stress due to a time-consuming and complicated process, these processes should be easy and quick to use. Although the fear of technical problems also exists in the long term, the results showed an increasing familiarity with the organizational process. In Section A.2.2 it has already been shown that the identification of construct-irrelevant factors plays an important role in CBA research (Wise 2019). This is a particular challenge in the case of CBA as a decentralized online exam. Thus, individual external influences on examinees can only be captured and considered with difficulty, if not at all. Organizational and technical problems are often not traceable and only affect individual examinees. While in centralized exams, an examiner is usually on-site to help, in decentralized exams, contact must first be established by examinees to examiners. In addition, it was confirmed that increased mental challenges due to a lack of familiarity are observed when a certain mode of exam is carried out for the first time (Hillier 2014; Hillier 2015). Different research results were found regarding computer anxiety and familiarity in CBA (Clariana / Wallace 2002; Matthíasdóttir / Arnalds 2016; Shermis / Lombard 1998). In our specific case, no relationship between mental challenges and technology affinity, which is closely related to computer anxiety and familiarity, has been demonstrated. Therefore, the first two recommendations address forming routines and improving familiarity regarding the specific mode of exam to reduce mental challenges. The first two recommendations for practice (RP; RP1 and RP2) fall into the phase of preparation for digital exams. Furthermore, it is recommended to implement easy-to-use problem-solving processes and support features. RP3 falls into the digital execution of exams phase. The formulated recommendations are shown in detail in Table 20.

The second factor identified was cheating. The difference in the occurrence of cheating between online exams and PBA, as well as possible causes, were investigated. The results showed that higher perceived cheating reduces the intention to participate in online exams. Overall, higher expected cheating was observed in online exams, with the lack of supervision during the exam named as the main reason. Due to GDPR concerns, real-time proctoring was not possible during the exams. In addition, examinees reported that they might feel disturbed by continuous video surveillance. This is particularly relevant since, according to examinees, cheating can never be eliminated. Examiners attempted to make cheating more difficult by adjusting the exam design to make the communication between examinees as well as the use of unauthorized aids more difficult. These design aspects included time pressure due to a higher scope of tasks, the use of randomized task pools, and a shift from examining factual knowledge to examining transfer knowledge. In Section A.2.2, the prevention and detection of cheating were addressed from a user-centric and a technical perspective. It was shown that cheating has a negative influence on the perception of CBA (Hillier 2014; Hillier 2015). LAUBSCHER ET AL. (2005) emphasized that detecting cheating should not disturb examinees during the exams. In this regard, many examinees expressed that they would feel disturbed by continuous video surveillance. Furthermore, examinees also assumed that no exam is free of cheating (Sindre / Vegendla 2015). In the current research, different technical approaches were discussed to carry out cheating detection after the exam took place (Kaya / Özel 2015; Kleerekoper / Schofield 2019; Laubscher et al. 2005;

Opgen-Rhein et al. 2018). The problem with these approaches are the data required. In particular, textual responses were evaluated for similarity and writing patterns. However, this approach was ineffective for short or closed question types (Kleerekoper / Schofield 2019). Study II found that versioning and randomized question pools are one way to make cheating more difficult. However, DERMO (2009) pointed out that individual examinees may consider this approach unfair. Here, the basic prerequisite is that the different tasks in a question pool have a comparable level of difficulty. Two primary types of cheating were addressed in the exams underlying Studies I and II: the use of unauthorized aids and communication between examinees. Considering the requirements for the content-related design of exams presented at the beginning of Section C.1, preventing cheating through an adapted exam design was found to be the most appropriate approach. Therefore, it is recommended to prevent cheating by modifying the task design in high-stakes exams (see Table 20). RP4 falls into the digital execution of exams phase.

The third factor identified was suitability for fair grading, which was measured by the differences between online exams and PBA regarding the perceived difficulty of the exam and the fair assessment of individual skills. Higher perceived suitability of fair grading increased the intention to participate. Fairness was assessed differently depending on the type of exam and the associated design. In particular, the difficulty of exams using an exam system was rated as significantly higher in the long term, while it was rated as quite easy during the first semester. This may indicate an adjustment in the exam design. Thus, the proportion of exams using mixed knowledge types and mixed question types was increased. Furthermore, the exams were increasingly implemented as open-book exams, as mentioned in the discussion of cheating. To prevent cheating, examinees also reported an adjustment in the wording of the tasks. For example, there was a deliberate attempt to avoid keywords from the lecture within the formulation of the exam tasks. This was considered problematic since students often orientated and prepared themselves using these keywords. Therefore, it was sometimes more difficult to precisely respond to the task. Due to the lack of delimitation, students were additionally put under time pressure and could not rate their performance during the exam. Examinees also stated that implementing of open-book exams increasingly led to the feeling that exams were more difficult, as the permitted aids increased the expected quality of the answers. A correlation was observed between the suitability of fair grading, mental challenges, and problems during the online exam. Both aspects reflect individual factors that influence examinees during the exam. It can be assumed that greater mental challenges and occurring problems reduce the individual exam performance and thus the perceived suitability for fair grading. In Section A.2.2, it was shown that fairness is an important criterion in the implementation of CBA. Important factors here were cheating, lack of technical equipment, and lack of familiarity with the mode of exam (Hillier 2014; Hillier 2015). The importance of cheating and mental challenges was already discussed in the context of the specific factors presented in Section C.1. In addition, it was shown that task design plays an important role. For example, inappropriate task design creates additional time pressure. As a result, examinees perceived exams more as time management-based than content-based (Matthíasdóttir / Arnalds 2016; Stephenson 2018). This effect was further enhanced by additional aids, which increased the required quality of the answers (Matthíasdóttir / Arnalds 2016; Stephenson 2018). Regarding the exam design, results from related research found that a close adaption of tasks between PBA and CBA is beneficial (Piaw 2012). This is intended to promote familiarity with the processing of the exam task. However, it can be assumed that due to the technical possibilities for authentic assessment and additional question types, appropriate CBA task design can deviate from that of PBA task design. For ICT-related tasks in particular, an authentic digital exam environment is perceived as advantageous (Higgins et al. 2002; Higgins et al. 2005; Ju et al. 2018; Piech / Gregg 2018). CANDRLIC ET AL. (2014) spoke of the use of adequate question types. Therefore, it is recommended that examiners should be trained in designing and implementing digital exam tasks (e.g., selection of digital question types) instead of just implementing PBA-based tasks digitally. RP5 falls into the phase of preparation for and execution of digital exams (see Table 20).

To reduce the bias of construct-irrelevant factors on examinee's high-stakes exam scores …		
Principle of …	**Mechanism**	**Rationale**
… promoted exam routine (RP1)	… encourage examinees to develop individual exam routines …	… because this may reduce mental challenges and promotes performance readiness.
… increased familiarity (RP2)	… train examinees in dealing with the respective exam process and system …	… because this can already increase familiarity before the execution and thus reduce stress.
… problem-solving (RP3)	… implement easy-to-perform problem-solving processes and support features …	… because this reduces the examinees' mental load, regardless of the occurrence of the problems.
… reduced cheating (RP4)	… reduce cheating by deliberate design of online exams …	… because cheating control by proctoring tools can increase students' mental stress in the exam.
… exam task design (RP5)	… train examiners to design digital exam tasks …	… because insufficient exam tasks make valid knowledge assessment difficult.

Table 20. Recommendations for Practice - Studies I and II

Digital Scoring of High-Stakes Exams

As Section A.2.2 shows, CBA often claims the advantage of higher test efficiency through automatic exam scoring, among other capabilities. Since the scoring of closed question types is often no longer a problem, the current research focuses on the scoring of open question types. While the technical perspective is often considered, the user-centric perspective remains largely unexamined. Study III investigated examinees' trust in using AI-based AES systems in high-stakes exams and the differences in trust regarding semi-automatic and fully automatic scoring processes. While in the semi-automatic scenario, the scoring was controlled by a human, no human post-scoring was performed in the automatic scenario. Overall, confidence in human scoring was highest, followed by confidence in semi-automatic scoring. Significantly lower confidence was demonstrated for automatic scoring. Furthermore, the examinees' characteristics (trustor), the system characteristics (trustee), and the exam characteristics (environment) were considered.

Among the examinees' characteristics, personality traits were identified as the most important factor for increasing trustworthiness of and trust in AI-based scoring. This factor was determined by agreeableness based on the Big Five Personality Traits Taxonomy (John et al. 2008), wherein higher agreeableness was found to lead to higher trustworthiness and higher trust. Agreeableness is expressed by the attributes of understanding, benevolence, and compassion. It can be assumed that people with these characteristics are more likely to see a fit between the task and the system. Regarding the former use of AI-based services and technology affinity, no influence on trustworthiness was proven. One possible reason for this can be that the examinees are not active but merely passive users. Thus, examinees have to trust the scoring process without being able to influence it. Section A.2.2 showed that the user-centric perspective of digital scoring of high-stakes exams is often neglected. However, regarding the user-centric perspective of the digital execution of exams, it was shown that personality traits influence the acceptance of CBA use. Here, a positive effect on the perceived ease of use of CBA and social influence was demonstrated for agreeableness (Terzis et al. 2012). Since acceptance plays an important role in determining the intention to use, the examinees' personality traits should be considered in the automatic scoring of essays. Therefore, it is recommended to address especially examinees with lower agreeableness to increase the trustworthiness of as well as the trust in AI-based AES systems (see Table 21). RP6 falls into the digital exam scoring phase.

Regarding the system characteristics, it was shown that the AI-based system was evaluated as suitable for scoring essays. A higher perception of the three characteristics (competence, reliability, and fairness) led to higher trustworthiness of the system. High values were observed for the competence and fairness of the system, while medium values were shown for the reliability of the system. Thus, AI-based systems are considered to have the ability to fairly evaluate essays, however, there are concerns about consistent scoring without contradictions. Competence and reliability were rated significantly higher for the semi-automatic system than for the automatic system. Thus, the human scorer in this scenario is attributed to a professional competence that exceeds the system's capabilities. Fairness was assessed as comparable for both scenarios and included objectivity, which is a core characteristic of automatic scoring. It can therefore be assumed that no increase in objectivity is expected from the additional human rater. As shown in Study III, people have more trust in human deciders to make important decisions, especially if they are more associated with a specific activity than automatic AI-based systems (Ashoori / Weisz 2019; Elson et al. 2021; Richardson / Clesham 2021). Previous research also concludes that semi-automatic systems are preferable (Shermis 2014, 2015; Buyrukoglu et al. 2019). The second recommendation (RP7) based on the results of Study III is that semi-automatic systems should be preferred over automatic systems for the digital scoring of essays in high-stakes exams, as this increases confidence in the correctness of the scores. Furthermore, the perception of the system characteristics significantly influences the trustworthiness of (semi-)automatic scoring systems. It is therefore recommended to provide examinees with information on the capabilities and suitability of the AES system and the scoring process (RP8). Recommendations RP7 and RP8 address the digital scoring of exams. Table 21 provides an overview of the recommendations.

The environmental characteristics comprised the task-related factors, namely, the perceived complexity of essay scoring and the perceived stakes of the scores for examinees. Since high values were observed for both factors, the essay scoring in exams was rated as both difficult and important. However, a significant negative influence on trustworthiness was only found for task complexity. A higher perceived complexity of essay scoring led to lower trustworthiness of the AI-based system. In our case, a higher perceived stake in the essay scores did not influence the trustworthiness of the AI-based AES system. One reason for this may be the good task-AI fit, as the system was rated as suitable for scoring essays overall. Thus, the stakes of an exam may not affect trust as long as the AES system is considered suitable. Since the provision of information on the capability and suitability of the system was already derived in RP8, no further recommendation arises from the results for the environmental characteristics.

To increase the trust in the digital scoring of essays in high-stakes exams …		
Principle of …	**Mechanism**	**Rationale**
… considered personality traits of examinees (RP6)	… take into account the personality traits of examinees…	… because less agreeable examinees show reduced trustworthiness of the system use.
… semi-automatic essay scoring (RP7)	… use semi-automatic systems for scoring essays …	… because an additional human scorer increases the trustworthiness by his perceived professional competence.
… provided process information (RP8)	… provide examinees with information on the characteristics of the AES system …	… because this helps to evaluate the suitability of the system for the task and thereby increases trustworthiness.

Table 21. Recommendations for Practice - Study III

Study IV investigated the design of a semi-AES system. In Section A.2.2, it was shown that related research focuses on improving scoring accuracy for different open question types. Following the results of Study III, examinees prefer AI-based semi-AES systems to fully automated systems. In this context, a system was designed to support examiners in scoring essay tasks. Based on the examiner's user story (scoring and documentation of exams) and the challenges they face in scoring essays (ensuring the principles of good grading, transparency and test efficiency), meta-requirements and three design principles for the semi-AES system were derived. These design principles were evaluated using a prototype implementation. The results showed that the examiners were largely satisfied with the functionalities and the user interface. The first two design principles comprised the two components of semi-automatic scoring of essays, namely, machine-learning-based pre-scoring and human post-scoring of essays. The AI-based pre-scoring was performed with a neural network that was trained with old exam scores and can use new exam scores for future exam scoring (supervised learning). The human post-scoring is based on a level of expectation, which is also represented in the system. This should make it easier for examiners to verify the pre-scoring. Overall, this approach addresses the challenges of objectivity, validity, reliability, and transparency in exam scoring, as well as test efficiency. The design principles can also be seen as a generalizable basis for comparable software artifacts with different levels of automation and other (semi-)open question types. The third design principle dealt with the documentation in the post-scoring process to promote transparency in scoring and communication with examinees. The documentation features included the use of comments and the direct assignment of the level of expectation to the corresponding parts of the essays. Since the design principles have already been documented within Study IV based on the chosen structure for Section C.1, they will be listed in Table 22. Recommendations RP9 to RP11 are assigned to the technical perspective of the digital exam scoring phase.

To support examiners in the scoring of essays in high-stakes exams …		
Principle of …	**Mechanism**	**Rationale**
…machine-learning-based pre-scoring (RP9)	… provide a machine-learning-based scoring mechanism that is able to generate automated pre-scoring drafts based on transparent evaluation criteria …	… because this can help examiners to reduce the scoring workload while increasing the objectivity of essay scoring.
… manual, simultaneous post-scoring (RP10)	… provide an easy-to-use user interface for manual post-scoring that uses the pre-scoring as basis to reduce the workload …	… because this can help to improve the overall scoring quality, required workload, students' acceptability of exam results, and acts as a manual verification step of the diagnostic quality of the machine-learning-based pre-scoring mechanism.
…documenting the post-scoring process (RP11)	… provide a traceable and transparent scoring process by documenting the examiner's decisions and the underlying evaluation criteria …	… because this can increase the transparency of the scoring process, might be mandatory for providing an explainable scoring process, and can help communicate the scoring results to the students.

Table 22. Recommendations for Practice - Study IV

Study V addressed the presentation of the results of an AI-based (semi-)AES. The focus was on the understandability and comprehensibility of the scores. Study III showed that the system's perceived capabilities influence the examinees' perceived trustworthiness of an AI-based AES system. Based on the user story of the examiners and the challenges of scoring essays identified in Study IV, it was concluded that the results must be documented. Since the lack of comprehensibility in AI-based systems is one of the core problems, research is increasingly focusing on XAI. Similar to the research in the field of AI, however, the technical perspective is primarily considered, which deals with the improving of scoring accuracy. A preliminary study within Study V showed that the comprehensibility of the XAI results and, thus, the trust in such systems is questionable. This concerns the examiners and examinees as actual users in particular. Therefore, different PBA aspects of essay scoring using common user interface design principles were evaluated. These included different content design elements (metrics for scoring) and visual design elements (color coding). The goal was to provide a user-centered explanation of the results of an AI-based AES system. It was shown that a stronger user-centered focus must be chosen when it comes to the understandable design of (X)AI-based results. In particular, the use of familiar documentation of the scoring known from PBA (e.g., color highlighting, and partial points) improved the comprehensibility among examinees. Problems arose when new, AI-related information was used. Although little research was carried out in this area in education, there are parallels to recommendations in the digital execution of exams from Section A.2.2. It was shown for the implementation of CBA that familiarity with the process had an important influence on the acceptance of examinees (Hillier 2014; Hillier 2015; Jeong 2014; Clariana / Wallace 2002; Boevé et al. 2015; Backes / Cowan 2019). PIAW (2012) went one step further and recommended that CBA follow PBA as closely as possible. This aspect was confirmed in this study. However, it should be noted that examinees have no experience with such systems and are therefore not familiar with interpreting the key figures of AI and the information they contain. In the future, a higher level of familiarity may also improve the understanding of AI-related key figures. As a first step, it is therefore recommended to provide user-centered documentation of the exam scoring for examinees to promote comprehensibility and, thus, confidence in the results. This documentation should consist of known information (e.g., the score or assignment to the level of expectation, as well as highlighting it within the text). Furthermore, when using additional AI-specific information (e.g., similarity), it is recommended to evaluate the comprehensibility beforehand and, if necessary, ensure the examinees are familiar with these key figures. Recommendations RP12 and RP13 are assigned to the technical and user-centric perspective of the digital exam scoring phase.

To improve examinee's comprehensibility of AI-based AES scores in high-stakes exams …		
Principle of …	**Mechanism**	**Rationale**
… user-centric documentation of scoring (RP 12)	… provide user-centric documentation of the essay scoring for examinees by using known information …	… because this can help examinees to understand the scoring and hereby increase the trust in the scoring.
… familiar scoring figures (RP 13)	… evaluate AI-specific key figures regarding their comprehensibility beforehand…	… because this can prevent examinees from being overloaded with the interpretation of the key figures.

Table 23. Recommendations for Practice - Study V

2 Conclusion

This dissertation investigated the digitization of high-stakes exams. First, a literature review was carried out in Section A.2. Then, five studies were conducted based on the identified research objectives.

Phase	Recommendation	
Preparation for the Digital Exam	RP1	Principle of **promoted exam routine**
	RP2	Principle of **increased familiarity**
	RP5	Principle of **exam task design**
Execution of the Digital Exam	RP3	Principle of **problem-solving**
	RP4	Principle of **reduced cheating**
Scoring of the Digital Exam	RP6	Principle of **considered personality traits of examinees**
	RP7	Principle of **semi-automatic essay scoring**
	RP8	Principle of **provided process information**
	RP9	Principle of **machine-learning-based pre-scoring**
	RP10	Principle of **manual, simultaneous post-scoring**
	RP11	Principle of **documenting the post-scoring process**
	RP12	Principle of **user-centric documentation of scoring**
	RP13	Principle of **familiar scoring figures**

Table 24. Recommendations for Practice - Overview

A total of 13 recommendations for practice were derived from the research, which can be assigned to the three stages of the digital assessment process (see Table 24). These recommendations are used in the following to answer the research questions from Section A.3.

Research Field: Execution of Digital Exams – User-Centric Perspective	
RQ1	Which factors must be considered in the spontaneous introduction of the execution of digital exams?

Related research showed that user-centric aspects such as anxiety, stress, cheating, and familiarity frequently influence the introduction of digital exams. These construct-irrelevant factors play a particularly important role in high-stakes exams since these exams involve the certification of certain competencies based on public criteria. Although the conducted research studies also showed that familiarity positively influences the perception of digital exams in the long term, examiners and institutions can also address the negative influences beforehand. In this context, three recommendations were derived for the phase before the conduction and two recommendations were derived for the conduction itself. It is important that examinees are informed about the organizational process of the exam (RP2) before it is conducted and that they are supported in this process to establish an exam routine (RP1). This should reduce the mental challenges resulting from the new exam process and the uncertainty due to the lack of familiarity. In addition, examiners should ensure that task implementation considers the new framework when switching from PBA to CBA (RP5). For example, digital exams offer the possibility of creating an authentic exam environment, which cannot be taken for granted. When transferring PBA tasks to CBA tasks, the appropriate implementation must be taken into account accordingly. Students also face problems during the exam, including organizational, technical, and content-related problems. Here, one must ensure that easy-to-use support features are provided regardless of decentralized or centralized implementation (RP3). These measures can also reduce mental challenges and increase the perceived suitability for fair grading. As with any exam, cheating must be avoided or at least made more difficult (RP4). Here, not only types of cheating known from PBA play a role, but also new cheating opportunities that arise through digital execution.

Research Field: Scoring of Digital Exams – User-Centric Perspective	
RQ2	Which factors influence the trust of examinees in AI-based (semi-)AES systems?

The user-centric perspective of digital high-stakes exam scoring was often neglected in previous research, where the focus in digital scoring was mostly on the technical perspective and the improvement of scoring accuracy. While closed question types do not pose a problem for automatic scoring, text quality is often considered from a structural and linguistic perspective in the scoring of essays and short answers. An increasing focus on content has emerged through the use of AI. However, research on AI-use showed that people place more trust in human decision-makers than machines when making important decisions. Nevertheless, semi-AES systems can also be used to perform pre-scoring to increase test efficiency. Thus, they serve as decision support systems that support the examiner. It was shown that examinees prefer semi-AES systems from a trust-based perspective (RP7). However, the assessment of trustworthiness is influenced by different factors. First, people have different personality traits that influence trust. In particular, people with a low level of agreeableness must be addressed by examiners during the introduction, as they usually have a lower level of trust in such systems (RP6). Second, a higher perception of competence, fairness, and reliability regarding the system leads to higher trustworthiness of the AES systems. Therefore, examiners and institutions should inform examinees about the functionality and suitability of the essay scoring system (RP8).

Research Field: Scoring of Digital Exams – Technical Perspective	
RQ3	How must AI-based (semi-)AES systems be technically designed to be perceived as useful?

The design aspects of a (semi-)AES system were addressed from a technical perspective. Previous research primarily investigated the selection of appropriate scoring criteria and the improvement of scoring accuracy. One focus of the dissertation was on the basic functionalities of a semi-AES system. It was shown that an AI-based pre-scoring should be complemented by the possibility of a human post-scoring (RP9 and RP10). This should increase test efficiency and, at the same time, address weaknesses in automatic scoring at the content level. In addition, documentation features were implemented to increase transparency and, thus, comprehensibility for the scoring (RP11). This also allows for the possibility of individual feedback. The second focus was on the presentation of AI-based scoring results for examinees. Since a major disadvantage of using AI is the lack of transparency, it is often compared to a black box. One approach to solve this problem is the use of XAI, which is supposed to explain its decisions and thus make them comprehensible. However, research in this area investigate new explanatory models that mostly address AI administrators rather than actual users, such as examiners and examinees. Therefore, the results are often not understandable. It is therefore recommended to focus more on the actual target group when presenting the explanations (RP12). This can be implemented by using known scoring key figures or highlighting from PBA. It is also recommended to familiarize users with the key figures used, especially when new AI-related metrics are used (RP13). This can improve the comprehensibility of and trust in the AES system.

3 Limitations and Future Perspectives

As with any research project, the results of this dissertation are subject to limitations. Since the limitations of the individual research contributions from Part B have already been addressed in the context of the respective studies, the limitations of the overall dissertation are presented in the following. Moreover, starting points for further research will be introduced.

The first limitation of this dissertation results from the weak systematization of the domain of digital testing in research and practice. As stated in the definition and demarcation of terms (Section A.2), many (sometimes incorrectly) synonymously used terms exist. These result from the (partly inconsistent) linguistic representations of digitization (e.g., digital, computer-based, web-based, online, electronic) and assessment (e.g., assessment, evaluation, testing, examination). By narrowing down to the central terms "digital" and "exam" as well as closely related terms such as "computer-based" and "assessment", an attempt was made to determine a consistent field of application. It should be noted that the results from Studies I and II are based on online exams (as a special form of computer-based exams). Although the recommendations for practice derived from these studies (Section C.1) can also be partially transferred to in-class computer-based exams, the characteristics of the respective form of digital exams must be considered. Therefore, empirical verification of the recommendations regarding in-class computer-based exams remains an objective for future research. From a research perspective, the prior development of a taxonomy to distinguish between different types of digital exams might be useful.

The second limitation arises from the methodological limitation of the introductory literature research of the dissertation for the identification of the current state of research and the derivation of research gaps (Section A.2). Due to the limited scope of the databases, potentially relevant literature may not have been included. In particular, the open framing of the search terms addresses the digitization of high-stakes exams on a broad scale. Future research should also consider the transfer possibilities of new research results from other disciplines. For example, research on anxiety in the field of psychology offers potential insights into the topic of test anxiety when taking digital exams. The increasing use of AI-based systems offers further potential. For example, traditional human-computer interaction is increasingly developing into human-AI interaction, since language models enable complex and individual interactions (e.g., using ChatGPT). Nevertheless, the identified articles reflect a current overview of previous research. Generally, one must also consider that the evaluation, systematization, and discussion of literature reviews always exhibit a certain subjectivity. However, reproducibility was ensured through extensive documentation of the procedure and the formulation of relevance criteria.

The third limitation results from the limited generalizability of the results due to the impact of the COVID-19 pandemic on the participants of the studies. This is especially true for the results in Section C.1, which are based on Studies I and II. In the context of conducting online exams, the study investigated examinees' perceptions of mental challenges. Since these are the examinees' individual, subjective perceptions, the generalizability of the results is only possible to a certain extent. This makes it difficult to classify pandemic-specific challenges, especially those at the beginning of the pandemic. Research results on digital exams during the pandemic are therefore often addressed in a separate research field, namely emergency remote assessments. Therefore, Study II explicitly transferred research findings from Study I to COVID-19-independent lessons learned. A possible transfer of individual results to other forms of digital assessment (e.g., in-class digital exams) must be examined in future research.

The fourth limitation concerns the generalizability of the recommendations for practice against the background of the target group. Since the studies were conducted at a German university, the results can primarily be applied to comparable institutions and stakeholder groups. Relevant in this context is the validity of the results in the case of regional or institution-dependent changes. In a regional, especially international, comparison, there are differences in the cultural and legal classifications of the digital implementation and AI-based scoring of high-stakes exams. For example, divergent data protection regulations or cultural views allow different ways of detecting cheating in exams (e.g., proctoring tools). The same concerns AI use in AES. On the one hand, there are countries where essay tasks play a minor role in exams and closed question types are primarily used. These can be evaluated without any problems, even without complex AI models. On the other hand, there are also privacy concerns regarding the use of datasets for training AI models. The institution under consideration is experiencing a similar situation. Thus, while educational institutions are similar in the basic idea of the CIA triad, this does not mean that digital exams are also comparable across institutions. A relevant example is the school context, where an overall grade may be based on numerous individual exams. Here, studies must consider a different understanding of high-stakes exams.

In conclusion, the need for research in the digitization of high-stakes exams primarily results from technological developments and increasing familiarity with ICT among the general population. These aspects influence the CIA triad presented in Section A.1 at all levels. For example, digital transformation will increasingly change the professional requirements of future employees, which will require an adaptation of the curriculum anchored competencies to be promoted and, consequently, of the instruction as well as the assessment. In addition, technological progress offers an increasing use of digital tools in instruction. This includes individualizing the learning process through AI-based tutoring and recommendation systems that provide goal-oriented feedback based on diagnostic and formative assessments. In the area of summative assessments, further research into the design, implementation, and scoring of exams is needed. In this context, LLM are said to have the greatest potential for a lasting impact on education. LLM consist of neural networks and are trained with large amounts of data to generate probability-based sequences of words as answers to tasks. They are characterized by a broad knowledge base as well as syntax and semantics similar to those of the human language (Kung et al. 2023; Rudolph et al. 2023). Since the end of 2022/beginning of 2023, LLM have emerged as a topic of public interest due to the introduction of Open AI's ChatGPT, which is considered one of the best-trained LLM to date (OpenAI 2023).

The topic of LLM is still in its infancy in current research. Especially in the educational context, AI-based texting is presented as a danger for decentralized assessment (e.g., term papers or online exams). So far, the focus has been primarily on the (unauthorized) use of ChatGPT from the perspective of examinees. Initial studies showed that tools like ChatGPT can be used to solve a variety of high-stakes exams to at least a sufficient degree. Although the subject domain and the knowledge type still influence the quality of the AI-generated answers, for individual domains, the answer quality is already equivalent to or better than that of human examinees (Lo 2023; Hobert et al. 2023; Kung et al. 2023; Bordt / Luxburg 2023). However, since AI-generated answers are already sufficient to pass exams in many subject domains, an increasing number of challenges are emerging for educational institutions (Susnjak 2022). Thus, it is becoming increasingly difficult to distinguish AI-generated responses from examinee-generated responses (Kasneci et al. 2023). The resulting implications for the design and implementation of exams must be addressed through practice and research. The challenges are contrasted with the potentials of LLM in the conceptualization and scoring of exams for institutions and examiners. However, current research has primarily described the potentials, while lacking empirical verification. First, the generative character of LLM may simplify the creation of exam tasks and supplementary exam materials (e.g., case studies). This allows examiners to, for example, create individual tasks and individual feedback in CAT based on different assessment criteria (see Section A.2), thus achieving higher assessment accuracy (Kasneci et al. 2023). Second, LLM may be used for AES, increasing the scoring accuracy of higher taxonomy levels and operators (Kasneci et al. 2023; Rudolph et al. 2023). The results of the AI-generated exam answers imply that LLM are generally capable of answering the exam task on the content level. This should enable LLM, depending on the subject domain and knowledge type, to correctly evaluate answers concerning their content. However, there are still challenges to be overcome before LLM can be used in education. One of the biggest problems of so-called generative AI is that it responds with partially plausible-sounding but incorrect information (OpenAI 2023). Since the answers are generated as probability-based word sequences, information is sometimes newly generated without any actual knowledge, and fictitious references are cited (Lo 2023; Hobert et al. 2023; Kasneci et al. 2023).

According to Kasneci et al. (2023), this carries the risk that users will accept the provided information of the LLM as true without critical examination. For example, bias and unfairness can occur as a result of the training data. These arise due to bias or underrepresentation of subgroups in the training data and have a negative impact on the results. Therefore, users need to be trained in the interpretation and evaluation of the results.

Appendix

Appendix A: Questionnaire Studies I and II 118

Appendix B: Guideline Follow-Up Interview Study II 121

Appendix C: Questionnaire Study III 122

Appendix D: Questionnaire Study V 126

Appendix A: Questionnaire Studies I and II

Bitte geben Sie Ihr Alter an. ***[Free Text]***		
[Free-text field]		
Bitte geben Sie Ihr Geschlecht an. ***[Single Choice]***		
männlich	*weiblich*	*divers*
Welcher Fakultät gehören Sie an? ***[Single Choice]***		
Fakultät für Agrarwissenschaften	*Juristische Fakultät*	
Fakultät für Biologie und Psychologie	*Philosophische Fakultät*	
Fakultät für Chemie	*Sozialwissenschaftliche Fakultät*	
Fakultät für Forstwissenschaften und Waldökologie	*Theologische Fakultät*	
Fakultät für Geowissenschaften und Geographie	*Universitätsmedizin*	
Fakultät für Mathematik und Informatik	*Wirtschaftswissenschaftliche Fakultät*	
Fakultät für Physik	*Sonstiges: [Free-text field]*	
Musste vor Beginn der Onlineprüfung ein Identifikationsverfahren durchgeführt werden? ***[Single Choice]***		
Es wurde kein Identifikationsverfahren durchgeführt.		
Es wurde ein Video-Ident-Verfahren eingesetzt (Login mit Nutzername sowie Fotoaufnahme von Gesicht und Studienausweis).		
Es wurde ein Live-Video-Verfahren eingesetzt (Live Identifikation per Videoübertragung, z. B. via Zoom).		
Sonstiges: [Free-text field]		
Wenn das Video-Ident-Verfahren eingesetzt wurde: **Bitte geben Sie an, inwiefern die folgenden Aussagen zutreffen.** ***[Likert Scale from „1 = Stimme überhaupt nicht zu“ to „7 = stimme voll und ganz zu“]***		
Die Nutzung des Video-Ident-Verfahrens hat ohne Probleme funktioniert.		
Die Nutzung des Video-Ident-Verfahrens war leicht verständlich.		
Ich bewerte das Video-Ident-Verfahren als sehr gut.		
Wenn das Video-Ident-Verfahren eingesetzt wurde: **Bitte geben Sie uns weiteres Feedback zum Einsatz des Video-Ident-Verfahrens.** ***[Free Text]***		
[Free-text field]		
In welchem Modul (Modulname) haben Sie die Onlineprüfung geschrieben? ***[Free Text]***		
[Free-text field]		
Wie wurde die Onlineprüfung durchgeführt? ***[Single Choice]***		
Download der Aufgabenstellung und Upload der Lösungen		
Beantwortung der Aufgabenstellung in einem E-Prüfungssystem (z. B. ILIAS)		
Sonstiges: [Free-text field]		
Welche Aufgabentypen wurden in der Onlineprüfung eingesetzt? ***[Multiple Choice]***		
Freitextaufgaben		
Single-Choice-Aufgaben (Jeweils genau eine korrekte Antwort)		
Multiple-Choice-Aufgaben (Keine, eine oder mehrere korrekte Antworten)		
Anordnungsaufgaben (In Reihenfolge bringen von Items)		
Zuordnungsaufgaben (Zuordnen von zwei oder mehreren Items)		
Lückentext-/Beschriftungsaufgaben		
Sonstiges: [Free-text field]		

Table 25. Questionnaire Studies I and II - Part 1

Welche Wissensarten/Kompetenzen wurden in der Onlineprüfung abgefragt bzw. überprüft? ***[Multiple Choice]***	
Faktenwissen (z. B. Jahreszahlen oder Bullet Point Listen aus den Veranstaltungsunterlagen)	
Transferwissen (z. B. Anwendung von erlerntem Wissen auf einen neuen Sachverhalt)	
Mathematische Fähigkeiten (z. B. Lösen von mathematischen Gleichungen o. Ä.)	
Modellierungs-/ Programmierfähigkeiten	
Juristische Gutachten	
Sonstiges: [Free-text field]	
Waren während der Prüfung Hilfsmittel erlaubt? Wenn ja, welche: ***[Multiple Choice]***	
Es waren keine Hilfsmittel erlaubt.	
Es gab keine Einschränkungen bezüglich der erlaubten Hilfsmittel.	
Formelsammlung	
Taschenrechner	
Lernunterlagen (z. B. eigene Zusammenfassungen, Vorlesungsfolien oder Bücher)	
Gesetzestexte (o. Ä.)	
Computersoftware (z. B. Word, Excel, R, SPSS etc.)	
Sonstiges: [Free-text field]	
Wenn Hilfsmittel erlaubt waren:* Bitte geben Sie an, inwiefern die folgende Aussage zutrifft oder nicht.** ***[Likert Scale from „1 = Stimme überhaupt nicht zu" to „7 = stimme voll und ganz zu"]	
Die erlaubten Hilfsmittel waren notwendig für die Beantwortung einzelner Aufgaben.	
Die erlaubten Hilfsmittel waren hilfreich für die Beantwortung einzelner Aufgaben.	
Haben Sie unerlaubte Hilfsmittel (z. B. Vorlesungsunterlagen oder die Bearbeitung in Gruppen) genutzt? ***[Single Choice]***	
Ja	*Nein*
Inwiefern sind die folgenden Probleme während der Onlineprüfung aufgetreten? ***[Likert Scale from „1 = Nicht aufgetreten" to „7 = In großem Umfang aufgetreten"]***	
Technische Probleme	
Organisatorische Probleme	
Inhaltliche Probleme	
Sonstige Probleme	
Haben Sie weitere Anmerkungen zur Durchführung oder zu technischen Problemen bei der Onlineprüfung? ***[Free Text]***	
[Free-text field]	
Bitte geben Sie an, inwiefern die folgenden Aussagen zutreffen oder nicht. ***[Likert Scale from „1 = Stimme überhaupt nicht zu" to „7 = stimme voll und ganz zu"]***	
Insgesamt habe ich mich vor der Onlineprüfung gut vorbereitet gefühlt.	
In Präsenzprüfungen in der Universität fällt mir die Konzentration auf die Prüfung leichter.	
Die Durchführung der Prüfungsleistung zuhause verursacht mehr Stress als die Durchführung in der Universität.	
Im Vergleich zu Präsenzprüfungen empfand ich die Onlineprüfung als schwieriger.	
Die Benotung der Onlineprüfung wird meinen tatsächlichen Wissensstand zu dem Thema widerspiegeln.	
In Relation zu dem Wissen der Kommilitonen wird meine Note meinen tatsächlichen Wissensstand zu dem Thema widerspiegeln.	
Ich wollte diese Prüfungsleistung lieber als Onlineprüfung anstatt einer Präsenzprüfung schreiben.	
Ich bin froh, dass die Prüfung online durchgeführt wurde, da ich Angst vor einer möglichen Corona-Infektion habe.	
Nach der Beendigung der Corona-Pandemie sollten Onlineprüfungen wieder durch Präsenzprüfungen ersetzt werden.	

Table 26. Questionnaire Studies I and II - Part 2

Ganz allgemein gesprochen: Glauben Sie, dass man den meisten Menschen vertrauen kann, oder dass man im Umgang mit anderen Menschen nicht vorsichtig genug sein kann? ***[Bipolar Rating Scale from 0 to 10]***
Man kann nicht vorsichtig genug sein. | Den meisten Menschen kann man vertrauen.
Glauben Sie, dass die meisten Menschen versuchen, sich fair zu verhalten oder versuchen die meisten Menschen, Sie auszunutzen, wenn sie die Gelegenheit dazu haben? ***[Bipolar Rating Scale from 0 to 10]***
Die meisten Menschen versuchen, mich auszunutzen. | Die meisten Menschen versuchen, sich fair zu verhalten.
Glauben Sie, dass die Menschen meistens versuchen, hilfsbereit zu sein, oder dass die Menschen meistens auf den eigenen Vorteil bedacht sind? ***[Bipolar Rating Scale from 0 to 10]***
Die Menschen sind meistens auf den eigenen Vorteil bedacht. | Die Menschen versuchen meistens, hilfsbereit zu sein.
Bitte geben Sie an, inwiefern Sie den folgenden Aussagen zustimmen. ***[Likert Scale from „1 = Stimme überhaupt nicht zu“ to „7 = stimme voll und ganz zu“]***
Ich glaube, dass der Anteil der Prüfungsteilnehmenden, die unerlaubte Hilfsmittel nutzen, bei Onlineprüfungen höher ist als bei Präsenzprüfungen.
Ich glaube, dass Studierende, die normalerweise nicht bei Prüfungen betrügen, bei Onlineprüfungen vermehrt auf unerlaubte Hilfsmittel zurückgreifen werden, da es nur eine eingeschränkte Aufsicht gibt.
Ich glaube, dass Studierende, die normalerweise nicht bei Prüfungen betrügen, bei Onlineprüfungen vermehrt auf unerlaubte Hilfsmittel zurückgreifen werden, da sie ansonsten einen Nachteil gegenüber betrügenden Studierenden erwarten.
Ich glaube, dass Studierende, die erst bei Onlineprüfungen auf unerlaubte Hilfsmittel zurückgegriffen haben, dies zukünftig auch bei Präsenzprüfungen verstärkt tun werden.
Im Folgenden geht es um Ihre Interaktion mit technischen Systemen. Mit ‚technischen Systemen‘ sind sowohl Apps und andere Software-Anwendungen als auch komplette digitale Geräte (z. B. Handy, Computer, Fernseher, Auto-Navigation) gemeint. Bitte geben Sie den Grad Ihrer Zustimmung zu folgenden Aussagen an. ***[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]***
Ich beschäftige mich gern genauer mit technischen Systemen.
Ich probiere gern die Funktionen neuer technischer Systeme aus.
In erster Linie beschäftige ich mich mit technischen Systemen, weil ich muss.
Wenn ich ein neues technisches System vor mir habe, probiere ich es intensiv aus.
Ich verbringe sehr gern Zeit mit dem Kennenlernen eines neuen technischen Systems.
Es genügt mir, dass ein technisches System funktioniert, mir ist es egal, wie oder warum.
Ich versuche zu verstehen, wie ein technisches System genau funktioniert.
Es genügt mir, die Grundfunktionen eines technischen Systems zu kennen.
Ich versuche, die Möglichkeiten eines technischen Systems vollständig auszunutzen.
Insgesamt bewerte ich Online-Klausuren als... ***[Bipolar Rating Scale from 1 to 7]***
behindernd | unterstützend
kompliziert | einfach
ineffizient | effizient
verwirrend | übersichtlich
langweilig | spannend
uninteressant | interessant
konventionell | originell
herkömmlich | neuartig

Table 27. Questionnaire Studies I and II - Part 3

Appendix B: Guideline Follow-Up Interview Study II

Klausursicherheit:
Gab es vor und während der Klausuren eine Art von Identitäts- und Betrugsüberprüfungen? Wenn ja: → *Welche?* → *Wie nützlichen waren diese?* → *Welche Maßnahmen würden Sie bei Onlineklausuren in Zukunft als zielführend erachten?*
Klausurdurchführung:
Wie wurden die Onlineklausuren bei Ihnen durchgeführt (Download/ILIAS/Sonstige)? → *Wie geeignet finden Sie die belegten Durchführungsarten?*
Wissensabfrage und Aufgabengestaltung:
Welche Wissensarten wurden abgefragt (Anwendung, Faktenwissen, Transfer)? ***Welche Aufgabentypen wurden eingesetzt (offen/geschlossen)?*** ***Waren Hilfsmittel erlaubt (max. Open-Book)?*** → *Wie geeignet fanden Sie die Aufgabentypen und die Wissensabfrage in Onlineklausuren?* → *Inwiefern hat sich die Klausurgestaltung von Präsenzklausuren unterschieden?* → *Was würden Sie in Zukunft bei Onlineklausuren ändern?*
Betrug:
Glauben Sie, dass bei Onlineprüfungen mehr Studierende betrügen als bei Präsenzprüfungen? Wenn ja: → *Woran könnte das liegen (eingeschränkte Aufsicht, sonst Nachteil)?* → *Wie könnte der Betrug bei Onlineklausuren in Zukunft reduziert werden?*
Mental:
Hatten Sie während der Onlineklausur Probleme sich zu konzentrieren? Wenn ja: → *Woran lag dies?* → *Wie sind Sie damit umgegangen?* → *Gab/Gibt es Möglichkeiten, diese Ursachen zu umgehen?* ***Hatten Sie während der Onlineklausur mehr Stress als in Präsenzklausuren? Wenn ja:*** → *Woran lag dies?* → *Wie sind Sie damit umgegangen?* → *Gab/Gibt es Möglichkeiten, diese Ursachen zu umgehen?*
Fair Grading:
Haben Sie Onlineprüfungen im Vergleich zu Präsenzklausuren als schwieriger empfunden? → *Haben Sie sich vor Onlineprüfungen gut vorbereitet gefühlt?* → *Haben Sie sich anders auf Onlineprüfungen vorbereitet als vor Präsenzklausuren?* ***Wie fair empfanden Sie die Benotung der Onlineklausuren?*** → *Wurde der tatsächliche Wissensstand zu einem Thema im Allgemeinen widergespiegelt?* → *Spielgelt die Note den tatsächlichen Wissensstand in Relation zu den Kommilitonen wider?*
Itention to Participate:
Wollten Sie die Prüfung während der Pandemie lieber als Online- anstatt einer Präsenzprüfung schreiben? → *Warum ja/nein?* ***Würden Sie nach Pandemie weiterhin an Onlineprüfungen teilnehmen wollen?*** → *Warum ja/nein?*
Sonstige Punkte:
Haben Sie sonstige Anmerkungen zu Onlinekausuren, die Ihnen während der Durchführung aufgefallen sind?

Table 28. Guideline Follow-Up Interview Study II

Appendix C: Questionnaire Study III

<table>
<tr><td colspan="3"><b>Bitte geben Sie Ihr Alter in Jahren an:</b>
<b><i>[Free Text]</i></b></td></tr>
<tr><td colspan="3"><i>[Free-text field]</i></td></tr>
<tr><td colspan="3"><b>Bitte geben Sie Ihr Geschlecht an.</b>
<b><i>[Single Choice]</i></b></td></tr>
<tr><td><i>männlich</i></td><td><i>weiblich</i></td><td><i>divers</i></td></tr>
<tr><td colspan="3"><b>Situational introduction (animated video)</b></td></tr>
<tr><td colspan="3">Sample photos:

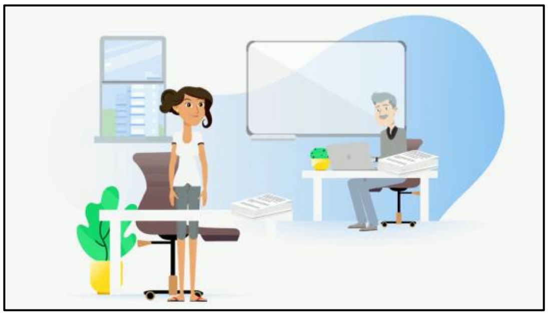

<b>Text:</b> <i>Stellen Sie sich vor, Sie studieren an einer Hochschule. Am Ende eines jeden Semesters müssen Sie an Pflichtklausuren teilnehmen, die Sie bestehen müssen. Wenn Sie diese Klausuren nicht bestehen, dann dürfen Sie Ihr Studium nicht fortsetzen.</i>
<i>Die Beantwortung der einzelnen Klausuraufgaben erfolgt ausschließlich in Form kurzer Textantworten, die zwischen einer viertel und einer halben Seite lang sind. Die Antworten beinhalten hierbei die Wiedergabe und den Transfer der gelernten Inhalte sowie das Erläutern anhand von Beispielen.</i>
<i>Die Korrektur erfolgt bisher im 4-Augenprinzip. Dies bedeutet, dass ein Mitarbeiter/eine Mitarbeiterin die Klausuren vorkorrigiert und der Professor/die Professorin diese Vorkorrektur noch einmal überprüft. Daher kann die Korrektur in großen Veranstaltungen mehrere Wochen dauern.</i></td></tr>
<tr><td colspan="3"><b>Bitte geben Sie für jede der folgenden Aussagen an, inwieweit Sie zustimmen.</b>
<b><i>[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]</i></b></td></tr>
<tr><td colspan="3"><i>Ich vertraue dem beschriebenen Korrekturprozess von Klausuren.</i></td></tr>
<tr><td colspan="3"><i>Ich hätte gerne die Möglichkeit, die korrigierenden Personen während der Korrekturdurchführung im Auge zu behalten.</i></td></tr>
<tr><td colspan="3"><i>Die Einsicht der finalen Klausurkorrektur durch die Prüfungsteilnehmer / Prüfungsteilnehmerinnen ist zwingend notwendig, um korrigierende Personen zu kontrollieren.</i></td></tr>
<tr><td colspan="3"><i>Personen, die Klausuren korrigieren, sollten stärker kontrolliert werden.</i></td></tr>
<tr><td colspan="3"><i>Für korrigierende Personen stehen die eigenen Interessen (z. B. ein möglichst geringer Korrekturaufwand) bei der Korrektur im Vordergrund.</i></td></tr>
<tr><td colspan="3"><b>Wie hoch schätzen Sie die Korrekturgenauigkeit (zwischen 0 und 100 %) in Prüfungsleistungen ein, wenn diese von Menschen durchgeführt wird:</b>
<b><i>[Free Text]</i></b></td></tr>
<tr><td colspan="3"><i>[Free-text field]</i></td></tr>
<tr><td colspan="3"><b>Bitte geben Sie für jede der folgenden Aussagen an, inwieweit Sie zustimmen.</b>
<b><i>[Likert Scale from „1 = Stimme überhaupt nicht zu“ to „5 = Stimme voll und ganz zu“]</i></b></td></tr>
<tr><td colspan="3"><i>Ich neige dazu, andere zu kritisieren.</i></td></tr>
<tr><td colspan="3"><i>Ich bin nachsichtig, vergebe anderen leicht.</i></td></tr>
<tr><td colspan="3"><i>Ich bin anderen gegenüber misstrauisch.</i></td></tr>
<tr><td colspan="3"><i>Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.</i></td></tr>
<tr><td colspan="3"><i>Ich bleibe auch in stressigen Situationen gelassen.</i></td></tr>
<tr><td colspan="3"><i>Ich reagiere leicht angespannt.</i></td></tr>
<tr><td colspan="3"><i>Ich mache mir oft Sorgen.</i></td></tr>
<tr><td colspan="3"><i>Ich werde selten nervös und unsicher.</i></td></tr>
</table>

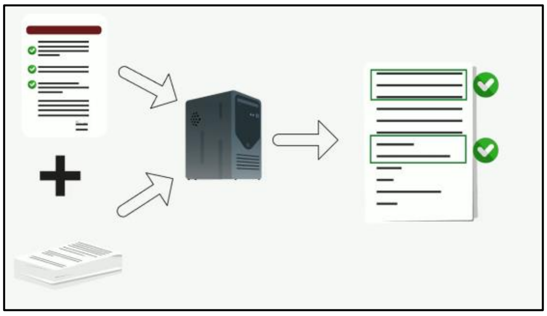
Table 29. Questionnaire Study III - Part1

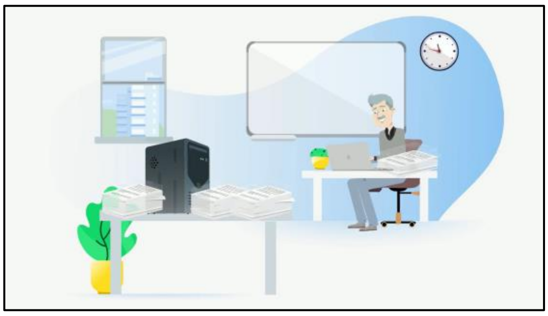
Im Folgenden geht es um Ihre Interaktion mit technischen Systemen. Mit ‚technischen Systemen‘ sind sowohl Apps und andere Software-Anwendungen als auch komplette digitale Geräte (z. B. Handy, Computer, Fernseher, Auto-Navigation) gemeint. Bitte geben Sie den Grad Ihrer Zustimmung zu folgenden Aussagen an. ***[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]***
Ich beschäftige mich gern genauer mit technischen Systemen.
Ich probiere gern die Funktionen neuer technischer Systeme aus.
In erster Linie beschäftige ich mich mit technischen Systemen, weil ich muss.
Wenn ich ein neues technisches System vor mir habe, probiere ich es intensiv aus.
Ich verbringe sehr gern Zeit mit dem Kennenlernen eines neuen technischen Systems.
Es genügt mir, dass ein technisches System funktioniert, mir ist es egal, wie oder warum.
Ich versuche zu verstehen, wie ein technisches System genau funktioniert.
Es genügt mir, die Grundfunktionen eines technischen Systems zu kennen.
Ich versuche, die Möglichkeiten eines technischen Systems vollständig auszunutzen.
Bitte geben Sie an, wie häufig Sie die folgenden Dienste nutzen: ***[Scale: „1 = Nie“, „2 = Täglich“, „3 = Fast täglich“, „4 = An 2 bis 3 Tagen pro Woche“, „5 = Ungefähr einmal pro Woche“ and „6 = Ungefähr einmal pro Monat“]***
Sprachassistenten *(z. B. Siri, Amazon Alexa etc.)*
Gesichtserkennung *(z. B. Entsperren des Smartphones)*
Individuelle Empfehlungen *(z. B. Kaufempfehlungen oder Musikvorschläge)*
Bitte geben Sie für jede der folgenden Aussagen an, inwieweit Sie zustimmen. ***[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]***
Bei unbekannten Personen sollte man sehr vorsichtig sein.
Bei den meisten Menschen kann man sich darauf verlassen, dass sie tun, was sie versprechen.
Heutzutage muss man vorsichtig sein, sonst wird man leicht ausgenutzt.
Die meisten Menschen sind ehrlich.
Die Korrektur von Prüfungsleistungen ist anspruchsvoll.
Für die Korrektur von Prüfungsleistungen braucht man ein Fachwissen, dass das Wissen zum reinen Beantworten einer Aufgabe übersteigt.
Die optimale Lösung einer Prüfungsleistungen ist immer eindeutig.
Bei der Bewertung von Prüfungsleistungen treten selten Fehler auf.
Die korrekte Bewertung von Prüfungsleistungen ist sehr wichtig.
Die Note in einer Prüfungsleistung hat langfristigen Einfluss auf das Leben der Studierenden.
Mir sind gute Noten wichtig.
Wenn ich eine schlechtere Note bekomme, als ich erwartet habe, denke ich nicht mehr lange daran.

Table 30. Questionnaire Study III - Part 2

Intervention: One AI scenario per participant (animated video)

Sample photos:





[Scenario 1 – Semi-automatic scoring]

Durch den Einsatz von künstlicher Intelligenz (auch KI genannt) kann dieser Korrekturvorgang auf wenige Tage verkürzt werden. ***Hierbei wird die Vorkorrektur durch die KI durchgeführt, während der Professor bzw. die Professorin diese Vorkorrektur nur noch stichprobenartig überprüft.*** *Für die Korrektur nutzt die KI einen vorher definierten Erwartungshorizont, der die Musterlösungen beinhaltet. Zudem greift das System auf vorherige Klausurkorrekturen zurück. Hierbei lernt das System, welche Antwortalternativen in Vergangenheit ebenfalls teilweise oder vollständig als korrekt bewertet wurden und wendet dieses Wissen auf die aktuelle Korrektur an. Der Vergleich findet dabei nicht nur stumpf auf Wortbasis statt, sondern berücksichtigt auch Synonyme, Wortkombinationen und Verneinungen. Hierdurch kann eine Überprüfung der inhaltlichen Korrektheit auch über Sätze hinweg gewährleistet werden. Zwar können der KI Fehler unterlaufen, allerdings haben Studien gezeigt, dass insbesondere bei hohem Korrekturaufwand auch menschlichen Prüferinnen und Prüfern in vergleichbarem Umfang Fehler unterlaufen.*

[Scenario 2 – automatic scoring]

Durch den Einsatz von künstlicher Intelligenz (auch KI genannt) kann dieser Korrekturvorgang auf wenige Tage verkürzt werden. ***Die KI übernimmt hierbei die vollständige Korrektur und Notenvergabe, sodass keine menschliche Nachkorrektur durch den Professor bzw. die Professorin stattfindet.*** *Für die Korrektur nutzt die KI einen vorher definierten Erwartungshorizont, der die Musterlösungen beinhaltet. Zudem greift das System auf vorherige Klausurkorrekturen zurück. Hierbei lernt das System, welche Antwortalternativen in Vergangenheit ebenfalls teilweise oder vollständig als korrekt bewertet wurden und wendet dieses Wissen auf die aktuelle Korrektur an. Der Vergleich findet dabei nicht nur stumpf auf Wortbasis statt, sondern berücksichtigt auch Synonyme, Wortkombinationen und Verneinungen. Hierdurch kann eine Überprüfung der inhaltlichen Korrektheit auch über Sätze hinweg gewährleistet werden. Zwar können der KI Fehler unterlaufen, allerdings haben Studien gezeigt, dass insbesondere bei hohem Korrekturaufwand auch menschlichen Prüferinnen und Prüfern in vergleichbarem Umfang Fehler unterlaufen.*

Bitte geben Sie für jede der folgenden Aussagen an, inwieweit Sie zustimmen. ***[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]***
Ich glaube, dass ein KI-basiertes Korrektursystem in meinem besten Interesse eingesetzt werden würde.
Das KI-basierte Korrektursystem kümmert sich um meine Interessen und nicht nur um die des Korrektors / der Korrektorin.
Während der KI-basierten Korrektur findet eine Bevorzugung einzelner Prüfungsteilnehmer / Prüfungsteilnehmerinnen statt.
Das KI-basierte Korrektursystem gewährleistet eine faire Bewertung der individuellen Leistung von Prüfungsteilnehmern / Prüfungsteilnehmerinnen.
Das KI-basierte Korrektursystem arbeitet zuverlässig.
Das KI-basierte Korrektursystem bewertet vergleichbare Klausurantworten unterschiedlicher Klausurteilnehmer / Klausurteilnehmerinnen gleich.
Ich kann mich darauf verlassen, dass das KI-basierte Korrektursystem fehlerfrei funktioniert.
Das KI-basierte Korrektursystem bewertet die Klausurantworten widerspruchsfrei.

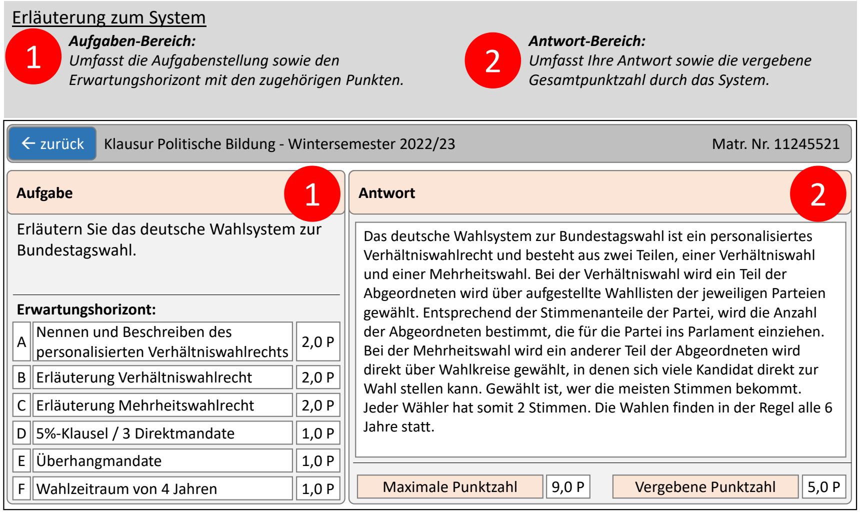
Table 31. Questionnaire Study III - Part 3

<table>
<tr><td colspan="2"><b>Bitte geben Sie für jede der folgenden Aussagen an, inwieweit Sie zustimmen.</b>
<b><i>[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]</i></b></td></tr>
<tr><td colspan="2"><i>Das KI-basierte System verfügt über fundierte Kenntnisse zur Korrektur von Klausuren.</i></td></tr>
<tr><td colspan="2"><i>Die Korrekturergebnisse des KI-basierten Korrektursystems sind genauso gut wie die einer hochkompetenten Person.</i></td></tr>
<tr><td colspan="2"><i>Das KI-basierte Korrektursystem bewertet die von mir abgegebenen Klausurantworten korrekt.</i></td></tr>
<tr><td colspan="2"><i>Das KI-basierte Korrektursystem nutzt das gesamte Wissen und die Informationen, die ihm zur Verfügung stehen, um eine Klausur zu korrigieren.</i></td></tr>
<tr><td colspan="2"><i>Der KI-basierte Korrekturprozess ist vertrauenswürdig.</i></td></tr>
<tr><td colspan="2"><i>Ich würde einen oder mehrere Aspekte des Korrekturverfahrens ändern, um den KI-Einsatz vertrauenswürdig zu machen.</i></td></tr>
<tr><td colspan="2"><i>Der KI-basierte Korrekturprozess wird zu einem fairen Ergebnis für die bewerteten Studierenden führen.</i></td></tr>
<tr><td colspan="2"><i>Prüfungsteilnehmer / Prüfungsteilnehmerinnen brauchen mehr Informationen darüber, wie das KI-basierte Korrektursystem bewertet, um dem Korrekturprozess vertrauen zu können.</i></td></tr>
<tr><td colspan="2"><i>Ich vertraue dem KI-basierten Korrekturprozess von Klausuren.</i></td></tr>
<tr><td colspan="2"><i>Ich hätte gerne die Möglichkeit, das KI-basierte Korrekturverfahren während der Korrekturdurchführung im Auge zu behalten.</i></td></tr>
<tr><td colspan="2"><i>Die Einsicht der finalen Klausurkorrekturen durch Prüfungsteilnehmer / Prüfungsteilnehmerinnen sind zwingend notwendig, um die KI-basierte Korrektur zu kontrollieren.</i></td></tr>
<tr><td colspan="2"><i>Das KI-basierte Korrektursystem sollte stärker kontrolliert werden.</i></td></tr>
<tr><td colspan="2"><i>Für KI-basierte Korrektursysteme stehen die eigenen Interessen (z. B. ein möglichst geringer Korrekturaufwand) bei der Korrektur im Vordergrund.</i></td></tr>
<tr><td colspan="2"><b>Wie hoch schätzen Sie die Korrekturgenauigkeit (zwischen 0 und 100 %) in Prüfungsleistungen ein, wenn diese mit dem beschriebenen KI-basierten Prozess durchgeführt wird:</b>
<b><i>[Free Text]</i></b></td></tr>
<tr><td colspan="2"><i>[Free-text field]</i></td></tr>
<tr><td colspan="2"><b>Würden Sie häufiger an einer Klausureinsicht teilnehmen, wenn das KI-basierte System eingesetzt wird:</b>
<b><i>[Single Choice]</i></b></td></tr>
<tr><td><i>Ja</i></td><td><i>Nein</i></td></tr>
<tr><td colspan="2"><b>Haben Sie weitere Anmerkungen zur KI-basierten Klausurkorrektur:</b>
<b><i>[Free Text]</i></b></td></tr>
<tr><td colspan="2"><i>[Free-text field]</i></td></tr>
</table>

Table 32. Questionnaire Study III - Part 4

Appendix D: Questionnaire Study V

Situational introduction (basic UI design)



Erläuterung zum System

Aufgaben-Bereich:
Umfasst die Aufgabenstellung sowie den Erwartungshorizont mit den zugehörigen Punkten.

Antwort-Bereich:
Umfasst Ihre Antwort sowie die vergebene Gesamtpunktzahl durch das System.

Klausur Politische Bildung - Wintersemester 2022/23 Matr. Nr. 11245521

Aufgabe

Erläutern Sie das deutsche Wahlsystem zur Bundestagswahl.

Erwartungshorizont:

A	Nennen und Beschreiben des personalisierten Verhältniswahlrechts	2,0 P
B	Erläuterung Verhältniswahlrecht	2,0 P
C	Erläuterung Mehrheitswahlrecht	2,0 P
D	5%-Klausel / 3 Direktmandate	1,0 P
E	Überhangmandate	1,0 P
F	Wahlzeitraum von 4 Jahren	1,0 P

Antwort

Das deutsche Wahlsystem zur Bundestagswahl ist ein personalisiertes Verhältniswahlrecht und besteht aus zwei Teilen, einer Verhältniswahl und einer Mehrheitswahl. Bei der Verhältniswahl wird ein Teil der Abgeordneten wird über aufgestellte Wahllisten der jeweiligen Parteien gewählt. Entsprechend der Stimmenanteile der Partei, wird die Anzahl der Abgeordneten bestimmt, die für die Partei ins Parlament einziehen. Bei der Mehrheitswahl wird ein anderer Teil der Abgeordneten wird direkt über Wahlkreise gewählt, in denen sich viele Kandidat direkt zur Wahl stellen kann. Gewählt ist, wer die meisten Stimmen bekommt. Jeder Wähler hat somit 2 Stimmen. Die Wahlen finden in der Regel alle 6 Jahre statt.

Maximale Punktzahl	9,0 P	Vergebene Punktzahl	5,0 P

Bitte geben Sie für jede der folgenden Aussagen an, inwieweit Sie zustimmen.
Durch die vom KI-basierten Korrektursystem bereitgestellte Darstellung der Korrekturergebnisse weiß ich, dass das System…
[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]

…zu meinem besten Interesse eingesetzt werden wird.

… sich um meine Interessen und nicht nur um die des Korrektors / der Korrektorin kümmert.

… eine faire Bewertung der Leistung von Prüfungsteilnehmern / Prüfungsteilnehmerinnen gewährleistet.

… die Klausurantworten zuverlässig bewertet.

… die Klausurantworten fehlerfrei bewertet.

… die Klausurantworten widerspruchsfrei bewertet.

… über fundierte Kenntnisse zur Klausurkorrektur verfügt.

… genauso gut wie eine hochkompetente Person korrigiert.

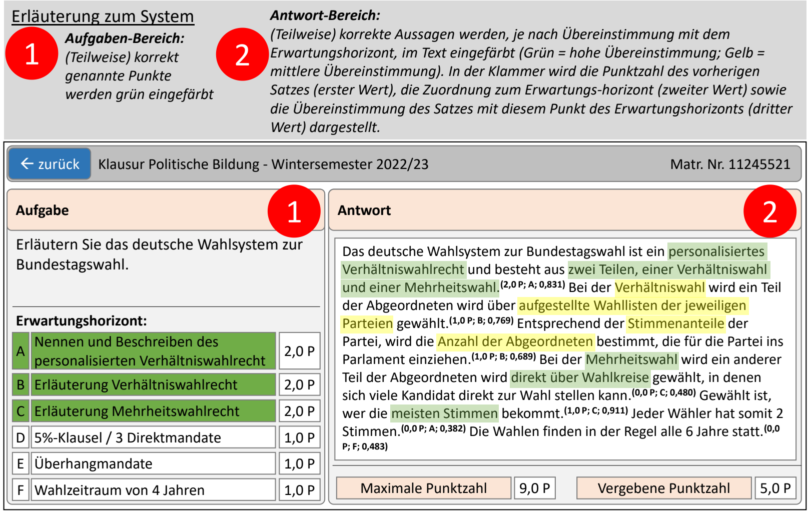
… die von mir abgegebenen Klausurantworten korrekt bewertet.

… die Klausurantworten vollständig nachvollziehbar bewertet.

… vertrauenswürdig ist.

Intervention: One modified UI design per participant

Sample photo:



Erläuterung zum System

Aufgaben-Bereich:
(Teilweise) korrekt genannte Punkte werden grün eingefärbt

Antwort-Bereich:
(Teilweise) korrekte Aussagen werden, je nach Übereinstimmung mit dem Erwartungshorizont, im Text eingefärbt (Grün = hohe Übereinstimmung; Gelb = mittlere Übereinstimmung). In der Klammer wird die Punktzahl des vorherigen Satzes (erster Wert), die Zuordnung zum Erwartungs-horizont (zweiter Wert) sowie die Übereinstimmung des Satzes mit diesem Punkt des Erwartungshorizonts (dritter Wert) dargestellt.

Klausur Politische Bildung - Wintersemester 2022/23 Matr. Nr. 11245521

Aufgabe

Erläutern Sie das deutsche Wahlsystem zur Bundestagswahl.

Erwartungshorizont:

A	Nennen und Beschreiben des personalisierten Verhältniswahlrecht	2,0 P
B	Erläuterung Verhältniswahlrecht	2,0 P
C	Erläuterung Mehrheitswahlrecht	2,0 P
D	5%-Klausel / 3 Direktmandate	1,0 P
E	Überhangmandate	1,0 P
F	Wahlzeitraum von 4 Jahren	1,0 P

Antwort

Das deutsche Wahlsystem zur Bundestagswahl ist ein personalisiertes Verhältniswahlrecht und besteht aus zwei Teilen, einer Verhältniswahl und einer Mehrheitswahl.$^{\text{(2,0 P; A; 0,831)}}$ Bei der Verhältniswahl wird ein Teil der Abgeordneten wird über aufgestellte Wahllisten der jeweiligen Parteien gewählt.$^{\text{(1,0 P; B; 0,769)}}$ Entsprechend der Stimmenanteile der Partei, wird die Anzahl der Abgeordneten bestimmt, die für die Partei ins Parlament einziehen.$^{\text{(1,0 P; B; 0,689)}}$ Bei der Mehrheitswahl wird ein anderer Teil der Abgeordneten wird direkt über Wahlkreise gewählt, in denen sich viele Kandidat direkt zur Wahl stellen kann.$^{\text{(0,0 P; C; 0,480)}}$ Gewählt ist, wer die meisten Stimmen bekommt.$^{\text{(1,0 P; C; 0,911)}}$ Jeder Wähler hat somit 2 Stimmen.$^{\text{(0,0 P; A; 0,382)}}$ Die Wahlen finden in der Regel alle 6 Jahre statt.$^{\text{(0,0 P; F; 0,483)}}$

Maximale Punktzahl	9,0 P	Vergebene Punktzahl	5,0 P

Table 33. Questionnaire Study V - Part 1

<table>
<tr><td colspan="6"><b>Bitte geben Sie für jede der folgenden Aussagen an, inwieweit Sie zustimmen.</b>
<b>Durch die vom KI-basierten Korrektursystem bereitgestellte Darstellung der Korrekturergebnisse weiß ich, dass das System…</b>
<b><i>[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]</i></b></td></tr>
<tr><td colspan="6"><i>...zu meinem besten Interesse eingesetzt werden wird.</i></td></tr>
<tr><td colspan="6"><i>… sich um meine Interessen und nicht nur um die des Korrektors / der Korrektorin kümmert.</i></td></tr>
<tr><td colspan="6"><i>… eine faire Bewertung der Leistung von Prüfungsteilnehmern / Prüfungsteilnehmerinnen gewährleistet.</i></td></tr>
<tr><td colspan="6"><i>… die Klausurantworten zuverlässig bewertet.</i></td></tr>
<tr><td colspan="6"><i>… die Klausurantworten fehlerfrei bewertet.</i></td></tr>
<tr><td colspan="6"><i>… die Klausurantworten widerspruchsfrei bewertet.</i></td></tr>
<tr><td colspan="6"><i>… über fundierte Kenntnisse zur Klausurkorrektur verfügt.</i></td></tr>
<tr><td colspan="6"><i>… genauso gut wie eine hochkompetente Person korrigiert.</i></td></tr>
<tr><td colspan="6"><i>… die von mir abgegebenen Klausurantworten korrekt bewertet.</i></td></tr>
<tr><td colspan="6"><i>… die Klausurantworten vollständig nachvollziehbar bewertet.</i></td></tr>
<tr><td colspan="6"><i>… vertrauenswürdig ist.</i></td></tr>
<tr><td colspan="6"><b>Fallen Ihnen weitere Informationen bzw. Gestaltungsmöglichkeiten ein, die die Nachvollziehbarkeit des Korrekturergebnisses verbessern könnten?</b>
<b><i>[Free Text]</i></b></td></tr>
<tr><td colspan="6"><i>[Free-text field]</i></td></tr>
<tr><td colspan="6"><b>Bitte geben Sie Ihr Alter in Jahren an:</b>
<b><i>[Free Text]</i></b></td></tr>
<tr><td colspan="6"><i>[Free-text field]</i></td></tr>
<tr><td colspan="6"><b>Bitte geben Sie Ihr Geschlecht an:</b>
<b><i>[Single Choice]</i></b></td></tr>
<tr><td colspan="2"><i>männlich</i></td><td colspan="2"><i>weiblich</i></td><td colspan="2"><i>divers</i></td></tr>
<tr><td colspan="6"><b>Bitte geben Sie Ihre Hauptfakultät an:</b>
<b><i>[Single Choice]</i></b></td></tr>
<tr><td colspan="3"><i>Fakultät für Agrarwissenschaften</i></td><td colspan="3"><i>Juristische Fakultät</i></td></tr>
<tr><td colspan="3"><i>Fakultät für Biologie und Psychologie</i></td><td colspan="3"><i>Philosophische Fakultät</i></td></tr>
<tr><td colspan="3"><i>Fakultät für Chemie</i></td><td colspan="3"><i>Sozialwissenschaftliche Fakultät</i></td></tr>
<tr><td colspan="3"><i>Fakultät für Forstwissenschaften und Waldökologie</i></td><td colspan="3"><i>Theologische Fakultät</i></td></tr>
<tr><td colspan="3"><i>Fakultät für Geowissenschaften und Geographie</i></td><td colspan="3"><i>Universitätsmedizin</i></td></tr>
<tr><td colspan="3"><i>Fakultät für Mathematik und Informatik</i></td><td colspan="3"><i>Wirtschaftswissenschaftliche Fakultät</i></td></tr>
<tr><td colspan="3"><i>Fakultät für Physik</i></td><td colspan="3"><i>Sonstiges: [Free-text field]</i></td></tr>
<tr><td colspan="6"><b>Im Folgenden geht es um Ihre Interaktion mit technischen Systemen. Mit ‚technischen Systemen‘ sind sowohl Apps und andere Software-Anwendungen als auch komplette digitale Geräte (z. B. Handy, Computer, Fernseher, Auto-Navigation) gemeint. Bitte geben Sie den Grad Ihrer Zustimmung zu folgenden Aussagen an.</b>
<b><i>[Likert Scale from „1 = Stimmt gar nicht“ to „6 = Stimmt völlig“]</i></b></td></tr>
<tr><td colspan="6"><i>Ich beschäftige mich gern genauer mit technischen Systemen.</i></td></tr>
<tr><td colspan="6"><i>Ich probiere gern die Funktionen neuer technischer Systeme aus.</i></td></tr>
<tr><td colspan="6"><i>Es genügt mir, dass ein technisches System funktioniert, mir ist es egal, wie oder warum.</i></td></tr>
<tr><td colspan="6"><i>Es genügt mir, die Grundfunktionen eines technischen Systems zu kennen.</i></td></tr>
</table>

Table 34. Questionnaire Study V - Part 2

References

(Abdul, Ashraf et al. 2018): Abdul, Ashraf; Vermeulen, Jo; Wang, Danding; Lim, Brian Y.; Kankanhalli, Mohan: *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda*. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Montreal, Canada.* (2018). pp. 1–18.

(Adadi / Berrada 2018): Adadi, A.; Berrada, M.: *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*. In: *IEEE Access* 6 (2018). pp. 52138–52160.

(Adnan et al. 2022): Adnan, M.; Uddin, M. I.; Khan, E.; Alharithi, F. S.; Amin, S.; Alzahrani, A. A.: *Earliest Possible Global and Local Interpretation of Students' Performance in Virtual Learning Environment by Leveraging Explainable AI*. In: *IEEE Access* 10 (2022). pp. 129843–129864.

(Ajzen 1991): Ajzen, I.: *The Theory of Planned Behavior*. In: *Organizational behavior and human decision processes* 50 (1991) 2. pp. 179–211.

(Al-Maqbali / Raja Hussain 2022): Al-Maqbali, A. H.; Raja Hussain, R. M.: *The Impact of Online Assessment Challenges on Assessment Principles during COVID-19 in Oman*. In: *Journal of University Teaching and Learning Practice* 19 (2022) 2. pp. 73–91.

(Alruwais et al. 2018): Alruwais, N.; Wills, G.; Wald, M.: *Advantages and challenges of using e-assessment*. In: *International Journal of Information and Education Technology* 8 (2018) 1. pp. 34–37.

(Alsaady et al. 2020): Alsaady, I.; Gattan, H.; Zawawi, A.; Alghanmi, M.; Zakai, H.: *Impact of COVID-19 Crisis on Exam Anxiety Levels among Bachelor Level University Students*. In: *Mediterranean Journal of Social Sciences* 11 (2020) 5. pp. 33–39.

(Ashoori / Weisz 2019): Ashoori, M.; Weisz, J. D.: *In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes*. Available online: http://arxiv.org/pdf/1912.02675v1 (accessed on 2023-07-10).

(Attali / Burstein 2005): Attali, Y.; Burstein, J.: *Automated Essay Scoring With e-rater® v. 2.0.*, ETS Research Report Series - ETS RR-04-45. 2005.

(Attali / Burstein 2006): Attali, Y.; Burstein, J.: *Automated Essay Scoring With e-rater® V.2*. In: *Journal of Technology, Learning, and Assessment* 4 (2006) 3. pp. 1–31.

(Attali / Powers 2008): Attali, Y.; Powers, D.: *A Developmental Writing Scale*, ETS Research Report Series - ETS RR-08-19. 2008.

(Bach et al. 2022): Bach, N. von dem; Baum, M.; Blank, M.; Ehmann, K.; Güntürk-Kuhl, B.; Pfeiffer, S.; Samray, D.; Seegers, M.; Sevindik, U.; Steeg, S.: *Umgang mit technischem Wandel in Büroberufen: Aufgabenprofile, lebendiges Arbeitsvermögen und berufliche Mobilität*, Wissenschaftliche Diskussionspapiere - Bundesinstitut für Berufsbildung (BIBB). 2022.

(Backes / Cowan 2019): Backes, B.; Cowan, J.: *Is the pen mightier than the keyboard? The effect of online testing on measured student achievement*. In: *Economics of Education Review* 68 (2019). pp. 89–103.

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

(Bao et al. 2021): Bao, Y.; Cheng, X.; de Vreede, T.; de Vreede, G.-J.: *Investigating the relationship between AI and trust in human-AI collaboration*. In: *Proceedings of the 54th Hawaii International Conference on System Sciences. Honolulu, Hawaii.* (2021). pp. 607–616.

(Barredo Arrieta et al. 2020): Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; Herrera, F.: *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. In: *Information Fusion* 58 (2020). pp. 82–115.

(Becker 2015): Becker, K.-D.: *Arbeit in der Industrie 4.0–Erwartungen des Instituts für angewandte Arbeitswissenschaft e.V.* In: *Botthof, A. & Hartmann, E. A. (eds.): Zukunft der Arbeit in Industrie 4.0* (2015). pp. 23–29. Wiesbaden: SpringerVieweg.

(Bertelsmann Stiftung 2020): Bertelsmann Stiftung: *Digitalisierung durchdringt die gesamte Arbeitswelt*. Available online: https://www.bertelsmann-stiftung.de/de/themen/aktuelle-meldungen/2020/august/digitalisierung-durchdringt-die-gesamte-arbeitswelt#link-tab-170523-14 (accessed on 2023-07-10).

(Billings 1997): Billings, C. E.: *Aviation Automation: The Search for A Human-Centered Approach.* 1. ed., Mahwah, New Jersey 1997.

(Birenbaum et al. 1992): Birenbaum, M.; Tatsuoka, K. K.; Gutvirtz, Y.: *Effects of Response Format on Diagnostic Assessment of Scholastic Achievement*. In: *Applied psychological measurement* 16 (1992) 4. pp. 353–363.

(Bloom et al. 1956): Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R.: *Taxonomy of educational objectives: the classification of educational goals - Handbook I: Cognitive domain*, New York, US 1956.

(Boevé et al. 2015): Boevé, A. J.; Meijer, R. R.; Albers, C. J.; Beetsma, Y.; Bosker, R. J.: *Introducing Computer-Based Testing in High-Stakes Exams in Higher Education: Results of a Field Experiment*. In: *PloS one* 10 (2015) 12. 1-13.

(Bordt / Luxburg 2023): Bordt, S.; Luxburg, U. von: *ChatGPT Participates in a Computer Science Exam*. Available online: https://arxiv.org/abs/2303.09461 (accessed on 2023-07-10).

(Borrego-Díaz / Galán-Páez 2022): Borrego-Díaz, J.; Galán-Páez, J.: *Explainable Artificial Intelligence in Data Science - From Foundational Issues Towards Socio-technical Considerations*. In: *Minds and Machines* 32 (2022). pp. 485–531.

(Burlak et al. 2006): Burlak, G. N.; Hernandez, J.-A.; Ochoa, A.; Munoz, J.: *The Use of Data Mining to Determine Cheating in Online Student Assessment* (2006) Proceedings of the Electronics, Robotics and Automotive Mechanics Conference (CERMA'06). Cuernavaca, Mexico. pp. 161–166.

(Burstein et al. 1996): Burstein, J.; Frase, L. T.; Ginther, A.; Grant, L.: *Technologies for Language Assessment*. In: *Annual Review of Applied Linguistics* 16 (1996). pp. 240–260.

(Burstein et al. 2004): Burstein, J.; Chodorow, M.; Leacock, C.: *Automated Essay Evaluation: The Criterion Online Writing Service*. In: *AI-Magazine* 25 (2004) 3. pp. 27–36.

(Butler-Henderson / Crawford 2020): Butler-Henderson, K.; Crawford, J.: *A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity*. In: *Computers & Education* 159 (2020). pp. 1–12.

(Buyrukoglu et al. 2019): Buyrukoglu, S.; Batmaz, F.; Lock, R.: *Improving marking efficiency for longer programming solutions based on a semi-automated assessment approach*. In: *Computer Applications in Engineering Education* 27 (2019) 3. pp. 733–743.

(Candrlic et al. 2014): Candrlic, S.; Katić, M. A.; Dlab, M. H. H.: *Online vs. Paper-based testing: A comparison of test results*. In: *Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). Opatija, Croatia.* (2014). pp. 657–662.

(Cassady / Johnson 2002): Cassady, J. C.; Johnson, R. E.: *Cognitive Test Anxiety and Academic Performance*. In: *Contemporary Educational Psychology* 27 (2002) 2. pp. 270–295.

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

(Chen et al. 2010): Chen, Y.-Y.; Liu, C.-L.; Lee, C.-H.; Chang, T.-H.: *An Unsupervised Automated Essay-Scoring System*. In: *IEEE Intelligent-Systems* 25 (2010) 5. pp. 61–67.

(Chen et al. 2020): Chen, L.; Chen, P.; Lin, Z.: *Artificial Intelligence in Education: A Review*. In: *IEEE Access* 8 (2020). pp. 75264–75278.

(Chufama / Sithole 2021): Chufama, M.; Sithole, F.: *The Pivotal Role of Diagnostic, Formative and Summative Assessment in Higher Education Institutions' Teaching and Student Learning*. In: *International Journal of Multidisciplinary Research and Publications* 4 (2021) 5. pp. 5–15.

(Chung et al. 2014): Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y.: *Empirical evaluation of gated recurrent neural networks on sequence modeling*. Available online: https://arxiv.org/abs/1412.3555 (accessed on 2023-07-10).

(Clariana / Wallace 2002): Clariana, R.; Wallace, P.: *Paper–based versus computer–based assessment: key factors associated with the test mode effect*. In: *British Journal of Educational Technology* 33 (2002) 5. pp. 593–602.

(Cluskey Jr et al. 2011): Cluskey Jr, G. R.; Ehlen, C. R.; Raiborn, M. H.: *Thwarting online exam cheating without proctor supervision*. In: *Journal of Academic and Business Ethics* 4 (2011) 1. pp. 1–7.

(Cohen 1992): Cohen, J.: *A Power Primer*. In: *Psychological-Bulletin* 112 (1992) 1. pp. 155–159.

(Cohen et al. 2018): Cohen, Y.; Levi, E.; Ben-Simon, A.: *Validating Human and Automated Scoring of Essays Against "True" Scores*. In: *Applied Measurement in Education* 31 (2018) 3. pp. 241–250.

(Conijn et al. 2022): Conijn, R.; Kleingeld, A.; Matzat, U.; Snijders, C.: *The fear of big brother: The potential negative side-effects of proctored exams*. In: *Journal of Computer Assisted Learning* 38 (2022) 6. pp. 1–14.

(Costa; Jr. / McCrae 2000): Costa, P. T., Jr.; McCrae, R. R.: *Neo Personality Inventory*. In: *Kazdin, A. E. (eds.): Encyclopedia of Psychology* 5 (2000). pp. 407–409. Oxford: Oxford University Press.

(Crawford et al. 2020): Crawford, J.; Butler-Henderson, K.; Rudolph, J.; Malkawi, B.; Glowatz, M.; Burton, R.; Magni, P.; Lam, S.: *COVID-19: 20 countries' higher education intra-period digital pedagogy responses*. In: *Journal of Applied Learning & Teaching* 3 (2020) 1. pp. 9–28.

(Danner et al. 2019): Danner, D.; Rammstedt, B.; Bluemke, M.; Lechner, C.; Berres, S.; Knopf, T.; Soto, C. J.; John, O. P.: *Das Big Five Inventar 2*. In: *Diagnostica* 65 (2019) 3. pp. 121–132.

(Deloatch et al. 2016): Deloatch, R.; Bailey, B. P.; Kirlik, A.: *Measuring effects of modality on perceived test anxiety for computer programming exams*. In: *Proceedings of the 47th ACM Technical Symposium on Computing Science Education. Memphis, USA.* (2016). pp. 291–296.

(Dengler / Matthes 2015): Dengler, K.; Matthes, B.: *Folgen der Digitalisierung für die Arbeitswelt: Substituierbarkeitspotenziale von Berufen in Deutschland*, IAB-Forschungsbericht - Institut für Arbeitsmarkt- und Berufsforschung (IAB). 2015.

(Dengler / Matthes 2018): Dengler, K.; Matthes, B.: *Substituierbarkeitspotenziale von Berufen: Wenige Berufsbilder halten mit der Digitalisierung Schritt*, IAB-Kurzbericht - Institut für Arbeitsmarkt- und Berufsforschung (IAB). 2018.

(Dermo 2009): Dermo, J.: *e-Assessment and the student learning experience: A survey of student perceptions of e-assessment*. In: *British Journal of Educational Technology* 40 (2009) 2. pp. 203–214.

(Dieckmann 2021): Dieckmann, L.: *Prüfungsstress: Datenschutz bei Online-Klausuren*. In: *c't - magazin für computertechnik / Recht* (2021) 1. pp. 174–175.

(Elsalem et al. 2020): Elsalem, L.; Al-Azzam, N.; Jum'ah, A. A.; Obeidat, N.; Sindiani, A. M.; Kheirallah, K. A.: *Stress and behavioral changes with remote E-exams during the Covid-19 pandemic: A cross-sectional study among undergraduates of medical sciences*. In: *Annals of Medicine and Surgery* 60 (2020). pp. 271–279.

(Elson et al. 2021): Elson, J. S.; Derrick, D. C.; Merino, L. A.: *An Empirical Study Exploring Difference in Trust of Perceived Human and Intelligent System Partners*. In: *Proceedings of the 54th Hawaii International Conference on System Sciences. Honolulu, Hawaii.* (2021). pp. 136–145.

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

(Fettke 2006): Fettke, P.: *State-of-the-Art des State-of-the-Art*. In: *Wirtschaftsinformatik* 48 (2006) 4. pp. 257–266.

(Fluck et al. 2009): Fluck, A.; Pullen, D.; Harper, C.: *Case study of a computer based examination system*. In: *Australasian Journal of Educational Technology* 25 (2009) 4. pp. 509–523.

(Foltz et al. 1999): Foltz, P. W.; Laham, D.; Landauer, T. K.: *The Intelligent Essay Assessor: Applications to Educational Technology*. In: *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning* 1 (1999) 2. pp. 939–944.

(Franke et al. 2019): Franke, T.; Attig, C.; Wessel, D.: *A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale*. In: *International Journal of Human-Computer Interaction* 35 (2019) 6. pp. 456–467.

(Frohm et al. 2008): Frohm, J.; Lindström, V.; Winroth, M.; Stahre, J.: *Levels of automation in manufacturing*. In: *Ergonomia - an International Journal of Ergonomics and Human Factors* 30 (2008) 3. pp. 1–28.

(Fulcher 2000): Fulcher, G.: *Computers in Language Teaching*. In: *Brett, P. & Motteram, G. (eds.): A special interest in computers: Learning and Teaching with Information and Communications Technologies.* (2000). pp. 93–107.

(Gamage et al. 2020): Gamage, K. A. A.; Silva, E. K. de; Gunawardhana, N.: *Online Delivery and Assessment during COVID-19: Safeguarding Academic Integrity*. In: *Education Sciences* 10 (2020) 11. pp. 301–324.

(German Chamber of Commerce and Industry 2020): German Chamber of Commerce and Industry: *Digitale Kompetenzen – Wahrnehmung und Anspruch: Wie viel Digitalität können deutsche Arbeitnehmer und welche Inhalte werden von Bildungseinrichtungen erwartet?* Available online: https://www.ihk-digitalkompetenz.de/wp-content/uploads/2020/10/Digitalkompetenz_Check_ Studie_Download.pdf (accessed on 2023-07-10).

(Goldberg 1990): Goldberg, L. R.: *An alternative" description of personality":The Big-Five factor structure*. In: *Journal of Personality & Social Psychology* 59 (1990) 6. pp. 1216–1229.

(Gregor et al. 2020): Gregor, S.; Kruse, L.; Seidel, S.: *Research Perspectives: The Anatomy of a Design Principle*. In: *Journal of the Association for Information Systems* 21 (2020) 6. pp. 1622–1652.

(Gregor / Hevner 2013): Gregor, S.; Hevner, A. R.: *Positioning and Presenting Design Science Research for Maximum Impact*. In: *MIS Quarterly* 37 (2013) 2. pp. 337–355.

(Grissom et al. 2016): Grissom, S.; Murphy, L.; McCauley, R.; Fitzgerald, S.: *Paper vs. Computer-Based Exams: A Study of Errors in Recursive Binary Tree Algorithms*. In: *Proceedings of the 47th ACM Technical Symposium on Computing Science Education. Memphis, USA.* (2016). pp. 6–11.

(Gunning et al. 2019): Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z.: *XAI-Explainable artificial intelligence*. In: *Science Robotics* 4 (2019) 37.

(Hair et al. 2018): Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E.: *Multivariate Data Analysis.* 8. ed., Englewood Cliffs, NJ 2018.

(Haque et al. 2023): Haque, A. B.; Islam, A. N.; Mikalef, P.: *Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research*. In: *Technological Forecasting & Social Change* (2023) 186. pp. 1–19.

(Harlen 2005): Harlen, W.: *Teachers' summative practices and assessment for learning–tensions and synergies*. In: *Curriculum Journal* 16 (2005) 2. pp. 207–223.

(Harlen / James 1997): Harlen, W.; James, M.: *Assessment and Learning: differences and relationships between formative and summative assessment*. In: *Assessment in Education: Principles, Policy and Practice* 4 (1997) 3. pp. 365–379.

(Harmon / Lambrinos 2008): Harmon, O. R.; Lambrinos, J.: *Are online exams an invitation to cheat?* In: *The Journal of Economic Education* 39 (2008) 2. pp. 116–125.

(Hartmann et al. 2021): Hartmann, P.; Hobert, S.; Schumann, M.: *The Intention to Participate in Online Exams – The Student Perspective*. In: *Proceedings of the 27th Americas Conference on Information Systems. Montreal, Canada.* (2021). pp. 1–10.

(Hartmann et al. 2022a): Hartmann, P.; Holthoff, N.; Hobert, S.; Schumann, M.: *(AI)N'T NOBODY HELPING ME? – DESIGN AND EVALUATION OF A MACHINE-LEARNING-BASED SEMI-AUTOMATIC ESSAY SCORING SYSTEM*. In: *Proceedings of the 30th European Conference on Information Systems. Timisoara, Romania.* (2022). pp. 1–16.

(Hartmann et al. 2022b): Hartmann, P.; Hobert, S.; Schumann, M.: *Trust, but Verify! - An Empirical Investigation of Students' Initial Trust in AI-Based Essay Scoring*. In: *Proceedings of the 28th Americas Conference on Information Systems. Minneapolis, USA.* (2022). pp. 1–10.

(Hartmann / Hobert 2023a): Hartmann, P.; Hobert, S.: *Explain AI-Based Essay Scorings without XAI - Empirical Investigation of an User-Centered UI Design for AI-Based AES Systems*. In: *Proceedings of the 29th Americas Conference on Information Systems. Panama City, Panama.* (2023). pp. 1–10.

(Hartmann / Hobert 2023b): Hartmann, P.; Hobert, S.: *FROM EMERGENCY REMOTE ASSESSMENT TO A NEW STATUS QUO? – LESSONS LEARNED FROM ONLINE ASSESSMENTS DURING THE COVID-19 PANDEMIC*. In: *Proceedings of the 31st European Conference on Information Systems. Kristiansand, Norway.* (2023). pp. 1–15.

(Heghedus et al. 2019): Heghedus, C.; Chakravorty, A.; Rong, C.: *Neural Network Frameworks. Comparison on Public Transportation Prediction*. In: *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Rio de Janeiro, Brazil.* (2019). pp. 842–849.

(Hermann et al. 2017): Hermann, T.; Hirschle, S.; Kowol, D.; Rapp, J.; Resch, U.; Rothmann, J.: *Auswirkungen von Industrie 4.0 auf das Anforderungsprofil der Arbeitnehmer und die Folgen im Rahmen der Aus- und Weiterbildung*. In: *Andelfinger, V. P. & Hänisch, T. (eds.): Industrie 4.0: Wie cyber-physische Systeme die Arbeitswelt verändern* (2017). pp. 239–253. Wiesbaden: Springer Gabler.

(Hevner et al. 2004): Hevner, A. R.; March, S. T.; Park, J.; Ram, S.: *Design Science in Information Systems Research*. In: *MIS Quarterly* 28 (2004) 1. pp. 75–105.

(Hewlett / Kahl-Andresen 2014): Hewlett, C.; Kahl-Andresen, A.: *Prüfungsökonomie statt Prüfungsqualität?* In: *Berufsbildung in Wissenschaft und Praxis* 43 (2014) 3. pp. 6–9.

(Higgins et al. 2002): Higgins, C.; Symeonidis, P.; Tsintsifas, A.: *The marking system for CourseMaster*. In: *ACM SIGCSE Bulletin* 34 (2002) 3. pp. 46–50.

(Higgins et al. 2005): Higgins, C. A.; Gray, G.; Symeonidis, P.; Tsintsifas, A.: *Automated Assessment and Experiences of Teaching Programming*. In: *Journal on Educational Resources in Computing* 5 (2005) 3. 1-21.

(Hillier 2014): Hillier, M.: *The very idea of e-Exams: student (pre)conceptions*. In: *Rhetoric and Reality: Critical perspectives on educational technology* (2014). pp. 77–88.

(Hillier 2015): Hillier, M.: *e-Exams with student owned devices: Student voices*. In: *Proceedings of the International Mobile Learning Festival 2015. Hong Kong SAR, China.* (2015). pp. 582–608.

(Hillier / Fluck 2013): Hillier, M.; Fluck, A.: *Arguing again for e-exams in high stakes examinations*. In: *Proceedings of Electric Dreams (ascilite). Sydney, Australia.* (2013). pp. 385–396.

(Hobert 2019): Hobert, S.: *Say hello to 'coding tutor'! design and evaluation of a chatbot-based learning system supporting students to learn to program*. In: *Proceedings of the 40th International Conference on Information Systems. Munich, Germany.* (2019). pp. 1–17.

(Hobert et al. 2023): Hobert, S.; Groth, M.; Nießner, T.; Wilhelmi, L.: *How Today's AI Content Generators Outperform Average Novice Students in Information Systems Exams*. In: *Proceedings of the 29th Americas Conference on Information Systems. Panama City, Panama.* (2023). pp. 1–5.

(Huang / Feng 2019): Huang, J.; Feng, Y.: *Optimization of recurrent neural networks on natural language processing*. In: *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition. Beijing, China.* (2019). pp. 39–45.

(Hung et al. 1993): Hung, S.-L.; Kwok, I.-F.; Chan, R.: *Automatic programming assessment*. In: *Computers & Education* 20 (1993) 2. pp. 183–190.

(Ilgaz / Afacan Adanır 2020): Ilgaz, H.; Afacan Adanır, G.: *Providing online exams for online learners: Does it really matter for them?* In: *Education and Information Technologies* 25 (2020) 2. pp. 1255–1269.

(ILIAS 2021): ILIAS: *The Open Source Learning Management System*. Available online: https://www.ilias.de/en/ (accessed on 2023-07-10).

(ILIAS 2022): ILIAS: *The Open Source Learning Management System*. Available online: https://www.ilias.de/en/ (accessed on 2023-07-10).

(Impey / Formanek 2021): Impey, C.; Formanek, M.: *MOOCS and 100 Days of COVID: Enrollment surges in massive open online astronomy classes during the coronavirus pandemic*. In: *Social Sciences and Humanities Open* 4 (2021) 1. pp. 1–9.

(Jang et al. 2022): Jang, Y.; Choi, S.; Jung, H.; Kim, H.: *Practical early prediction of students' performance using machine learning and eXplainable AI*. In: *Education and Information Technologies* 27 (2022). pp. 12855–12889.

(Jeong 2014): Jeong, H.: *A comparative study of scores on computer-based tests and paper-based tests*. In: *Behaviour & Information Technology* 33 (2014) 4. pp. 410–422.

(Johar / Kumar 2016): Johar, S.; Kumar, U.: *Transforming Assessment: New Pedagogies for the Digital Age*. In: *Kumar, U. (eds.): The Wiley Handbook of Personality Assessment* (2016). pp. 399–414. Hoboken: Wiley-Blackwell.

(John et al. 2008): John, O. P.; Naumann, L. P.; Soto, C. J.: *Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues*. In: *John, O. P., Robins, R. W. & Pervin L. A. (eds.): Handbook of personality: Theory and research* (2008). pp. 114–158. New York: Guilford Press.

(Joint Committee on Standards for Educational and Psychological Testing 2014): Joint Committee on Standards for Educational and Psychological Testing: *Standards for Educational an Psychological Testing*, Washington D.C., US 2014.

(Joint Information Systems Committee 2007): Joint Information Systems Committee: *Effective Practive with e-Assessment - An overview of technologies, policies and practice in further and higher education* 2007.

(Ju et al. 2018): Ju, A.; Mehne, B.; Halle, A.; Fox, A.: *In-class coding-based summative assessments: tools, challenges, and experience*. In: *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education. Larnaca, Cyprus.* (2018). pp. 75–80.

(Kaggle 2012): Kaggle: *The Hewlett Foundation: Short Answer Scoring - Develop a scoring algorithm for student-written short-answer responses*. Available online: https://www.kaggle.com/c/asap-sas/overview (accessed on 2023-07-05).

(Kaiser 1974): Kaiser, H. F.: *An index of factorial simplicity*. In: *Psychometrika* 39 (1974). pp. 31–36.

(Kalogeropoulos et al. 2013): Kalogeropoulos, N.; Tzigounakis, I.; Pavlatou, E. A.; Boudouvis, A. G.: *Computer-based assessment of student performance in programing courses*. In: *Computer Applications in Engineering Education* 21 (2013) 4. pp. 671–683.

(Kasneci et al. 2023): Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; Kasneci, G.: *ChatGPT for good? On opportunities and challenges of large language models for education.* In: *Learning and Individual Differences* (2023) 103. pp. 1–9.

(Kaya / Özel 2015): Kaya, M.; Özel, S. A.: *Integrating an online compiler and a plagiarism detection tool into the Moodle distance education system for easy assessment of programming assignments*. In: *Computer Applications in Engineering Education* 23 (2015) 3. pp. 363–373.

(Kelly / Columbus 2020): Kelly, A. P.; Columbus, R.: *CHALLENGES FACING AMERICAN HIGHER EDUCATION*, College in the Time of Coronavirus - American Enterprise Institute (AEI). 2020.

(Kerr et al. 2013): Kerr, D.; Mousavi, H.; Iseli, M. R.: *Automatic Short Essay Scoring Using Natural Language Processing to Extract Semantic Information in the Form of Propositions*, Natioanl Center for Research on Evaluation, Standards, and Student Teaching (CRESST) - CRESST Report 831. 2013.

(Khosravi et al. 2022): Khosravi, H.; Buckingham Shum, S.; Chen, G.; Conati, C.; Tsai, Y.-S.; Kay, J.; Knight, S.; Martinez-Maldonado, R.; Sadiq, S.; Gašević, D.: *Explainable Artificial Intelligence in education*. In: *Computers and Education: Artificial Intelligence* 3 (2022). pp. 1–22.

(King et al. 2009): King, C. G.; Guyette Jr., R. W.; Piotrowski, C.: *Online Exams and Cheating: An Empirical Analysis of Business Students' Views*. In: *The Journal of Educators Online* 6 (2009) 1. pp. 1–11.

(Kleerekoper / Schofield 2019): Kleerekoper, A.; Schofield, A.: *The False-Positive Rate of Automated Plagiarism Detection for SQL Assessments*. In: *Proceedings of the 2019 Conference on United Kingdom & Ireland Computing Education Research. Canterbury, United Kingdom.* (2019). pp. 1–6.

(Kleinhans / Schumann 2015): Kleinhans, J.; Schumann, M.: *Increasing testing efficiency through the development of an IT-based adaptive testing tool for competency measurement*. In: *Interactive Technology and Smart Education* 12 (2015) 4. pp. 242–255.

(Kocielnik et al. 2019): Kocielnik, R.; Amershi, S.; Bennett, P. N.: *Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems*. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, Scotland.* (2019). pp. 1–14.

(Kumar / Boulanger 2020): Kumar, V.; Boulanger, D.: *Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value*. In: *Frontiers in Education* 5 (2020). pp. 1–22.

(Kung et al. 2023): Kung, T. H.; Cheatham, M.; Medenilla, A.; Sillos, C.; Leon, L. de; Elepano, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; Tseng, V.: *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models*. In: *PloS Digital Health* 2 (2023) 2. pp. 1–12.

(Küppers / Schroeder 2018): Küppers, B.; Schroeder, U.: *Students' Perceptions of e-Assessment*. In: *Proceedings of the Open Conference on Computers in Education 2018. Linz, Austria.* (2018). pp. 275–284.

(Lai / Liou 2010): Lai, C.-Y.; Liou, W.-C.: *Implementation of E-Learning and Corporate Performance: An Empirical Investigation*. In: *International Journal of Advanced Corporate Learning* 3 (2010) 1. pp. 4–10.

(Lajis / Aziz 2010): Lajis, A.; Aziz, N. A.: *NL scoring technique for the assessment of learners' understanding*. In: *Proceedings of the 2010 Second International Conference on Computer Research and Development. Kuala Lumpur, Malaysia.* (2010). pp. 379–383.

(Lajis / Aziz 2012): Lajis, A.; Aziz, N. A.: *NL scoring and Bloom competency test: an experimental result*. In: *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication. Kuala Lumpur, Malaysia.* (2012). pp. 1–5.

(Laubscher et al. 2005): Laubscher, R.; Olivier, M. S.; Venter, H. S.; Eloff, J. H. P.; Rabe, D. J.: *The Role of Key Loggers in Computer-based Assessment Forensics*. In: *Proceedings of the 2005 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries. White River, South Africa.* (2005). pp. 123–130.

(Laugwitz et al. 2008): Laugwitz, B.; Held, T.; Schrepp, M.: *Construction and Evaluation of a User Experience Questionnaire*. In: *Proceedings of the 4th Symposium of the Austrian HCI and Usability Engineering Group. Graz, Austria.* (2008). pp. 63–76.

(Lazarinis et al. 2010): Lazarinis, F.; Green, S.; Pearson, E.: *Creating personalized assessments based on learner knowledge and objectives in a hypermedia Web testing application*. In: *Computers & Education* 55 (2010) 4. pp. 1732–1743.

(Lee / See 2004): Lee, J. D.; See, K. A.: *Trust in Automation: Designing for Appropriate Reliance*. In: *Human factors* 46 (2004) 1. pp. 50–80.

(Li et al. 2008): Li, X.; Hess, T. J.; Valacich, J. S.: *Why do we trust new technology? A study of initial trust formation with organizational information systems*. In: *The Journal of Strategic Information Systems* 17 (2008) 1. pp. 39–71.

(Lo 2023): Lo, C. K.: *What is the Impact of ChatGPT on Education? A Rapid Review of the Literature*. In: *Education Sciences* 13 (2023) 4. pp. 410–425.

(Lombardi / Oblinger 2017): Lombardi, M. M.; Oblinger, D. G.: *Authentic learning for the 21st century: An overview*, EDUCAUSE: Learning Initiative - ELI Paper 1. 2017.

(Madsen / Gregor 2000): Madsen, M.; Gregor, S.: *Measuring Human-Computer Trust*. In: *Proceedings of the 11th Australasian Conference on Information Systems. Brisbane, Australia.* (2000). pp. 1–12.

(Maguire et al. 2010): Maguire, K. A.; Smith, D. A.; Brallier, S. A.; Palm, L. J.: *Computer-based testing: A comparison of computer-based and paper-and-pencil assessment*. In: *Academy of Educational Leadership Journal* 14 (2010) 4. pp. 117–125.

(March / Smith 1995): March, S. T.; Smith, G. F.: *Design and natural science research on information technology*. In: *Decision Support Systems* 15 (1995) 4. pp. 251–266.

(Marinoni et al. 2020): Marinoni, G.; Van’t Land, H.; Jensen, T.: *The impact of Covid-19 on higher education around the world*, IAU Global Survey Report - International Association of Universities. 2020.

(Matthíasdóttir / Arnalds 2016): Matthíasdóttir, Á.; Arnalds, H.: *e-assessment: students’ point of view*. In: *Proceedings of the 17th International Conference on Computer Systems and Technologies 2016. Palermo, Italy.* (2016). pp. 369–374.

(Mayer et al. 1995): Mayer, R. C.; Davis, J. H.; Schoorman, F. D.: *An Integrative Model of Organizational Trust*. In: *Academy of Management Review* 20 (1995) 3. pp. 709–734.

(McCabe 2005): McCabe, D. L.: *Cheating among college and university students: A North American perspective*. In: *International Journal for Educational Integrity* 1 (2005) 1.

(Miller 2011): Miller, C.: *Aesthetics and e-assessment: the interplay of emotional design and learner performance*. In: *Distance Education* 32 (2011) 3. pp. 307–337.

(Mitchell et al. 2003): Mitchell, T.; Aldridge, N.; Broomhead, P.: *Computerised Marking of Short-Answer Free-Text Responses*. In: *Proceedings of the 2003 IAEA Conference. Manchester, UK.* (2003).

(Mogey / Fluck 2015): Mogey, N.; Fluck, A.: *Factors influencing student preference when comparing handwriting and typing for essay style examinations*. In: *British Journal of Educational Technology* 46 (2015) 4. pp. 793–802.

(Mohler / Mihalcea 2009): Mohler, M.; Mihalcea, R.: *Text-to-text Semantic Similarity for Automatic Short Answer Grading*. In: *Proceedings of the 12th Conference of the European Chapter of the ACL. Athens, Greece.* (2009). pp. 567–575.

(Mohri et al. 2018): Mohri, M.; Rostamizadeh, A.; Talwalkar, A.: *Foundations of Machine Learning.* 2. ed., Cambridge, MA, US 2018.

(Nicholson 2007): Nicholson, P.: *A History of E-Learning: Echoes of the pioneers*. In: *Fernández-Manjón, B., Sánchez-Pérez, J. M., Gómez-Pulido, J. A., Vega-Rodríguez, M. A. & Bravo-Ródriguez, J. (eds.): Computers and Education: E-learning, From Theory to Practice* (2007). pp. 1–11. Dordrecht: Springer.

(Nielsen 1994): Nielsen, J.: *Enhancing the Explanatory Power of Usability Heuristics*. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Boston, US.* (1994). pp. 152–158.

(Noorbehbahani et al. 2022): Noorbehbahani, F.; Mohammadi, A.; Aminazadeh, M.: *A systematic review of research on cheating in online exams from 2010 to 2021*. In: *Education and Information Technologies* 27 (2022). pp. 8413–8460.

(Ocak / Karakus 2021): Ocak, G.; Karakus, G.: *Undergraduate students’ views of and difficulties in online exams during the COVID-19 pandemic.* In: *Themes in eLearning* 14 (2021) 14. pp. 13–30.

(OECD 2020): OECD: *Remote Online Exams in Higher Education During the COVID-19 Crisis*, OEC Education Policy Perspectives. 2020.

(Oleson et al. 2011): Oleson, K. E.; Billings, D. R.; Kocsis, V.; Chen, J. Y. C.; Hancock, P. A.: *Antecedents of trust in human-robot collaborations*. In: *Proceedings of the 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support. Miami Beach, US.* (2011). pp. 175–178.

(OpenAI 2023): OpenAI: *ChatGPT: Optimizing Language Models for Dialogue*. Available online: https://openai.com/blog/chatgpt/ (accessed on 2022-07-10).

(Opgen-Rhein et al. 2018): Opgen-Rhein, J.; Küppers, B.; Schroeder, U.: *An application to discover cheating in digital exams*. In: *Proceedings of the 18th Koli Calling International Conference on Computing Education Research. Koli, Finland.* (2018). pp. 1–5.

(Owunwanne et al. 2010): Owunwanne, D.; Rustagi, N.; Dada, R.: *Students perceptions of cheating and plagiarism in higher institutions*. In: *Journal of College Teaching & Learning* 7 (2010) 11. pp. 59–68.

(Parhizgar 2012): Parhizgar, S.: *Testing and Technology: Past, Present and Future*. In: *Theory and Practice in Language Studies* 2 (2012) 1. pp. 174–178.

(Pearson Education Ltd 2019): Pearson Education Ltd: *Pearson Test of English Academix: Automated Scoring*, PTE Academic. 2019.

(Peffers et al. 2007): Peffers, K.; Tuunanen, T.; Rothenberger, M. A.; Chatterjee, S.: *A Design Science Research Methodology for Information Systems Research*. In: *Journal of Management Information Systems* 24 (2007) 3. pp. 45–77.

(Pellegrino 2010): Pellegrino, J. W.: *The design of an assessment system for the Race to the Top: A learning sciences perspective on issues of growth and measurement*, Princeton: Educational Testing Service. 2010.

(Piaw 2012): Piaw, C. Y.: *Replacing paper-based testing with computer-based testing in assessment: Are we doing wrong?* In: *Procedia - Social and Behavioral Sciences* 64 (2012). pp. 655–664.

(Piech / Gregg 2018): Piech, C.; Gregg, C.: *BlueBook: A computerized replacement for paper tests in computer science*. In: *Proceedings of the 49th ACM Technical Symposium on Computer Science Education. Baltimore, US.* (2018). pp. 562–567.

(Prados et al. 2011): Prados, F.; Soler, J.; Boada, I.; Poch, J.: *An automatic correction tool that can learn*. In: *Frontiers in Education Conference. Rapid City, US.* (2011). F1D-1-F1D-5.

(Prisacari / Danielson 2017): Prisacari, A. A.; Danielson, J.: *Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use*. In: *Computers in Human Behavior* 77 (2017). pp. 1–10.

(Quinlan et al. 2009): Quinlan, T.; Higgins, D.; Wolff, S.: *Evaluating the Construct-Coverage of the e-rater® Scoring Engine*, ETS Research Report Series - ETS RR-09-01. 2009.

(Ramesh / Sanampudi 2021): Ramesh, D.; Sanampudi, S. K.: *An automated essay scoring systems: a systematic literature review*. In: *Artificial Intelligence Review* 55 (2021). pp. 2495–2527.

(Rausch et al. 2019): Rausch, A.; Kögler, K.; Seifried, J.: *Validation of Embedded Experience Sampling (EES) for Measuring Non-cognitive Facets of Problem-Solving Competence in Scenario-Based Assessments*. In: *Frontiers in Psychology* 10 (2019). pp. 1–16.

(Reddy et al. 2018): Reddy, K. J.; Menon, K. R.; Thattil, A.: *Academic Stress and its Sources among University Students*. In: *Biomedical and Pharmacology Journal* 11 (2018) 1. pp. 531–537.

(Ribera / Lapedriza 2019): Ribera, M.; Lapedriza, A.: *Can we do better explanations? A proposal of User-Centered Explainable AI*. In: *Joint Proceedings of the ACM IUI 2019 Workshops. Los Angeles, US.* (2019). pp. 1–7.

(Richardson / Clesham 2021): Richardson, M.; Clesham, R.: *Rise of the machines? The evolving role of AI technologies in high-stakes assessment*. In: *London Review of Education* 19 (2021) 1. pp. 1–13.

(Rudolph et al. 2023): Rudolph, J.; Tan, S.; Tan, S.: *ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?* In: *Journal of Applied Learning & Teaching* 6 (2023) 1. pp. 342–362.

(Rytkönen / Myyry 2014): Rytkönen, A.; Myyry, L.: *Student Experiences on Taking Electronic Exams at the University of Helsinki*. In: *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2014. Tampere, Finland.* (2014). pp. 2114–2121.

(Sartor 2020): Sartor, G.: *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*. Available online: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf (accessed on 2022-07-10).

(Schlippe et al. 2023): Schlippe, T.; Stierstorfer, Q.; Koppel, M. ten; Libbrecht, P.: *Explainability in Automatic Short Answer Grading*. In: *Proceedings of 2022 3rd International Conference on Artificial Intelligence in Education Technology. Birmingham, UK.* (2023). pp. 69–87.

(Schmidgall / Powers 2017): Schmidgall, J. E.; Powers, D. E.: *Technology and High-stakes Language Testing*. In: *Schmidgal, J. E. & Powers, D. E. (eds.): The handbook of technology and second language teaching and learning* (2017). pp. 317–331. Hoboken: Wiley-Blackwell.

(Schrepp et al. 2014): Schrepp, M.; Hinderks, A.; Thomaschewski, J.: *Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios*. In: *Proceedings of the Third International Conference of Design, User Experience, and Usability. Heraklion, Greece.* (2014). pp. 383–392.

(Schrepp 2021): Schrepp, M.: *Data Analysis Tools - Two Excel-Sheets that make the analysis of your results easy*. Available online: https://www.ueq-online.org/Material/Data_Analysis_Tools.zip (accessed on 2023-07-10).

(Sharma / Jayagopi 2018): Sharma, A.; Jayagopi, D. B.: *Automated Grading of Handwritten Essays*. In: *Proceedings of the 2018 16th International Conference on Frontiers in Handwritten Recognition. Niagara Falls, US.* (2018). pp. 279–284.

(Shermis 2014): Shermis, M. D.: *State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration*. In: *Assessing Writing* 20 (2014). pp. 53–76.

(Shermis 2015): Shermis, M. D.: *Contrasting state-of-the-art in the machine scoring of short-form constructed responses*. In: *Educational Assessment* 20 (2015) 1. pp. 46–65.

(Shermis / Lombard 1998): Shermis, M. D.; Lombard, D.: *Effects of computer-based test administrations on test anxiety and performance*. In: *Computers in Human Behavior* 14 (1998) 1. pp. 111–123.

(Siau / Wang 2018): Siau, K.; Wang, W.: *Building Trust in Artificial Intelligence, Machine Learning, and Robotics*. In: *Cutter Business Technology Journal* 31 (2018) 2. pp. 47–53.

(Simmons / Holliday 2019): Simmons, C.; Holliday, M. A.: *A comparison of two popular machine learning frameworks*. In: *Journal of Computing Sciences in Colleges* 35 (2019) 4. pp. 20–25.

(Sindre / Vegendla 2015): Sindre, G.; Vegendla, A.: *E-exams versus paper exams: A comparative analysis of cheating-related security threats and countermeasures*. In: *Proceedings of the Norwegian Information Security Conference. Trondheim, Norway.* (2015). pp. 1–12.

(Ständige Wissenschaftliche Kommission 2022): Ständige Wissenschaftliche Kommission: *Digitalisierung im Bildungssystem: Handlungsempfehlungen von der Kita bis zur Hochschule*, Gutachten der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz (SWK) 2022.

(Standing Conference of the Ministers of Education and Cultural Affairs 2022): Standing Conference of the Ministers of Education and Cultural Affairs: *Jahresbericht der Kultusministerkonferenz zur Bildung in der digitalen Welt* 2022.

(Stephenson 2018): Stephenson, B.: *An Experience Using On-Computer Programming Questions During Exams*. In: *Proceedings of the 23rd Western Canadian Conference on Computing Education. Victoria, Canada.* (2018). pp. 1–6.

(Stowell / Bennett 2010): Stowell, J. R.; Bennett, D.: *Effects of Online Testing on Student Exam Performance and Test Anxiety*. In: *Journal of Educational Computing Research* 42 (2010) 2. pp. 161–171.

(Summons 1997): Summons, P.: *Automated assessment and marking of spreadsheet concepts*. In: *Proceedings of the 2nd Australasian Conference on Computer Science Education. Melbourne, Australia.* (1997). pp. 178–184.

(Susnjak 2022): Susnjak, T.: *ChatGPT: The End of Online Exam Integrity?* Available online: https://arxiv.org/abs/2212.09292 (accessed on 2023-07-10).

(Sweller 2003): Sweller, J.: *Evolution of human cognitive architecture*. In: *The Psychology of Learning and Motivation* 43 (2003). pp. 215–266.

(Taghipour / Ng 2016): Taghipour, K.; Ng, H. T.: *A Neural Approach to Automated Essay Scoring*. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, US.* (2016). pp. 1882–1891.

(Taras 2005): Taras, M.: *Assessment - summative and formative - some theoretical reflections*. In: *British Journal of Educational Studies* 53 (2005) 4. pp. 466–478.

(Terzis et al. 2012): Terzis, V.; Moridis, C. N.; Economides, A. A.: *How student's personality traits affect Computer Based Assessment Acceptance: Integrating BFI with CBAAM*. In: *Computers in Human Behavior* 28 (2012) 5. pp. 1985–1996.

(Terzis et al. 2013): Terzis, V.; Moridis, C. N.; Economides, A. A.: *Continuance acceptance of computer based assessment through the integration of user's expectations and perceptions*. In: *Computers & Education* 62 (2013). pp. 50–61.

(Terzis / Economides 2011): Terzis, V.; Economides, A. A.: *The acceptance and use of computer based assessment*. In: *Computers & Education* 56 (2011) 4. pp. 1032–1044.

(Thomas et al. 2002): Thomas, P.; Price, B.; Paine, C.; Richards, M.: *Remote electronic examinations: student experiences*. In: *British Journal of Educational Technology* 33 (2002) 5. pp. 537–549.

(Tierney et al. 2011): Tierney, R. D.; Simon, M.; Charland Julie: *Being Fair: Teachers' Interpretations of Principles for Standards-Based Grading*. In: *The Educational Forum* 75 (2011) 3. pp. 210–227.

(Timmis et al. 2016): Timmis, S.; Broadfoot, P.; Sutherland, R.; Oldfield, A.: *Rethinking assessment in a digital age: opportunities, challenges and risks*. In: *British Educational Research Journal* 42 (2016) 3. pp. 454–476.

(Tinoca 2012): Tinoca, L.: *Promoting e-assessment quality in higher education: a case study in online professional development*. In: *Proceedings of the 2012 International Conference on Information Communication Technologies in Education. Rhodes, Greece.* (2012). pp. 213–223.

(Tuah / Naing 2021): Tuah, N. A. A.; Naing, L.: *Is Online Assessment in Higher Education Institutions during COVID-19 Pandemic Reliable?* In: *Siriraj Medical Journal* 73 (2021) 1. pp. 61–68.

(UNESCO 2020): UNESCO: *Global monitoring of school closures caused by COVID-19*. Available online: https://en.unesco.org/covid19/educationresponse (accessed on 2023-07-10).

(vom Brocke et al. 2009): vom Brocke, J.; Simons, A.; Niehaves, B.; Niehaves, B.; Reimer, K.; Plattfaut, R.; Cleven, A.: *Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process*. In: *Proceedings of the 17th European Conference on Information Systems. Verona, Italy.* (2009). pp. 1–14.

(vom Brocke et al. 2015): vom Brocke, J.; Simons, A.; Riemer, K.; Niehaves, B.; Plattfaut, R.; Cleven, A.: *Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research*. In: *Communications of the Association for Information Systems* 37 (2015). pp. 205–224.

(Warnecke / Burchard 2020): Warnecke, T.; Burchard, A.: *Klausur mit Überwachungskamera im WG-Zimmer - Debatte um Uniprüfungen im Digitalsemester*. Available online: https://www.tagesspiegel.de/ wissen/klausur-mit-uberwachungskamera-im-wg-zimmer-7606918.html (accessed on 2023-07-10).

(Webster / Watson 2002): Webster, J.; Watson, R. T.: *Analyzing the Past to Prepare for the Future: Writing a Literature Review*. In: *MIS Quarterly* 26 (2002) 2. pp. xiii–xxiii.

(Weller 2004): Weller, M.: *Learning objects and the e-learning cost dilemma*. In: *Open Learning: The Journal of Open, Distance and e-Learning* 19 (2004) 3. pp. 293–302.

(Wessel et al. 2019): Wessel, D.; Attig, C.; Franke, T.: *ATI-S - An Ultra-Short Scale for Assessing Affinity for Technology Interaction in User Studies*. In: *Proceedings of Mensch und Computer 2019. Hamburg, Germany.* (2019). pp. 147–154.

(Wiannastiti et al. 2018): Wiannastiti, M.; Sujarwo, S.; Matthew, R. T.; Oktriono, K.: *Students' writing test using an integrated-multimedia website: a case study of english professional class*. In: *Proceedings of the 6th International Conference on Information and Education Technology. Osaka, Japan.* (2018). pp. 135–138.

(Wise 2019): Wise, S. L.: *Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating*. In: *Education Inquiry* 10 (2019) 1. pp. 21–33.

(Wu et al. 2011): Wu, K.; Zhao, Y.; Zhu, Q.; Tan, X.; Zheng, H.: *A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type*. In: *International Journal of Information Management* 31 (2011) 6. pp. 572–581.

(Yusuf / Al-Banawi 2013): Yusuf, N.; Al-Banawi, N.: *The Impact Of Changing Technology: The Case Of E-Learning*. In: *Contemporary Issues in Education Research* 6 (2013) 2. pp. 173–180.

(Zhang et al. 2021): Zhang, L. Y.; Petersen, A. K.; Liut, M.; Simion, B.; Alaca, F.: *A Multi-Course Report on the Experience of Unplanned Online Exams*. In: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education. Virtual Event, US.* (2021). pp. 17–23.

(Zhao et al. 2017): Zhao, S.; Zhang, Y.; Xiong, X.; Botelho, A.; Heffernan, N.: *A Memory-Augmented Neural Model for Automated Grading*. In: *Proceedings of the 2017 4th ACM Conference on Learning @ Scale. Cambridge, US.* (2017). pp. 189–192.

Assurance Upon Admission of the Doctoral Examination

Doctoral Programm *Wirtschaftswissenschaften*

I confirm,

1. that the dissertation "Digitization of High-Stakes Exams - Empirical Insights and Design Recommendations for the Digital Execution and Scoring of Exams" that I submitted was produced independently without assistance from external parties, and not contrary to high scientific standards and integrity,
2. that I have adhered to the examination regulations, including upholding a high degree of scientific integrity, which includes the strict and proper use of citations so that the inclusion of other ideas in the dissertation are clearly distinguished,
3. that in the process of completing this doctoral thesis, no intermediaries were compensated to assist me neither with the admissions or preparation processes, and in this process,
    - no remuneration or equivalent compensation were provided
    - no services were engaged that may contradict the purpose of producing a dissertation
4. that I have not submitted this dissertation or parts of this dissertation elsewhere.

I am aware that false claims (and the discovery of those false claims now, and in the future) with regards to the declaration for admission to the doctoral examination can lead to the invalidation or revoking of the doctoral degree.

Göttingen, 10.07.2023 Philipp Hartmann

Overview of Author Contribution on the Conducted Research Studies

Author	Contribution *(marked Author)*
Study I: "The Intention to Participate in Online Exams - The Student Perspective" *(Hartmann et al. 2021) - AMCIS 2021 - published*	
Hartmann, P. *Hobert, S.* *Schumann, M.*	Conzeptualization, methodology, formal analysis, investigation, data curation, writing, visualization
Study II: "From Emergency Remote Assessment to a New Status Quo? – Lessons Learned From Online Assessments During the COVID-19 Pandemic" *(Hartmann/Hobert 2023 b) - ECIS 2023 - published*	
Hartmann, P. *Hobert, S.*	Conzeptualization, methodology, formal analysis, investigation, data curation, writing, visualization
Study III: "Trust, but Verify! - An Empirical Investigation of Students' Initial Trust in AI-Based Essay Scoring" *(Hartmann et al. 2022 b) - AMCIS 2022 - published*	
Hartmann, P. *Hobert, S.* *Schumann, M.*	Conzeptualization, methodology, formal analysis, investigation, data curation, writing, visualization
Study IV: "(AI)n't Nobody Helping Me? - Design And Evaluation of a Machine-Learning-Based Semi-Automatic Essay Scoring System" *(Hartmann et al. 2022 a) – ECIS 2022 - published*	
Hartmann, P. *Holthoff, N.* *Hobert, S.* *Schumann, M.*	Conzeptualization, writing, visualization, supervision
Study V: "Explain AI-Based Essay Scorings without XAI – Empirical Investigation of an User-Centered UI Design for AI-Based AES Systems" *(Hartmann/Hobert 2023 a) - AMCIS 2023 - published*	
Hartmann, P. *Hobert, S.*	Conzeptualization, methodology, formal analysis, investigation, data curation, writing, visualization

Göttingen, 10.07.2023 Philipp Hartmann