TEST TO TOOL

CREATING VALUE FROM EDUCATIONAL ASSESSMENT DATA FOR LEARNERS AND EDUCATORS

EVA DE SCHIPPER

TEST TO TOOL

Creating Value from Educational Assessment Data
for Learners and Educators

Eva de Schipper

TEST TO TOOL

Creating Value from Educational Assessment Data
for Learners and Educators

DOCTORAL THESIS

to obtain
the degree of doctor at the University of Twente,
on the authority of the Rector Magnificus,
prof. dr. ir. A. Veldkamp,
on account of the decision of the Doctorate Board
to be publicly defended
on Thursday the 28th of May 2026 at 14:30 hours

by

Eva de Schipper
born on the 23rd of March 1994 in Kupang, Indonesia

This dissertation has been approved by:
Promotor:
Prof. dr. ir. B.P. Veldkamp
Co-promotor:
Prof. dr. R.C.W. Feskens

Cover design: Eva de Schipper
Cover background illustration: Rocketpixel on Magnific
Printed by: Gildeprint – www.gildeprint.nl
ISBN (print): 978-90-365-7222-4
ISBN (digital): 978-90-365-7223-1
DOI: https://doi.org/10.3990/1.9789036572231

MANUSCRIPT COMMITTEE:

Chair / secretary	Prof. dr. T. Bondarouk	University of Twente
Promotor	Prof. dr. ir. B.P. Veldkamp	University of Twente
Co-promotor	Prof. dr. R.C.W. Feskens	Cito & University of Twente
Members	Prof. dr. J. Braeken	University of Oslo
	Prof. dr. T.H.S. Eysink	University of Twente
	Dr. Qiwei He	Georgetown University
	Prof. dr. I. Molenaar	Radboud University
	Dr. E.C. Roelofs	University of Twente

Contents

CHAPTER 1

Introduction

Data are everywhere. Their volume increases rapidly in many fields, including educational assessment, which relies heavily on the use of data. When a learner interacts with a test question (*item*), data are created from which information about the learner and the item can be derived. In psychometrics (the science, theory, and technology of measurement; e.g., Brennan, 2023), such data are used to estimate the ability of the learner as well as characteristics of the item such as its difficulty. Item characteristics are used to assemble tests, and when combined with information about the learner, this data can also be used for the personalization of tests. Traditional examples include computerized adaptive testing and multistage testing.

The ongoing transition from paper-based to digital testing has resulted in larger amounts and different types of assessment data being available. Testing digitally not only allows capturing student answers (*responses*), but also indicators of the process that a student went through while answering (*process data*), such as the time the student spent answering, how often the student rewrote a response before submitting, where on the screen the student clicked, or even which part of the screen the student looked at at a specific point in time. Data resulting from the digitization of educational processes can be used to provide additional support and insights to learners and teachers, and thus data-informed teaching is gaining traction in recent years (e.g., Schenke & Meijer, 2018; Schildkamp & Poortman, 2022).

The worldwide increase in data availability has gone hand in hand with staggering improvements in the computational processing power of available hardware, such as the introduction of GPUs (graphics processing units) for their parallel computing abilities, as well as scientific breakthroughs in the modeling of data, allowing for the possibility of increasingly complex and often real-time analyses (Wetenschappelijke Raad voor het Regeringsbeleid [WRR], 2021). An example of this phenomenon is the recent widespread adoption of generative AI.

This combination of increased availability of data, improved computational processing power, and scientific modeling breakthroughs offers opportunities to use data from educa-

tional assessment in new ways in order to gain additional insights into learning processes as well as to offer support to various educational stakeholders (Saab, 2025). At the same time, new challenges and dangers arise, such as concerns around pedagogical autonomy (Saab, 2025), model biases, the use of learners' data, and learners' rights to privacy (e.g., Vervaart, 2025). Societal debate has also occasionally centered around such challenges in recent years, discussing, for example, which data should be saved, and who should (and should not) have access to them (e.g., Hagen, 2025).

This dissertation aims to make use of the opportunities that educational assessment data offer in a responsible and practice-oriented manner, and to add value to educational practice for students and teachers by doing so. The main research question is therefore:

"How can we responsibly create additional value from educational assessment data?"

Within the studies of this dissertation, we make use of different types of educational assessment data and a wide array of methods in order to explore possible answers to this question. To ensure a strong connection to educational practice and increase its relevance and potential for real-world application, we involve teachers and students in the research where possible.

1.1 AI in education

While I wrote this dissertation, AI made its way from the domain of science into the lives of anyone with access to the internet and a device. This is so much the case that the Netherlands Scientific Council for Government Policy views AI as a new system technology (Wetenschappelijke Raad voor het Regeringsbeleid [WRR], 2021), on par with electricity or the internal combustion engine. When their report came out in 2021, no labs existed in the Netherlands around the use of AI in education. Since then, NOLAI (The National Education Lab AI) has been founded with large-scale funding from the Dutch government, and many other smaller AI hubs and communities have also come into existence.

The impact of AI on education cannot be overstated. More and more AI tools that use and produce data (or models that were trained on data) are being developed, implemented, and used. For example, the OECD recently reported that about one-third of teachers have used AI in their work, with strong variation between countries (OECD, 2025). At the same time, the proportion of AI tools in education that is based on existing knowledge about learning or teaching remains underwhelming (Molenaar, 2024).

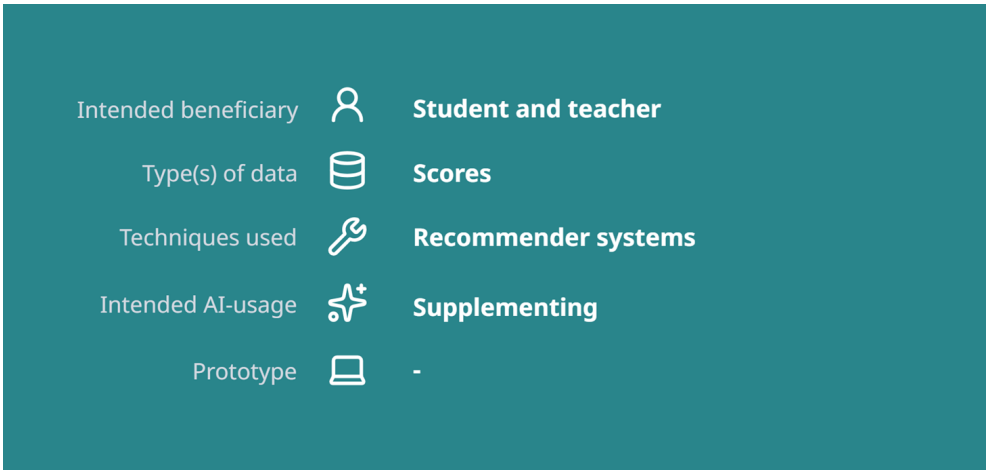
Complicating factors for the integration of AI tools into education are laws and policies around AI. Researchers, developers, and entrepreneurs in the European Union now also must take the AI Act into account, in addition to the already implemented GDPR. For example, under GDPR, data may only be used for the goal for which they were collected, and for children under the age of 16, parental consent is required for data collection. The

AI Act classifies the use of AI in education as high-risk. This classification means that no biometric data may be used, a register must be kept documenting AI use, and proper oversight is required. Upcoming changes in the AI Act and other relevant legislation in upcoming years bring uncertainty to ed-tech companies and stakeholders in education alike. Additionally, legislation varies internationally, complicating the use of AI models and tools developed in different parts of the world, such as by American "Big Tech" companies.

A current societal concern regarding AI is that increasing automation will lead to job loss or change the distribution of available jobs. This is a valid concern, especially considering that this is already becoming a reality in some places and areas of work, such as the American tech industry (Burleigh, 2026). It is still unclear how exactly the introduction of AI will impact jobs in the educational domain. However, the field of education has a large number of tasks that are difficult to automate (Molenaar, 2024). This is why Molenaar argues that AI in education should augment human intelligence rather than replace it (Cukurova et al., 2019). In this perspective, humans and AI can collaborate to achieve more than either could alone. Human work can be partially replaced by, supplemented with, and enriched through the use of AI. In the studies within this dissertation, we make use of several techniques that are viewed as AI, such as recommender systems and natural language processing techniques (Wetenschappelijke Raad voor het Regeringsbeleid [WRR], 2021). When describing the individual studies of this dissertation in the following section, I indicate whether the techniques are used to partially replace, supplement, or enrich human work.

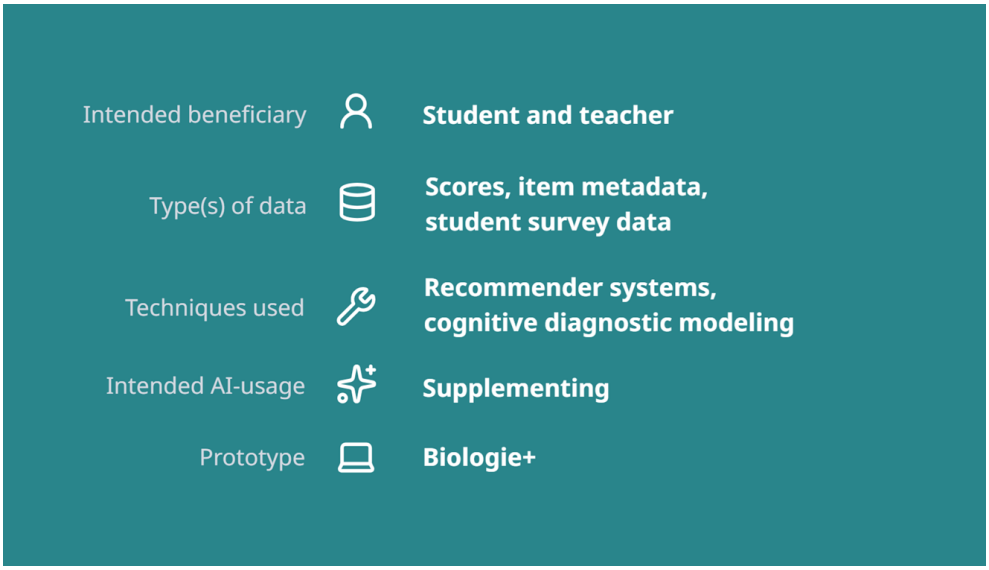
1.2 Included studies

Chapter 2



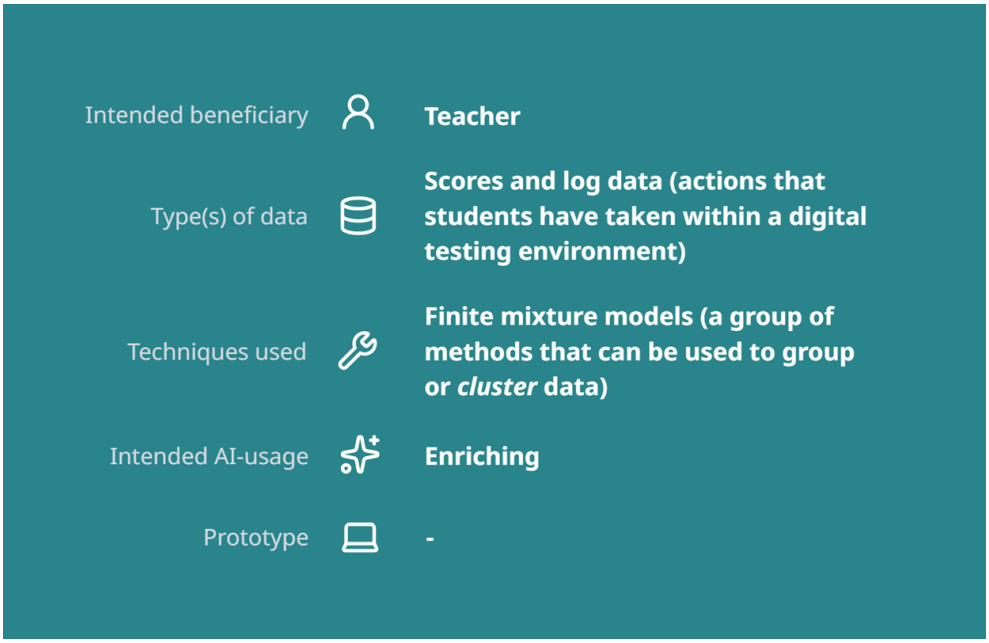
Recommender systems are the algorithms used by companies such as Spotify, Amazon, and Netflix to recommend products (e.g., songs and series) to their users. **Chapter 2** investigates whether these algorithms can be used to recommend personalized practice tests to students based on scores they achieved on an earlier test. We apply five different recommender systems to existing data from central examinations at the end of Dutch secondary education to predict scores for student responses from earlier years, and then compare these scores to those that the students actually received. In this way, we assess the theoretical feasibility of using recommending systems for providing practice tests to students.

Chapter 3



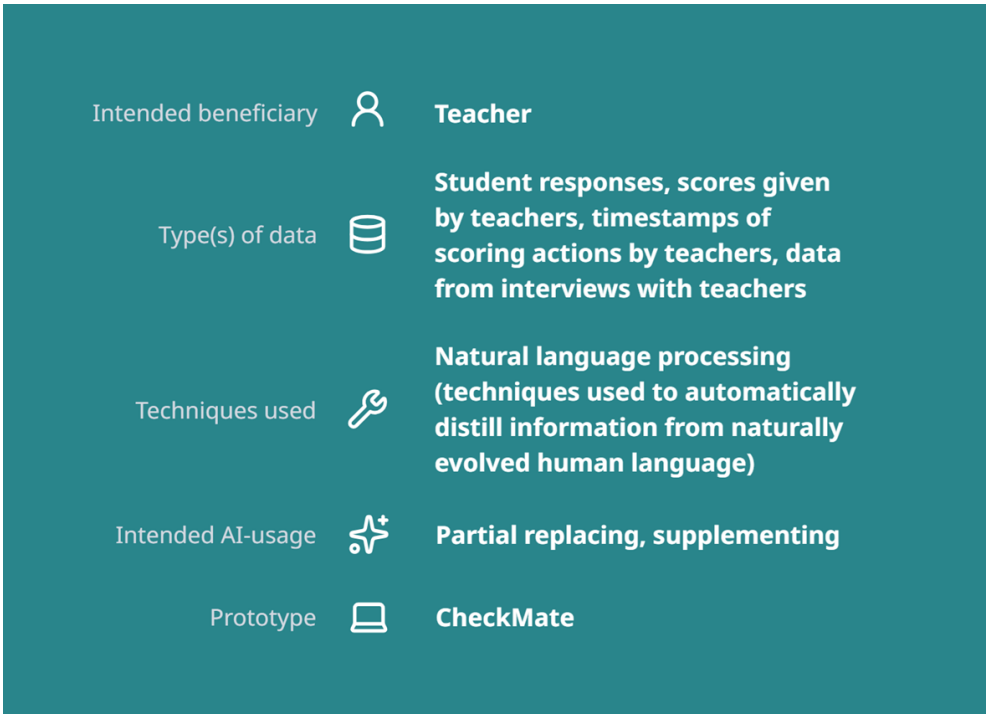
The next chapter picks up where the previous left off. After ascertaining the theoretical feasibility of using recommender systems for providing personalized practice tests in Chapter 2, in **Chapter 3** we develop a new type of recommender system using *cognitive diagnostic modeling* (CDM, a category of methods within psychometrics that assesses whether a student has achieved mastery on specific topics), and experiment with the application of it within the classroom using a digital prototype called *Biologie+* (see Section 1.3). In the experiment, some students receive a personalized practice test and others a fixed practice test. We investigate whether students with a personalized practice test achieve larger learning gains, and how the students in each experimental condition experience the practice session.

Chapter 4



In **Chapter 4**, we not only make use of scores that students received for their responses on items, but also of data derived from recording their actions within a digital assessment environment. For one mathematics item that was administered in a large-scale survey in France, we try to uncover the solution strategy that students used to solve the item using the log data resulting from their actions. This type of information could be used by teachers to determine pedagogical actions towards their students, and potentially evaluate their own teaching leading up to the assessment.

Chapter 5



Where Chapter 2 and 3 focus on aiding students in their learning processes, and Chapter 4 focuses on gaining insights into student processes to aid teachers, **Chapter 5** aims to support teachers in their duty of grading tests. Teachers worldwide experience high working pressure and increased administrative burden, which contributes to increased staff turnover and teacher shortages (Dupriez et al., 2016; Spruyt et al., 2023). Scoring open-ended questions is a difficult and time consuming task, which can also lead to the choice around which types of questions to include in a test being a pragmatic one as opposed to a choice based on the most valid way to measure student abilities. Furthermore, random error introduced by inconsistent scoring can impact the reliability of construct measurements. This chapter contains practice-oriented qualitative and quantitative research into how scoring student answers to open-ended questions can be made more efficient and consistent, hopefully eventually leading to 1) lighter administrative burdens for teachers, 2) opportunities for more valid construct measurements, and 3) more reliable construct measurements. This study is executed using the digital prototype CheckMate (see Section 1.3)

1.3 Prototypes: CheckMate & Biologie+

For two of the studies in this dissertation, web-based prototypes were developed and used to collect data. The prototypes were developed by the prototyping team of CitoLab, the research and innovation department of the Cito Foundation. Team Prototypes is an interdisciplinary team, combining expertise in various disciplines such as educational measurement, user experience, software, and graphic design in order to develop innovative new measurement instruments and improve on existing ones in an iterative and user-focused process. The resulting prototypes are used as research instruments, as is exemplified in Chapter 3 and Chapter 5 of this dissertation.

CheckMate is a prototype developed to help teachers score open-ended questions efficiently and consistently. The development process and functionalities of this prototype are described in detail in Chapter 5. However, it is useful to mention that outside the scope of this study, research and development on CheckMate has continued in the form of a co-creation project funded by NOLAI in which CitoLab collaborates with two Dutch secondary education schools (Parmant Aloysius and Roncalli Bergen op Zoom). The project has a duration of three years (2024 - 2027) and aims to contribute to knowledge on all factors surrounding scoring assistance, and to encourage the implementation of such technology in schools in the Netherlands.

The prototype Biologie+ is an exam preparation tool that teachers can use with their students while working towards their central Biology examinations. In the student interface of the web application, students are administered several tests: two diagnostic tests and two practice tests. After each of the diagnostic tests, they are offered a report in which they can view their performance on the different content domains of Biology. In the teacher interface of the application, teachers can view their students' progress in individual student reports and class reports.

In order to save teachers time while participating in the study, all tests are automatically scored using techniques developed while building the prototype CheckMate (further described in Section 1.3 and Chapter 5). However, students can alter their own scores if they feel that the automated score is incorrect. It is in the interest of the students to score their own tests fairly, because there are no consequences attached to their performance and realistic scores lead to a more informative diagnostic report. The scores were nevertheless checked by the researchers and when necessary by content experts at Cito Foundation before being used as data.

An important requirement throughout the development of Biologie+ was to ensure that teachers and students benefited directly from participating in the study, and were not burdened unnecessarily. Therefore, providing services to the participants, such as automated scoring, a user-friendly interface, and instant reporting was a primary focal point. These services were not necessary to execute the experimental study in Chapter 3,

but they were highly appreciated by all participating teachers – all of them indicating that they would gladly use the application again in a following year, if offered. Biologie+ serves as an example of how research can benefit participants directly if properly integrated into their daily practice.

1.4 Final note

The studies in this dissertation contribute to the available scientific knowledge on how we can use data from educational assessment to assist stakeholders in educational practice. However, results such as these are only valuable once they are thoroughly tested by these stakeholders and implemented in a user-friendly and ethical manner (Saab, 2025). At the same time, successful implementation of data-based tools cannot be achieved without solid scientific research on the impact of technology and educational interventions (Saab, 2025). My hope is that this dissertation contributes to building bridges between scientific research, educational technology development, and educational practice.

CHAPTER 2

Personalized and automated feedback in summative assessment using recommender systems

This chapter has been published as:

De Schipper E, Feskens R. & Keuning J. (2021) Personalized and Automated Feedback in Summative Assessment Using Recommender Systems. *Frontiers in Education, 6.* https://doi.org/10.3389/feduc.2021.652070

Abstract

In this study we explore the use of recommender systems as a means of providing automated and personalized feedback to students following summative assessment. The intended feedback is a personalized set of test questions (*items*) for each student that they could benefit from practicing with. Recommended items can be beneficial for students as they can support their learning process by targeting specific gaps in their knowledge, especially when there is little time to get feedback from instructors. The items are recommended using several commonly used recommender system algorithms, and are based on the students' scores in a summative assessment. The results show that in the context of the Dutch secondary education final examinations, item recommendations can be made to students with an acceptable level of model performance. Furthermore, it does not take a computationally complex model to do so: a simple baseline model which takes into account global, student-specific and item-specific averages obtained similar performance to more complex models. Overall, we conclude that recommender systems are a promising tool for helping students in their learning process by combining multiple data sources and new methodologies, without putting additional strain on their instructors.

2.1 Introduction

Feedback is an important and widely researched factor in improving educational outcomes of students due to its potential to indicate what is needed to bridge the gap between where students are and where they are going (e.g., Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008). It can be defined as "... information provided by an agent regarding aspects of one's performance or understanding" (see Hattie & Timperley, 2007). The concept of feedback is central to the field of formative educational assessment. Formative assessment aims to improve student attainment of the learning material through participation in the assessment (e.g., Black & Wiliam, 2009; Heritage, 2007; Shute, 2008). Formative assessment takes place during the learning process, unlike its counterpart summative assessment, which usually takes place at the end of a learning process and aims to make a decision or judgment about the student's skill or knowledge level (Dixson & Worrell, 2016). Feedback on the basis of summative assessment is much less common for several reasons, including the aforementioned timing of the assessment at the end of a learning process. Other important reasons include the fast and standardized nature of summative assessment: the assessment usually needs to be followed quickly by a judgment call or decision, and the assessment design is optimized for quantitative decision making (e.g. grading, or a pass/fail judgment), as opposed to for providing a basis for helpful feedback. However, feedback can be useful to students in the context of summative assessment as well, and its data provides a wealth of information about their knowledge or skill level. This paper explores a method of providing personalized and automated feedback to students on the basis of summative assessment data.

Students who are preparing for educational assessment can benefit from practice material that is attuned to their individual needs. This can be considered topic contingent feedback, defined by Shute (2008) as "feedback providing the learner with information relating to the target topic currently being studied". Students tend to study by rereading (Karpicke et al., 2009), and most students are poor at judging their mastery of the material (Dunlosky & Lipko, 2007). Furthermore, an extensive meta-analysis on practice testing by Adesope et al. (2017) found that practice tests are more beneficial for learning than all other comparison conditions (e.g., Callender & McDaniel, 2009; Roediger III & Karpicke, 2006b), especially when the practice and final test formats are identical. This is true regardless of whether the students are supplied with corrective feedback (information on the correctness of their answers) afterwards. Practicing with personalized practice material, such as a set of test questions (*items*), can therefore be a valuable addition to the learning process.

Providing a personalized set of test items can be seen as a recommendation problem. Recommender systems (e.g., Koren et al., 2022) are often used these days by commercial parties such as Netflix, Amazon, and Spotify to recommend items (movies, products,

songs, etc.) to users (customers). There have been many applications of recommender systems in the field of education, particularly within the context of e-learning systems. Manouselis et al. (2011) provide an introduction to recommender systems for Technology Enhanced Learning settings and Rivera et al. (2018) give a more recent overview of applications in education. The most common application is to recommend learning materials and resources, such as books, papers, and courses to students (e.g., Aher & Lobo, 2013; Bobadilla et al., 2009; Bokde et al., 2015; Ghauth & Abdullah, 2010; Khribi et al., 2008; Liang et al., 2006; Luo et al., 2010; O'Mahony & Smyth, 2007; Tang & McCalla, 2005; Vialardi et al., 2009; Zaíane, 2002).

Recommender systems are not yet widely used in educational assessment contexts. One example of an application of recommender systems within the framework of formative assessment is a recent paper by Bulut et al. (2020), who developed an intelligent recommender system (IRS) that can be used to produce individualized test administration schedules for students. It is easy to draw parallels between educational assessment data and the more commercial contexts in which recommender systems are most often applied: users can be likened to students and products or items to test questions. Assessment data can look rather similar to the rating matrices that stem from the more commercial applications, especially when using an incomplete test design (where different students are subjected to different test items). Recommender systems could be used for presenting students with a personalized set of practice items, giving feedback such as "Students like you struggled with these exam questions" or "In the past, you have struggled with exam questions such as these". Essentially, this entails using an algorithm with which items are selected on the criterion that similar students did *not* have affinity with them. Practicing with the recommended questions and focusing on the learning material covered in them could help students study effectively by closing gaps in their knowledge in a targeted way.

Presumably one reason for the lack of applications of recommender systems within educational assessment is that the established field of computerized adaptive testing (CAT) is concerned with the similar task of providing test-takers with subsequent test questions (e.g., Van der Linden, Glas, et al., 2000). One difference between these methods lies in their purpose: in a CAT, the primary goal for item selection is related to test optimization (to accurately estimate the student's ability), whereas the aim of item selection in this paper is related to optimization of the learning process (to maximize the student's grasp of the material). Recommender systems are more suited to this purpose than the dominant modeling paradigm in the field of CAT (*item response theory*, or IRT). IRT assumes that the performance of a student on a test is dependent on the latent ability of the student and characteristics of the test items (e.g., Embretson & Reise, 2013; Hambleton et al., 1991). When using IRT to generate new test items for students in a CAT, students who have the same estimated latent ability will be given the same test item. Different recommender algorithms take different additional information into account (such as the

similarities between the score patterns of different students) and therefore have increased potential for personalization.

This paper explores an application of recommender systems for summative assessment. Specifically, we consider recommender systems as a means of providing automated and personalized feedback to students based on their scores in a summative assessment. The current study differs from the existing literature on recommender systems in education in two ways. First, this study applies recommender systems in a high-stakes summative assessment context. Unlike in a pervasive digital learning environment or in a formative assessment context, the available information per student is limited to a single summative assessment. Second, the information that is used to generate recommendations differs between this study and most related literature, and with it the items that are recommended. The majority of the aforementioned studies use explicit evaluation ratings by students or teachers on the quality or usefulness of the items that are to be recommended. In this study, we use the students' scores on an exam as input for the recommendation algorithms. Thai-Nghe et al. (2010) used similar input (students' scores on their first attempt on a task) in their application of recommender systems but focused on the task of predicting student achievement rather than providing recommendations to students.

The objectives of this study are to determine whether recommender systems can successfully be used to recommend practice questions to students following a high-stakes summative assessment, and if so, which algorithms are most suited to this purpose. To this end, we compare the performance of several types of such recommender algorithms. The remainder of the text is structured as follows: in the methods section, we introduce recommender systems and describe the algorithms with which we will recommend items to students, as well as the data that we use. The results section evaluates the performances of these algorithms. In the conclusion, the performances of the algorithms are compared. We then go on to discuss the practical relevance and potential of the results in the discussion.

2.2 Methods

2.2.1 Data

The data used in the current study come from the final examinations for secondary school students in the Netherlands. These examinations are obligatory and provide the students with access to higher education such as university when completed with a sufficient result. After taking their secondary school exams, students in the Netherlands get the opportunity to take one resit exam. Reasons for a student to take this opportunity are to (a) improve their overall grade (e.g. in order to improve the odds of getting accepted into the higher education of their choice) or (b) obtain a passing grade for a course for which that is a

diploma prerequisite. The resit exam takes place two to four weeks after the grading of the initial exam, during which the student must prepare. Due to these time constraints, it is necessary that any feedback to the student is delivered within this time frame.

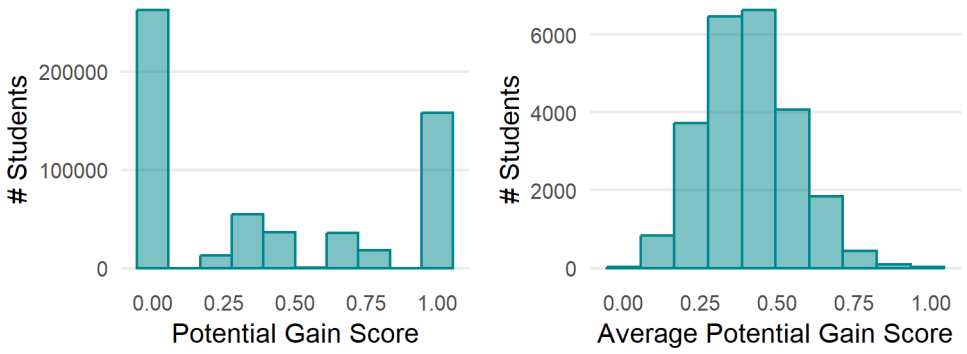
We focus on digital exam data for mathematics in the year 2019 from students in grade 10 of a vocational track in Dutch secondary education. The subject of mathematics was chosen because these exams feature a relatively large amount of items with a maximum score of higher than one. Items such as these, for which partial scores can be obtained, give us more finely grained information about the student's ability. There are nine test versions (*booklets*), each with 22 to 26 test questions (*items*), taken by between 2,453 and 2,945 students (24,167 in total). There are 72 unique items in total, and all booklets have at least a third of their items in common with at least one other booklet, ensuring that all booklets are (indirectly) linked to each other through items that they have in common. All students are subjected to only one of the booklets, which means that they face only a subset of all available items during their exam. This means that, for all students, there is a set of items that they have not yet been faced with due to the nature of the test design. These are items that could be recommended to students to practice with.

The responses that students gave during their first exam have been scored, either automatically (multiple choice questions) or by their teachers. The items have a maximum score of between one and five, and no decimal scores are given. The score of a student on an item is rescaled to a measure we will call *potential gain*. First, the score is divided by the maximum obtainable score on the item. This gives us the proportion of the possible score that is obtained on the item. This proportion is then subtracted from one. This rescaled score can be interpreted as the percentage of the maximum item score that was *not* obtained by the student. We use this as a measure of how much can still be learned by the student concerning the course material that was covered by the item. This is also the measure that we will try to predict using recommender algorithms, and upon which recommendations will be based.

The left side of figure 2.1 displays the distribution of the potential gain scores in the data. It is more common to score either full credit or no credit than it is to score partial credit on the exam items. There is a noticeable lack of scores in the range $0.1 - 0.2$, and very few scores in the ranges $0.4 - 0.5$ and $0.8 - 0.9$. This is due to the possible scores than can be acquired on the items. There is no item in the data that would enable a student to receive 90% of the points, thus the range of $0.1 - 0.2$ to gain is empty. A similar situation is true for the ranges $0.4 - 0.5$ and $0.8 - 0.9$. These can only be attained on an item with a maximum score of 5, and there is only one such item in the data.

As illustrated in the right side of figure 2.1, the average potential gain scores of the students is normally distributed with a mean of 0.41. It is more common for students to have a low average potential gain score, and there are very few students with an average

Figure 2.1: Distribution of Potential Gain Scores in the Data.



potential gain score above 0.8. These observations make sense in the light of the context of final examinations: most students will get a decent average score and pass the exam (and therefore have a low average potential gain score).

2.2.2 Software

All analyses are performed in the programming language Python 3 (Van Rossum & Drake, 2009), using the Surprise library (Hug, 2020). Graphical representations of the data and the results are made using the R programming language (R Core Team, 2025). The scripts (written in the Python 3 language) that detail how the recommendations are produced and evaluated are included as appendices.

2.2.3 Recommender Systems

Recommender systems can roughly be divided into content-based filtering methods, collaborative filtering (*CF*) methods and hybrid approaches which combine both these designs (Melville & Sindhwani, 2017). CF methods use matrix factorization techniques to characterize users and content and make statements about which content and which users are similar (Koren et al., 2009). They can be further subdivided into *neighborhood-based* or *memory-based* and *model-based* approaches. In neighborhood-based techniques, a subset of users or items are chosen based on their similarity to the active user or item, and a weighted combination of their ratings is used to produce predictions for the active user's rating on the active item (e.g., Breese et al., 2013; Melville & Sindhwani, 2017). Model-based techniques provide recommendations by estimating parameters of statistical models for user ratings (e.g., Billsus, Pazzani, et al., 1998; Koren et al., 2009). Content-based filtering methods (e.g., Balabanović & Shoham, 1997; Lang, 1995; Mooney & Roy, 2000)

provide recommendations by comparing representations of content describing an item to representations of content that interests the user (e.g., Melville & Sindhwani, 2017).

In this study, we compare the performance of several types of such recommendation methods. The input for the recommender algorithms is the set of potential gain scores, accompanied by their student identifier and the label of the item. The output will be a set of recommended items for each student. Firstly, two baselines are included with which the performance of more advanced algorithms can be compared. The first baseline is an algorithm that predicts a random score for students on the items by drawing from a normal distribution. The mean and standard deviation for this distribution are estimated using the observations in the training data. The performance achieved by this baseline will be considered the absolute minimum that must be achieved by the other algorithms. The other baseline algorithm predicts scores by taking into account the overall average score, the average score on the item and the average score of the student. We include this baseline algorithm to be able to assess the added benefit of more complicated and more computationally heavy methods. For the score of student u on item i, the baseline estimate is:

$$b_{ui} = \mu + b_u + b_i \tag{2.1}$$

Where μ is the overall average score, and the parameters b_u and b_i denote the deviations from μ of student u and item i respectively. b_u and b_i are estimated by solving the least squares problem detailed by Koren (2010). We alternatingly use the following equations 10 times for all items and all students, using 0 as the starting value for both b_u and b_i.

$$b_i = \frac{\sum_{u:(u,i)\in K} (r_{ui} - \mu - b_u)}{\lambda_i + |\{u|(u,i) \in K\}|} \tag{2.2}$$

$$b_u = \frac{\sum_{i:(u,i)\in K} (r_{ui} - \mu - b_i)}{\lambda_u + |\{i|(u,i) \in K\}|} \tag{2.3}$$

Where r_{ui} is the score of student u on item i, and K is the set of student-item pairs for which the score is known. The regularization parameters λ_i and λ_u are used to avoid overfitting (Koren, 2010) and are set to 10 and 15 respectively.

Next, we include two neighborhood-based CF methods: user-based and item-based. When user-based CF (UBCF) predicts a score for a student on an item, it relies more heavily on information from students with a similar score pattern. Item-based CF (IBCF) is similar, but is based on the similarities between items, not between students. To compute the similarities between students or between items, we use the cosine similarity measure, which is the cosine of the angle between two vectors. Values range between -1 and 1, where -1 is perfectly dissimilar and 1 is perfectly similar.

For students u and u', the cosine similarity is:

$$sim(u, u') = cos(\theta) = \frac{\mathbf{r}_u \cdot \mathbf{r}_{u'}}{||\mathbf{r}_u|| \cdot ||\mathbf{r}_{u'}||} = \sum_i \frac{r_{ui} \cdot r_{u'i}}{\sqrt{\sum_i r_{ui}^2} \cdot \sqrt{\sum_i r_{u'i}^2}} \tag{2.4}$$

The UBCF algorithm will predict a score for a student on an item by taking the average of the scores that the 40 most similar students obtained on that item. The IBCF algorithm predicts the score by taking the average of the scores that the student in question received on the 10 items that are most similar to the item in question.

Neighborhood-based CF methods are simple and therefore transparent, and they are known to perform well (Desrosiers & Karypis, 2011). IBCF is much faster and more scalable than UBCF and can result in similar or better performance (Deshpande & Karypis, 2004; Sarwar et al., 2001). For both these methods, we apply a type of centering in which we reduce all scores by their baseline estimates. In a more commercial context, centering is applied, for instance, to correct for the general tendencies of users to give higher or lower scores. In this context we use centering to control for the students' general ability levels and the items' general difficulty levels. Using UBCF, the rating $\hat{r}$ for student u on item i is estimated as

$$\hat{r}_{ui} = b_{ui} + \frac{\sum\limits_{v \in N_i^k(u)} \mathsf{sim}(u, v) \cdot (r_{vi} - b_{vi})}{\sum\limits_{v \in N_i^k(u)} \mathsf{sim}(u, v)}, \tag{2.5}$$

where b_{ui} is the baseline estimate for student u on item i, and k is the number of similar students (v) taken into account (40). Using IBCF, the rating $\hat{r}$ for student u on item i is estimated as

$$\hat{r}_{ui} = b_{ui} + \frac{\sum\limits_{j \in N_u^k(i)} \mathsf{sim}(i, j) \cdot (r_{uj} - b_{uj})}{\sum\limits_{j \in N_u^k(i)} \mathsf{sim}(i, j)}, \tag{2.6}$$

where k is the number of similar items taken into account (10). For more detail, please refer to Section 2.2 by Koren (2010).

Finally, we include the more complex Singular Value Decomposition (SVD) approximation method. SVD approximation is a model-based algorithm matrix factorization algorithm (Golub & Reinsch, 1971; Koren et al., 2009), built upon much older ideas (Eckart & Young, 1936). The idea behind such models is that attitudes or preferences of a user can be determined by a small number of hidden factors. To this end, it reduces the rating matrix into two smaller matrices through extracting a smaller number of features. SVD approximation is more scalable (as are other model-based approaches) (Cacheda et al., 2011; Sarwar et al., 2002) and therefore may perform better than UBCF, depending on the scale and sparsity of the data.

More details on this application of the SVD algorithm can be found in the Surprise library documentation (Hug, 2020), which builds upon work by Funk (2006), Koren et al. (2009), and Ricci et al. (2011).

2.2.4 Evaluation

For the evaluation of the recommender algorithms, we use k-fold cross validation. This means that the data is randomly split into k parts and in each run $k-1$ parts are used for training and the remaining part is used for testing. After all k runs, each part was used as the test set exactly once. We have chosen $k = 3$. This leaves us with an average number of 7.99 items in the test sets per student.

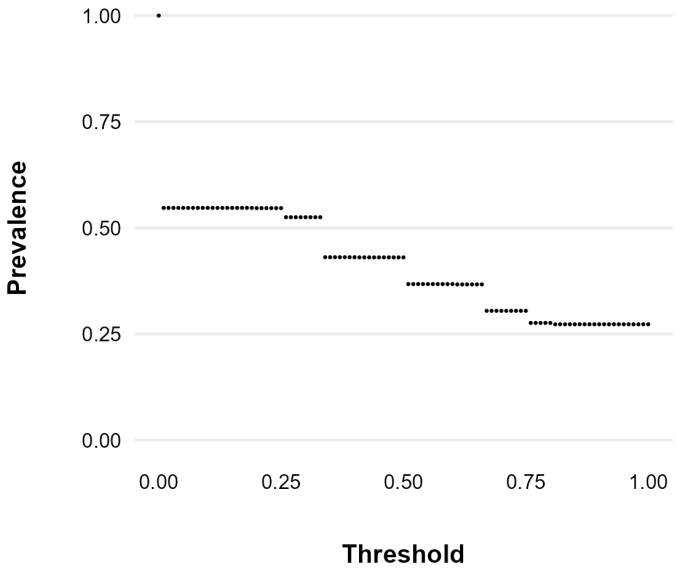
The training data is simply a random subset (66.6% percent) of the complete data. The rest of the data is used as test data. For each student, the information that is available about them in the training data is used for predicting their scores on items in the test set. The predicted score on an item determines whether the item is recommended to the student. It is by comparing observed scores in the test set with the predicted scores on those items that we can evaluate the recommender algorithm (Breese et al., 2013).

When testing the performance of an algorithm on the test set, we use the eight highest predicted potential gain scores for each student. This value is chosen so that we use as many items as we can while still ensuring that most students have enough observations available in the test set. For each user, we determine which observed scores (in the top eight items with the highest predicted scores in the test set) are above a specified threshold. These items are considered relevant to the student, and should be recommended by the algorithm. We also identify which predicted scores are above the threshold. These are items that the algorithm would recommended to the student. We vary the value of the threshold from 0 to 1 in increments of 0.01. A threshold of 0.50, for example, indicates that a *good* item is an item on which a student's predicted score is less than half of the maximum achievable score.

We make use of a confusion matrix that contains four important pieces of information: the number of recommendations that ought to have been recommended (true positives), the number of items that have correctly not been recommended (true negatives), the number of items that have been recommended that ought not to have been (false positives) and the number of items that have not been recommended but ought to have been (false negatives).

Several evaluation metrics can be derived from this confusion matrix. Precision tells us how often items recommended by the classifier are correct recommendations. This is the most important metric for our current context, as the idea is to help a student spend their time wisely and not lead them to spend time on learning materials that they have already mastered. Recall (also known as the true positive rate) informs us on the proportion of useful recommendations that have actually been recommended. High performance in

Figure 2.2: The prevalence, or the proportion of potential gain scores higher than or equal to the different threshold values.



terms of recall would be preferable, because we would be able to offer the students more recommendations, potentially uncovering more of the topics that a student has issues with. We consider recall to be less important than the correctness of the recommended items (precision).

The false positive rate (*FPR*) tells us how many of the items that should not be recommended (due to a low score) are recommended nonetheless. By combining the FPR with the recall (also known as the *true positive rate*, or *TPR*), receiver operating characteristic (ROC) curves can be made, which are used in many fields to assess the performance of a classifier.

The prevalence tells us the proportion of scores in the data above the predefined threshold. It depends on the chosen threshold and influences the number of items that can be recommended to students in practice. Figure 2.2 shows the prevalence belonging to different threshold values between 0 and 1 (in increments of 0.01. For example, 27% of the potential gain scores ($158,270$ out of $579,622$ observations) exceed a threshold of 0.90.

2.3 Results

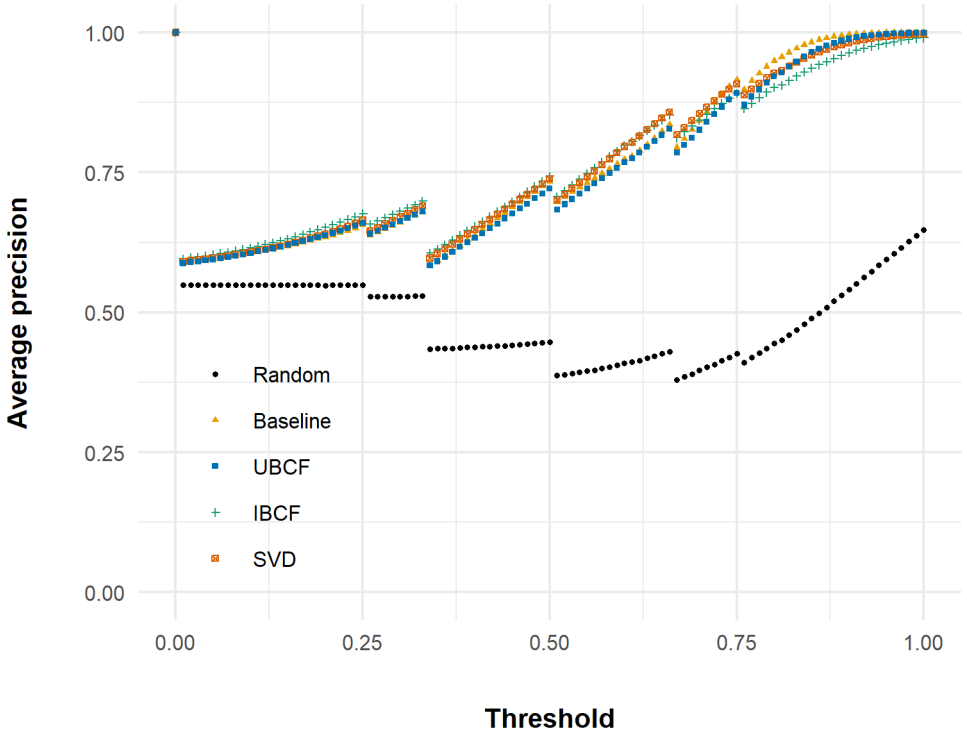
Three-fold cross-validation was applied to verify the comparability of the results for different segments of the data. There was no substantial difference between folds in terms of precision and recall. The largest difference between two folds in terms of precision was 0.0128, which occurred using the UBCF model with a threshold of 0.73. The largest difference in recall between two folds was 0.0153, using the UBCF model with a threshold of 0.68. These differences in precision and recall are negligible.

Figure 2.3 compares the average performance in terms of precision of the five different models, evaluated for threshold values ranging from 0 to 1 in increments of 0.01. Several things can be observed upon inspecting the graphs. Firstly, it should be noted that the performance of the models does not increase continuously with the thresholds, but instead acts in a stepwise fashion. This can be attributed to the stepwise nature of the proportion of potential gain scores equal to or higher than the threshold in the data (the prevalence, see figure 2.2). This stepwise pattern can also be observed in the performance of the models in terms of all other evaluation metrics that are still to come. Next, it is clear that the random model has a lower precision for all threshold values, and that all other models follow a similar trend in precision. The differences between the four other models are rather small, though they increase slightly for higher threshold values. For threshold values lower than 0.62, IBCF performs the best, whereas for threshold values from 0.62 up to and including 0.73, SVD approximation outperforms the other models. For threshold values higher than or equal to 0.72 the baseline model has the highest precision.

Figure 2.4 compares the average recall for the five different models, evaluated for the different threshold values. In general, a higher threshold leads to lower performance in terms of recall for all models. The random model outperforms the other models when the threshold exceeds 0.72. Similarly to their performance in terms of precision, the performance in terms of recall for the four other models is very similar for most threshold values. It differs more for threshold values at both ends of the scale. For threshold values equal to or below 0.39, the baseline model has the best performance in terms of recall. For threshold values between 0.40 and 0.71, UBCF usually outperforms the other models, although the differences are very small.

The ROC curves for the five models are displayed in figure 2.5. ROC curves evaluate the false positive rate and the true positive rate (or recall) for different threshold values. A larger area under the curve (*AUC*) indicates better classification performance, where 1 is the best possible AUC value and 0 the worst. A random model is expected to have an AUC of 0.5, which means that it is not able to distinguish classes from each other. The random model included in this study has an AUC of 0.530. The four other models' AUCs differ only marginally: the baseline model performs best with an AUC of 0.762, after which comes UBCF (AUC $= 0.758$), SVD approximation (AUC $= 0.747$) and finally

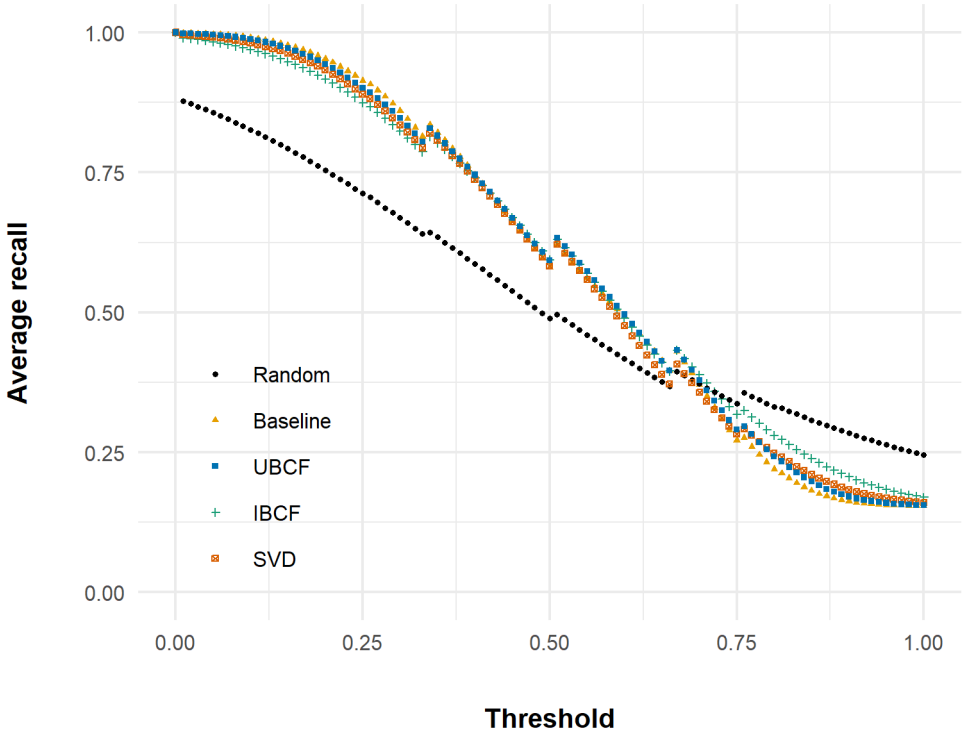
Figure 2.3: Precision per model for different thresholds, averaged over students and cross-validation folds.



IBCF (AUC $= 0.741$).

Although the average number of recommended items is not necessarily used for evaluating the models, it is nevertheless informative for assessing their practical purpose. Therefore, figure 2.6 displays the average number of items recommended to students in the test sets. The maximum possible number of recommended items is 8, because the models were evaluated for the top 8 items with the highest predicted scores in the test set. In figure 2.6, it can be seen that higher threshold values lead to a lower number of recommended items (on average). This is to be expected, because a higher threshold means a lower prevalence, and therefore a smaller number of items that qualify to be recommended to students. The random model recommends more items to students (on average) when the threshold is more than or equal to 0.37. The other four models are fairly similar, although they differ a little bit more on each end of the threshold scale. For low thresholds (less than or equal to 0.36), the baseline model recommends the most items, whereas for high thresholds (higher than or equal to 0.40), the IBCF model recommends the most.

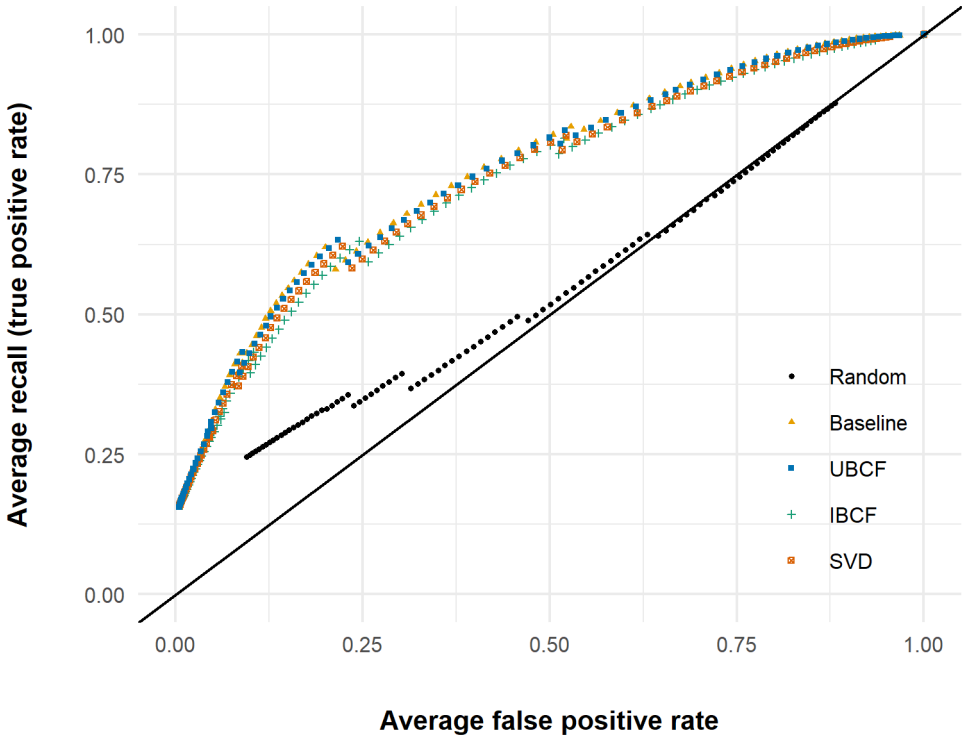
Figure 2.4: Recall per model for different thresholds, averaged over students and cross-validation folds.



2.4 Conclusion

In this study, we recommended sets of items to students using several commonly used recommender system algorithms. Findings indicate that there are no substantial differences between the tested models, except for the random model which was used as a baseline and generally performs a lot worse. A surprising result is that the other baseline model, which predicts scores by taking into account the global, student-specific, and item-specific averages, performs similarly to the more complex models that were included. This suggests that a) the relationships between items and those between users do not add valuable information towards predicting scores that is not already captured in aforementioned baseline and b) that the prediction is not substantially improved by using matrix factorization techniques. Given similar performance, a simpler model is preferable for reasons of transparency (it is easier to explain to the target audience) and computational cost. The benefit of a lower computational cost is especially pronounced when comparing the baseline model to user-based collaborative filtering: the computational complexity of

Figure 2.5: False positive rate versus recall (ROC curve) per model for different thresholds, averaged over students and cross-validation folds.

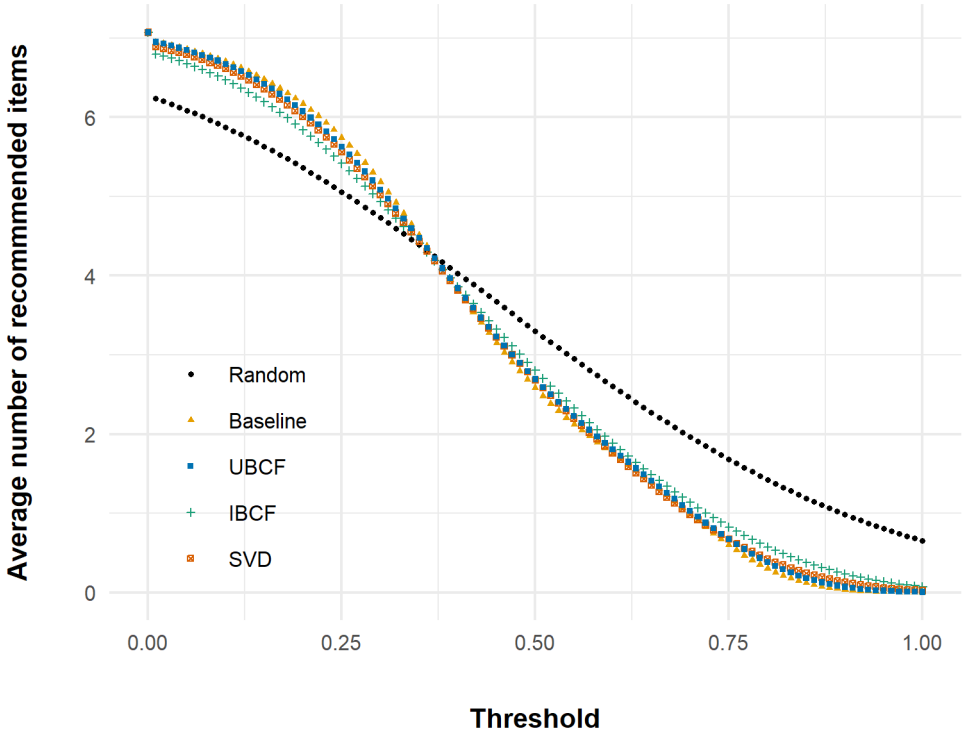


the user-user similarity matrix increases near quadratically with the number of users.

A second conclusion that can be drawn is that a trade-off has to be made between performance in terms of precision and performance in terms of recall, when choosing a threshold for what constitutes a good potential gain score. In this particular context precision is of higher importance and thus an acceptable level of precision will be considered as a primary prerequisite condition. As the precision increased along with an increase in threshold, we would for this reason not prefer the lower thresholds that were included in the study. Besides the fact that lower thresholds result in lower precision, choosing a low threshold also means that a student may be recommended more items that are already well within their capabilities. This renders the recommendations of less practical use to the students.

Choosing a high threshold leads to a lower model performance in terms of recall. When the threshold exceeds 0.72, the four more informed models even perform worse in terms of recall than the random model. This is undesirable. Using a high threshold also introduces a more practical problem: the prevalence of items above the threshold

Figure 2.6: Number of recommended items per model and threshold, averaged over students and cross-validation folds



decreases substantially. A reduced performance in terms of recall adds to this issue. For example: the prevalence of observations above a threshold of 0.90 is 0.27, and the recall for the baseline model is 0.16. This means that of the 27% of items that can (on average) be recommended to students, only 16% is recovered by the model. This decreases the amount of useful recommendations per student substantially. In situations where more material is available, prevalence becomes less important because there is still enough to recommend, but performance in terms of recall remains important because it impacts the coverage of the feedback over the content domain.

All in all, it seems reasonable to choose a threshold somewhere in the middle of the scale. Intuitively, a threshold of 0.5 makes sense: an item would be recommended to the student when the student is expected to score fewer than half of the points on the item (missing more than half of the possible points). The average precision in the test sets for the baseline model is 0.73, and the average recall 0.58. A reason for using a threshold of 0.50 as opposed to a threshold of 0.51 is a 6% drop in prevalence, which has an influence on the number of items that are relevant to recommend. In practice, the number of items

that are available to recommend can also be informative for choosing a threshold.

2.5 Discussion

The aim of the present study was to explore the use of recommender systems as a means of providing automated and personalized feedback to students based on their scores in a summative assessment, in the form of a set of test items that a student could benefit from practicing with. The results showed that in the context of the final examinations in Dutch secondary education, item recommendations can be made to students with an acceptable level of performance in terms of precision and recall. Furthermore, it does not take a complicated model to provide useful recommendations: a simple baseline model which takes into account global, student-specific and item-specific averages obtained similar performance to the more complex models (IBCF, UBCF and SVD approximation).

Recommended practice questions can be beneficial for students as they can support their learning process by targeting specific gaps in their knowledge, especially when there is little time to get feedback from instructors. More generally, recommender systems may be useful in any educational context where 1) direct teacher intervention is infeasible (e.g. due to scale or time constraints), 2) appropriate input data is available, and 3) it is plausible that the recommendation of practice material is useful to the student. They can be a powerful tool in guiding students in the direction of better understanding without putting additional strain on their teachers, by making use of data that is already available.

This study has taken a first step towards practical implementation by assessing the theoretical feasibility of applying recommender systems in a summative assessment context. An important step to take before implementation is to perform a validation study, in which the impact of recommending practice items on learning outcomes (e.g. increase in score during a resit exam, learning satisfaction, amount of studying) is assessed. Ghauth and Abdullah (2010) performed such a practical validation, by testing experimentally whether learning gain had improved more for the group that received recommendations. Theoretical evaluation can give us information about the quality of the predictions (an important condition), but it cannot give us information about the (perceived) usefulness of the provided feedback to the target group.

When preparing to use a recommender system in practice, one should evaluate the model performance in terms of evaluation metrics that are important for the specific context. Naturally, the level of performance on those metrics that is acceptable for practical use is debatable. Certainly, we would not wish for students to spend valuable time studying topics that they have already mastered. At the same time it is true that in summative assessment, the students often do not get any systematic feedback at all. To many students, especially to those that do not know where to begin, feedback may be helpful even if it is not perfect. In the context of preparing for a resit for secondary education final ex-

aminations, we expect students would be helped with any reliable personalized feedback, especially considering that the time pressure makes it difficult for their instructors to offer feedback on the performance of the student on their first exam.

A key aspect of recommender system algorithms is the measure that is optimized in order to provide recommendations. In this study, the measure that was optimized was the potential gain score, which is the estimated proportion of the maximum obtainable score the student is expected not to obtain. In other words: the items that are recommended to the students are the ones that are expected to be the most difficult for them. In this particular context, the students need to master all of the material, as they will be faced with a high-stakes summative assessment. Therefore, it is appropriate that they should practice the things they are not yet adapt with, regardless of the difficulty of those items. In other situations, being faced with the items that are predicted to be the most challenging for them may be an unnecessarily demotivating experience for students. The choice to optimize for different measures may lead to many other interesting applications of recommender systems in an educational context. Other potentially rich directions for future studies and applications include using meta-data of the learning materials in hybrid recommender algorithms, and including student non-response in the input data as potentially relevant information by using e.g. the SVD++ algorithm (Koren, 2008).

In this study, the performance of several recommender algorithms for recommending practice test questions to students was assessed. The results suggest that recommender systems can provide useful feedback to students, especially in contexts where teacher intervention is infeasible (i.e. due to time constraints). Many other directions for future research and applications are possible concerning the use of recommender systems in an educational context. Overall, we conclude that recommender systems are a promising tool for helping students in their learning process by combining multiple data sources and new methodologies.

2.6 Appendix 1: script for producing recommendations

```
import pickle
import pandas as pd
import numpy
import surprise as sp
import surprise.model_selection as spm
from evaluate import metrics, KFold_evaluate
```

```
df = pd.read_pickle("./df2.pkl")
reader = sp.Reader(rating_scale=(0, 1))
data = sp.Dataset.load_from_df(df[["KandidaatID", "ItemID",
"ToGain"]], reader)

# If the student has more than the cutoff to gain,
# the item is recommended.
cutoff = numpy.linspace(0, 1, 101)

# Number of folds for cross-validation
K = 3

# Number of predicted ratings to take into account.
n = [8]

# 1) Random Items

print("Random")
Random = sp.NormalPredictor()
evaluations = KFold_evaluate(data, Random, K, cutoff, n, "Random")

# 2) Baseline: Predicting baseline estimate for given user and item.
Baseline = sp.BaselineOnly()
evaluations = evaluations.append(KFold_evaluate(data, Baseline, K,
cutoff, n, "Baseline"), ignore_index = True)

# 3) User-Based Collaborative Filtering with Cosine Similarity
sim_options = {"nam": "cosine"}

UBCF = sp.KNNBaseline(sim_options = sim_options)
evaluations = evaluations.append(KFold_evaluate(data, UBCF, K, cutoff,
n, "UBCF"), ignore_index = True)

# 4) Item-Based Collaborative Filtering with Cosine Similarity
sim_options = {"name": "cosine",
               "user_based": False}

IBCF = sp.KNNBaseline(k = 10, sim_options = sim_options)
evaluations = evaluations.append(KFold_evaluate(data, IBCF, K, cutoff,
```

```
n, "IBCF"), ignore_index = True)

# 5) Singular Value Decomposition
SVD = sp.SVD()
evaluations = evaluations.append(KFold_evaluate(data, SVD, K, cutoff,
n, "SVD"), ignore_index = True)
```


2.7 Appendix 2: script for evaluating recommendations

```
from collections import defaultdict
from surprise import Dataset
from surprise import SVD
import pandas as pd
import numpy
import random
from surprise.model_selection import KFold
import surprise.model_selection as spm
import json

def Average(lst):
    return sum(lst) / len(lst)

def KFold_evaluate(data, algo, K, threshold, n, modelname):
    """
    Do KFold cross-validation given the data, an algorithm and threshold.

    parameters:
        data - data as accepted by the surprise package.
        algo - a prediction algorithm in the surprise package.
        K (int) - number of folds.
        threshold (list of floats) - What constitutes a 'good' rating.
        n (list of integers) - how many of the top predicted ratings
        to take into account.

    returns:
```

```
        evaluations - list (length K) of dictionaries.
    """

    col_names = ["n_rec", "precision", "recall", "fpr", "model",
    "fold", "n", "threshold"]

    results = pd.DataFrame(columns = col_names)

    # define a cross-validation iterator
    kf = spm.KFold(n_splits = K, random_state = 42)

    fold = 1

    for trainset, testset in kf.split(data):

        algo.fit(trainset)
        predictions = algo.test(testset)

        for t in threshold:
            for N in n:

                evalmetrics = metrics(predictions, threshold = t,
                n = N, fold = fold, modelname = modelname)
                evalmetrics["model"] = modelname
                evalmetrics["fold"] = fold
                evalmetrics["n"] = N
                evalmetrics["threshold"] = t

                results.loc[len(results)] = evalmetrics

        fold += 1

    return results

def metrics(predictions, threshold, n, fold, modelname):
    """Return evaluation metrics averaged over users."""

    # First map the predictions to each user.
    user_est_true = defaultdict(list)
```

```
for uid, _, true_r, est, _ in predictions:
    user_est_true[uid].append((est, true_r))

n_recs = dict()
precisions = dict()
recalls = dict()
fprs = dict()

json_dump = []

for uid, user_ratings in user_est_true.items():

    # Sort user ratings by estimated value
    user_ratings.sort(key=lambda x: x[0], reverse=True)

    tp = 0
    fn = 0
    fp = 0
    tn = 0

    for (est, true_r) in user_ratings[:n]:

        if (true_r >= threshold) & (est >= threshold):
            tp += 1
        elif (true_r >= threshold) & (est < threshold):
            fn += 1
        elif (true_r < threshold) & (est >= threshold):
            fp += 1
        else:
            tn += 1

    json_dump.append({"tp": tp, "fp": fp, "tn": tn, "fn": fn})

    # Number of recommended items (predicted positives).
    n_recs[uid] = tp + fp

    # Precision: Proportion of recommended items that are relevant.
    precisions[uid] = tp / (tp + fp) if (tp + fp) != 0 else 1
```

```
        # Recall: proportion of relevant items that are recommended.
        recalls[uid] = tp / (tp + fn) if (tp + fn) != 0 else 1

        # False positive rate
        fprs[uid] = fp / (fp + tn) if (fp + tn) != 0 else 1

    n_rec = sum(n_rec for n_rec in n_recs.values()) / len(n_recs)
    precision = sum(prec for prec in precisions.values()) /
        len(precisions)
    recall = sum(rec for rec in recalls.values()) / len(recalls)
    fpr = sum(fpr for fpr in fprs.values()) / len(fprs)

    evaluations = {"n_rec": n_rec, "precision": precision,
                   "recall": recall, "fpr": fpr}

    return evaluations
```

CHAPTER 3

Using a cognitive diagnostic modeling based recommender system for providing practice tests: effects on learning gain and student experience

This chapter has been submitted for publication.

De Schipper, E., Maas, L., Feskens, R., & Veldkamp, B. P. (2025). *Using a cognitive diagnostic modeling based recommender system for providing practice tests: effects on learning gain and student experience*.

Abstract

In this study, we introduce a theory-driven recommender system based on cognitive diagnostic modeling (CDM) to recommend personalized practice tests to students preparing for their final examinations. The CDM-recommender has the potential to provide pedagogically meaningful test question recommendations based on estimated domain mastery of students. The study consists of two parts. In the first part, we compare the technical performance of the CDM-recommender to those of other recommendation algorithms using historical exam data from Dutch biology assessments. The CDM-recommender performs as well as other recommender systems. The second part of the study describes an experiment in which students prepare for their high-stakes final exams using either a fixed or personalized practice test generated by the CDM-recommender. The two groups showed no significant differences in learning gain, and the students perceived the personalized tests as more difficult (which they were by design). These findings suggest that CDM-based personalization may not yield immediate benefits in high-stakes exam contexts without sustained engagement or learner agency.

3.1 Introduction

There are notable benefits to studying using practice tests for student learning and retention (Adesope et al., 2017; Bjork, Bjork, et al., 2011; Callender & McDaniel, 2009; Roediger III & Karpicke, 2006a), and students often find practice tests useful in preparation for exams (Cassady & Gridley, 2005). Practice tests not only enhance long-term memory through retrieval but also aid in identifying gaps in knowledge, especially when feedback is provided (Adesope et al., 2017). Despite these advantages, students frequently default to less effective strategies such as rereading and highlighting (Karpicke et al., 2009). A reason for this may be that students often struggle to accurately judge their own understanding and learning progress (Dunlosky & Lipko, 2007), potentially leading to students not knowing what material to practice or incorrectly selecting practice test materials.

The personalization of practice tests can further improve learning by providing students with tailored opportunities to bridge the space between their current understanding and their learning goals (e.g., Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008). Automated personalization systems can adapt to learners' needs without placing additional burden on instructors. Prior research shows that personalized practice tools can improve learner performance and intrinsic motivation (Alamri et al., 2020), particularly when students are given agency in selecting or shaping their tasks (Bernacki et al., 2021; Kulasegaram & Rangachari, 2018). Moreover, alignment in format between formative and summative assessments can facilitate transfer and recall of target information (Cassady & Gridley, 2005).

Recommender systems, which are widely used in commercial contexts to recommend products to users (Ricci et al., 2015), can offer a scalable approach to personalizing learning by recommending relevant educational materials. In recent decades, their use in educational settings grew steadily, particularly in online learning environments where they help recommend resources such as books, videos, and courses to learners (e.g., Aher & Lobo, 2013; Bobadilla et al., 2009; Chen et al., 2018; Ghauth & Abdullah, 2010; Huang et al., 2019; Khribi et al., 2008; Luo et al., 2010; Tang & McCalla, 2005; Zaíane, 2002; Zhang, 2024).

The application of recommender systems in educational assessment, and more specifically, in recommending practice test questions (*items*), remains limited, despite successful demonstration of this application by De Schipper et al. (2021). A plausible reason for the relative lack of literature on this application is the existence of a rich array of methods for the purpose of individualized assessment originating from the field of computerized adaptive testing (CAT, e.g., Van der Linden, Glas, et al., 2000). Recommender systems can offer additional flexibility by their potential to incorporate various modeling techniques, including the item response theory models commonly used in CAT, and their ability to

take contextual factors such as content domains into account.

In this study, we introduce a theory-driven recommender system based on psychometric estimates from the field of *cognitive diagnostic modeling* (CDM) to recommend personalized practice tests to students preparing for their final examinations. CDMs can differentiate between students with the same total sumscore on a test, offering additional information about their strengths and weaknesses through the estimation of mastery and non-mastery of different skills or attributes of interest (Ma & de la Torre, 2016). Unlike other data-driven recommender systems, which infer patterns from scores only, the CDM-recommender introduced in this study can incorporate the additional item metadata of what content domain is being covered by each item in a test. This enables more pedagogically grounded recommendations tailored to the content that a student has not yet mastered.

In the first part of the study, we introduce and explain the new CDM-recommender, and compare its technical performance to other recommendation algorithms on existing student data. In the second part, we deploy the developed recommender in an experimental setting to assess its practical impact on student learning outcomes and experiences. Students receive either a personalized or a fixed set of practice items in preparation for their final exam. This experiment is detailed in the second part of the study.

We address the following research questions:

- Can the CDM-recommender provide suitable personalized practice tests to students in the context of the Dutch secondary education final examinations?
- Do students provided with a personalized set of practice items achieve larger learning gains compared to students provided with a fixed set?
- Does the experience of taking the practice test differ for students in these two groups?

Students' progress is tracked through log data of pre-tests, practice tests, and post-tests, and supplemented with surveys capturing their experiences and perceptions.

3.2 Theoretical performance of CDM-recommender

This section discusses the methods, results and conclusion of the first part of this study, in which a new CDM-recommender is introduced and its technical performance compared to those of other recommender algorithms on existing student data.

3.2.1 Methods

3.2.1.1 Context and data

This study takes place within the context of Dutch secondary education, where students are required to take standardized final examinations, which account for half of a student's final grade per subject. Passing these high-stakes summative assessments is a prerequisite for admission to higher education programs. Some of the exams are made public each year, and students in later years can use these as material to practice with in preparation for their final examinations. The students are between 16 and 18 years old when they take the exams.

In the theoretical part of this study, we use historical data from two Biology exams from 2022 for students in a vocational track of secondary education (*vmbo-basis*) in order to test the performance of the recommendation algorithms. The input for the algorithms is a students' score per item. To quantify how much a student might benefit from additional practice on a given item, we computed a *potential gain score* by dividing a student's score by the maximum obtainable scores on the item and then subtracting the result from one. This metric represents the proportion of the maximum obtainable score which the student did not obtain on the item.

The dataset contained $1,448$ students and 72 unique exam items, of which 68 had a maximum score of one point and four had a maximum score of two points. Each student in the data has taken an exam which contained 36 items. The data contain $25,711$ potential gain scores of zero (where a student has achieved the maximum score on an item), 552 potential gain scores of 0.5 (where a student has achieved a score of one on an item with a maximum score of two) and $26,584$ potential gain scores of one (where a student has not scored any points on an item).

3.2.1.2 The CDM-recommender

The **CDM-recommender**, similarly to the other recommenders, uses data on a student-item level in order to predict unknown scores and to recommend items to students. However, it also incorporates item metadata detailing the content domain that an item covers. The Sequential G-DINA model (Ma & de la Torre, 2016) that is the foundation of the CDM-recommender needs a Q-matrix as input: this is a matrix that indicates which item measures which content domain. The model is used to estimate mastery profiles (indications on whether a student masters a domain or not), as well as item response probabilities for each possible profile a student could have. For each item, these response probabilities indicate the probability of a student obtaining a score given a specific mastery profile.

For each student, the sequential G-DINA model is used to estimate a mastery profile for six different content domains within Biology, such as *Plants, animals and their cohe-*

sion. For example, a student could have the profile 010011, indicating that the student has been estimated to master domains two, five, and six, but not domains one, three, and four. When we recommend practice items to a student, we select the items for which the model predicts the highest probability of getting the question incorrect, given the estimated mastery profile of the student. A disadvantage of the CDM-recommender is that a student must answer a sufficient number of items per domain in order to enable estimation of their mastery profiles. Advantages are that it has a low computational cost, and that it may provide more relevant content to the user due to taking more information into account.

3.2.1.3 Other recommendation algorithms

The performance of the CDM-recommender was tested on historical student data and compared to the performance of the following other recommender algorithms: a baseline model, an IRT-based recommendation algorithm (the *IRT-recommender*), a matrix factorization model using singular value decomposition (the *SVD-recommender*), and an algorithm that recommended items to students randomly. The latter was added as a benchmark model for the minimal performance to be attained.

The **baseline model** predicts a potential gain score by combining the overall average, student average and item average of the potential gain scores. The potential gain score for a student on an item is estimated as described by De Schipper et al. (2021). Such a baseline model is simple to implement, has a low computational cost, and requires no domain-specific knowledge from the user.

The **IRT-recommender** works similarly to the baseline model, but uses an extended nominal response model. In our case, this model reduces to the Rasch model for dichotomous items and partial credit model (PCM) for polytomous items. The item parameters are estimated using conditional maximum likelihood, after which the student parameters are estimated using maximum likelihood estimation. The student abilities and item parameters are then used to predict potential gain scores for the students on items which they have not faced. For dichotomous items, modeled with the Rasch model, the ordering of the potential gain scores is the same regardless of student ability. This means that the items with highest potential gain scores are the same items for each student. For polytomous items, modeled with the PCM, the ordering of the potential gain scores can depend on student ability. Therefore, the items with the highest potential gain score are not necessarily the same items for each student. The IRT-recommender is relatively simple to implement and has a low computational cost.

The **SVD-recommender** is a model-based matrix factorization algorithm (Koren et al., 2009), meaning that it reduces the potential gain scores of the student into scores on a smaller amount of unobserved latent factors, and predicts potential gain scores by using the factor loadings of the student and of the item on these factors. The idea behind such

models is that attitudes or preferences of a user can be determined by a small number of hidden factors. To this end, it reduces the rating matrix into two smaller matrices through extracting a smaller number of features. More details on this application of the SVD-algorithm can be found in the Surprise library documentation (Hug, 2020). Matrix factorization algorithms generally perform well in terms of accuracy when data is large and dense enough, but have a relatively high computational cost.

The **random recommender** predicts a random potential gain score for students on the items by drawing from a normal distribution with a mean and standard deviation that are estimated using the observations in the training data.

3.2.1.4 Evaluating the recommender algorithms

In order to test how well the recommender algorithms predict scores for students on unknown items, we used data from students who have taken these exams in the past and performed k-fold cross-validation. We can evaluate the model by comparing the predicated scores with the observed scores on the k test sets (Breese et al., 2013).

It is common in cross-validation for the data to be split into subsets randomly. This was not possible in the current study, because in order for the CDM-recommender to estimate mastery probabilities for the students on all domains, it needed observed scores on items from all domains for all students. To ensure that in each training set, each student had observed scores on items from all domains, we opted for a stratified sampling method in which about $1/3$ of the items per domain are removed in order to produce a test set. Each third of the data is used as a test set once, in the usual manner of k-fold cross-validation where $k = 3$.

The performance of the recommenders was evaluated in terms of precision and recall. Precision indicates the proportion of recommended items that were correctly recommended, whereas recall indicates the proportion of relevant (recommendable) items that were in fact recommended. The performance of the models in terms of these measures is also dependent on the threshold that we choose to recommend an item. An item is recommended if its estimated potential gain score exceeds a given threshold. To evaluate the algorithms across a range of sensitivities, we tested 21 threshold values between zero and one (in increments of 0.05). Higher threshold values improve the precision of a recommendation algorithm but reduce the recall, and vice versa.

In the context of students studying for their high-stakes central examinations, we argue that precision is more important than recall: a lower precision may lead to students unnecessarily spending time on topics which they have already mastered, while a lower recall may not be such an issue because teachers have other learning resources available that can be used. Low recall can be problematic when the total number of items that can be recommended is low, which is why we pay specific attention to this in the experimental part of this study. Lastly, we also inspect variability between the folds of the k-fold

cross-validation to examine the recommenders' sensitivity to changes in the data.

3.2.1.5 Software

Most of the code relating to the recommendation algorithms was written in the programming language Python 3 (Van Rossum & Drake, 2009), using the surprise library (Hug, 2020). The R programming language (v4.4.1; R Core Team 2025) was used for statistical testing as well as to make graphical representations of the data and the results. Furthermore, the IRT-recommender and CDM-recommender also made use of the R packages dexter (Maris et al., 2024) and CDM (George et al., 2016), respectively.

3.2.2 Results

3.2.2.1 Precision

All recommenders besides the random recommender scored similarly in terms of precision, as can be seen in Figure 3.1. For all thresholds, the difference in precision is negligible: translated to the student level, the difference between choosing one of the other three recommenders amounts to a difference of less than one item per student (in the current test data of 12 items per student). All recommenders outperform the random recommender at all thresholds except a threshold of zero, with higher thresholds having increasingly high performance.

3.2.2.2 Recall

Like their performance in terms of precision, all models besides the random recommender also score similarly in terms of recall, as shown in Figure 3.2. For most thresholds above 0.5, the IRT-recommender performs slightly better in terms of recall than the baseline and SVD-recommenders. For some of the thresholds starting at 0.5, the CDM-recommender slightly outperforms the other recommenders. Notably, the random recommender starts outperforming the other recommenders for high thresholds (starting at 0.7 for the baseline recommender, 0.75 for the CDM and SVD-recommenders, and 0.8 for the IRT-recommender).

3.2.2.3 Variability between folds

The CDM-recommender had the biggest maximum difference in precision ($\Delta_{precision} = 0.0482$) and recall ($\Delta_{recall} = 0.0859$) between the folds, both occurring at a recommendation threshold of 0.6. The reason for the largest differences between folds occurring for the CDM-recommender is most likely because some domains have very few items in the training set, leading to less reliable estimation of mastery profiles. The correlations

Figure 3.1: Recommendation algorithm performance in terms of precision using different values of the estimated potential gain score as thresholds for recommending an item.

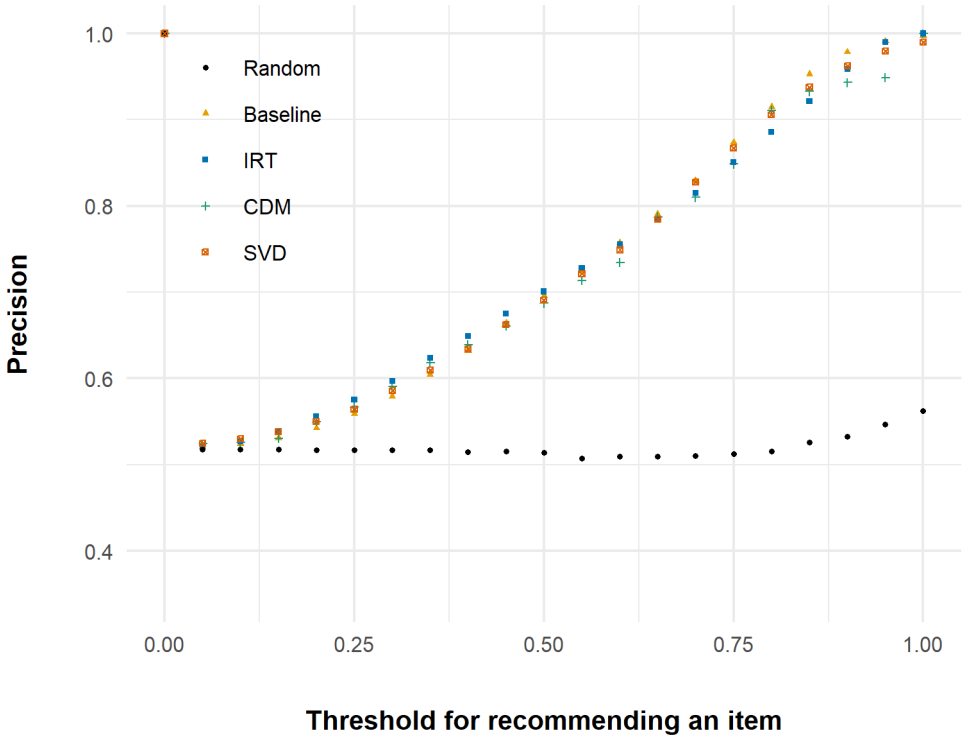


Figure 3.2: Recommendation algorithm performance in terms of recall using different values of the estimated potential gain score as thresholds for recommending an item.

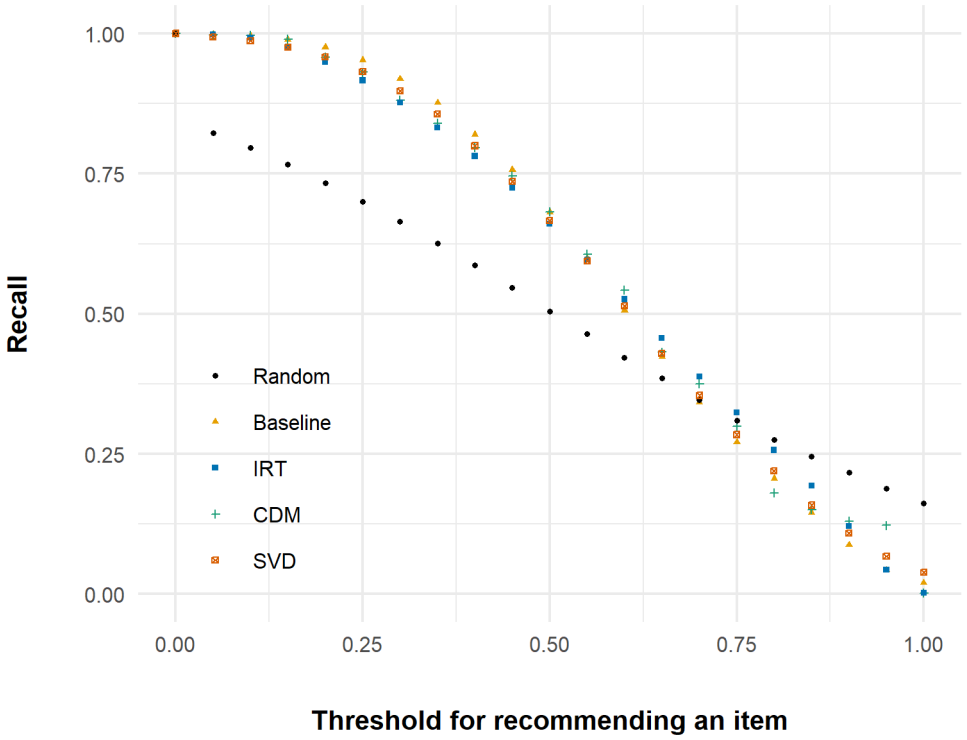


Table 3.1: Correlations between mastery status on each domain estimated using the complete data versus domain mastery status estimated using training sets 1, 2, and 3. The content domains are: Learning ability in the subject of Biology (K3), Cells are the base (K4), Plants, animals, and their cohesion (K6), Maintaining the body (K9), Responding to stimuli (K11), and From generation to generation (K12).

Training set #	K3	K4	K6	K9	K11	K12
1	0.518	0.567	0.636	0.566	0.598	0.663
2	0.400	0.537	0.645	0.614	0.434	0.654
3	0.529	0.623	0.684	0.676	0.689	0.714

Table 3.2: Correlations between item response probabilities per mastery profile estimated using the complete data versus item response probabilities per mastery profile estimated using training sets 1, 2, and 3. Here, π_j indicates the conditional probability to obtain score j given a certain mastery profile.

Training set #	π_0	π_1	π_2
1	0.983	0.988	0.918
2	0.976	0.982	0.875
3	0.987	0.988	0.999

between estimated mastery statuses per domain in the training sets versus the complete data shown in Table 3.1 give some indication on the extent to which the number of items per domain influenced the quality of the CDM recommendations. Naturally, the more items are removed from the training set, the lower the correlations between the results of the two models. If the item set that is used to estimate mastery profiles increases, the estimates will be more reliable, which is likely to result in better recommendations. This also means that the results that are found for the CDM-recommender in the current study are a minimal performance measure, not a realistic portrayal of the performance in practice.

We also correlated the item response probabilities π_{jc} (the conditional probability to obtain score j given a certain mastery profile c) that were estimated based on each training set with item response probabilities that were estimated based on the complete data. We found high correlations (all between 0.9 and 1, as shown in Table 3.2), indicating that the training sets and complete data led to highly similar item parameter estimates.

The SVD-recommender is known to also be impacted by the size (Koren et al., 2009) and composition (Adomavicius & Zhang, 2012) of the training data. Both the baseline model and IRT-recommender are less impacted by the size and composition of the training sets, because they estimate fewer model parameters, and are able to use all available items that the students have answered for each estimate.

3.2.3 Conclusion

All recommenders outperform the random recommender for most chosen thresholds in terms of both precision and recall. The CDM-recommender performs similarly to all other (non-random) recommenders. The performance of the CDM-recommender is expected to improve with larger amounts of data per domain. We conclude that the CDM-recommender shows adequate performance and can therefore be employed in the experimental part of this study to construct practice tests for the participating students.

3.3 Experimental study: the effects of the CDM-recommender in practice

This section discusses the second part of this study, in which we implemented the CDM-recommender in an experimental real-life situation for recommending a personalized practice test to students, and assessed its effects in terms of student learning gain and student experience.

3.3.1 Methods

A web-based application called *Biologie+* was built for the purpose of this study, allowing us to control which students were placed in which experimental condition and to collect log data on the behavior of the participating students. In Biologie+'s teacher environment, the teacher could log in with their email address, add students, unlock tests for the students, and view reports based on the results of their students (see Figure 3.3). In the student environment, students could log in with a personal code that was provided in the teacher environment when the teacher added students (see Figure 3.4), take the tests that had been unlocked by their teacher (see Figure 3.5), and view reports based on their own performance on the pre- and post-test in the application.

Students logged into Biologie+ during three separate Biology lessons. During the first lesson, students took a test which was a central examination from an earlier year. The students' scores were used as calibration data for the CDM-recommender in order to assemble personalized practice tests for students in the experimental condition. Furthermore, this test was used as a pre-test for the purposes of this study. On the second occasion, students studied the exam material by taking either a personalized or fixed practice test depending on the student's randomly allocated experimental condition. The construction of the practice tests is discussed in Section 3.3.1.1. After this second session, the students were administered a short survey on their experiences with the difficulty and the content of the practice test, as well as on how they felt during the session. During the third session, the students took another whole exam, which was used as a post-test

Figure 3.3: An example of a student report available to a teacher using Biologie+, showing a grade for the entire test, the proportion of points the student has achieved for each domain, and an indication of the proportion of the points that most students in the past achieved for the same domain on the same test.

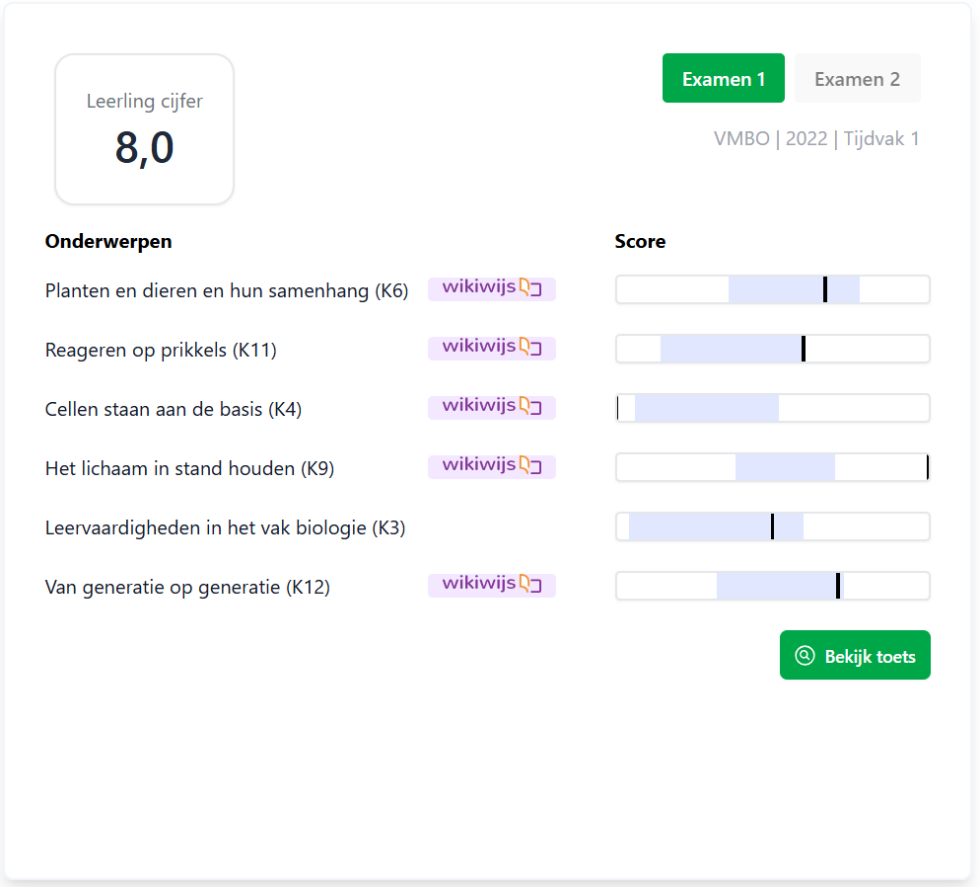
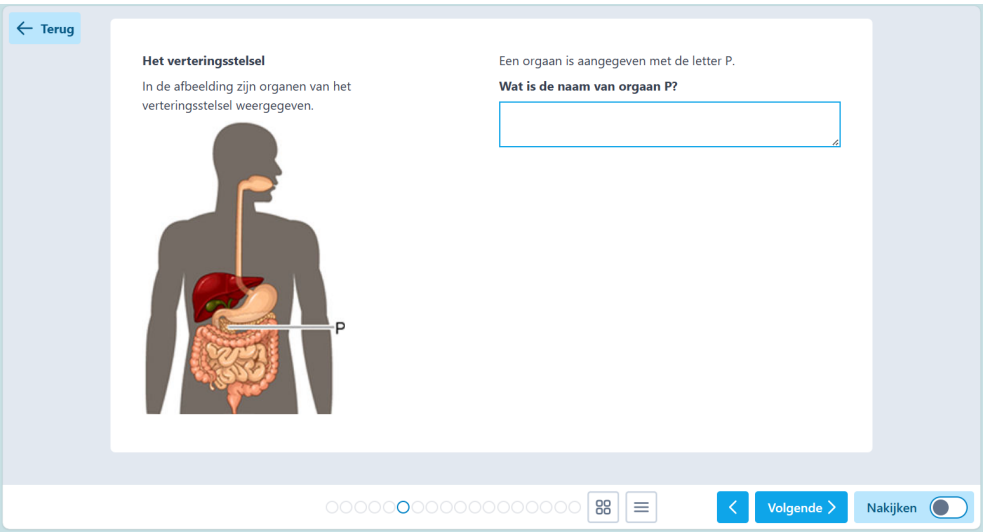


Figure 3.4: An example of the landing page for students in Biologie+. On this page, students could navigate to the next test (when enabled by their teacher), as well as view their rapports for completed tests.

Figure 3.5: An example of the student-facing test administration environment in Biologie+. In this environment, students could navigate through a test, view the items, respond to items, and score their answers.



for the purposes of this study. The practice test for the experimental condition that was *not* allocated to the student (e.g., a personalized test for students that had been in the control condition) was provided as a supplementary test in the application, to ensure that no student was put at a learning disadvantage regardless of the outcome of the current study. Students were free to use this test in their studies as desired, and their data were not used for this study.

During the pre- and post-test, students were instructed to take the test individually without help from their teacher, their learning materials or the internet. During the practice tests, students were allowed to use all of these resources and instructed to make use of the opportunity to study the material. However, in order to prevent interaction effects within the study, students were not allowed to discuss the material with their peers.

3.3.1.1 Construction of practice tests

All practice tests consisted of 18 items, which is approximately half of the length of a regular central examination for Biology for these students. The reason that the practice test was made shorter than the regular exam was to give the students time to delve into the learning material which the items were about. Students in the control condition were given a practice test that was assembled prior to the study. The contents of the fixed practice test were designed to be comparable to the students' upcoming central examinations in terms of 1) the ratio in which the items spanned different topics, 2) the ratio of open-ended to closed-ended items, 3) the average difficulty of the test, and 4) the standard deviation of the difficulty of the test.

The personalized practice tests for students in the experimental condition were constructed by using the CDM-recommender to choose 18 items from a set of 85 available items. These 85 items were in the central examinations in the years 2021 and 2023, had no known performance issues, collectively covered all content domains, and had an average p-value (proportion correct) of 0.47. Most of the items (78) had a maximum score of one. Seven items had a maximum score of two. Only one item of this set was included in the fixed practice test, so overlap between the practice tests in the experimental conditions was minimal. In order to assemble a personalized practice test for the students in the experimental condition, the CDM-recommender estimated mastery profiles for the students using the scores on the pre-test, used these to estimate the potential gain scores on the available 85 items, and obtained the 18 items with the highest potential gain scores. The Q-matrix included only unidimensional items (each item measured only one domain). The pre-test which was used to estimate the mastery profiles contained two items for domain K3, four items for domain K4, six items for domain K6, 12 items for domain K9, five items for domain K11, and seven items for domain K12. A more detailed description of the CDM-recommender can be found in Section 3.2.1.2.

Since 50.3% of the potential gain scores in the historical student data were score 1,

and 1% were score 0.5, the prevalence of items that were relevant to recommend was 51.3% with a threshold of 0.5 or lower and 50.3% with a threshold higher than 0.5. Combined with the results that were achieved in the theoretical study in terms of recall, we expected that there would be, on average, enough items available to be recommended to the students as part of their practice test.

3.3.1.2 Learning gain

The student's scores on the pre- and post-tests were converted to a grade between one and ten (ten being the highest), taking the difficulty of the tests into account by adjusting the calculation with the method that is also customary in the student's final examinations. Grades on the pre-test were calculated as follows: $grade = (score/maximum) * 9 + N$, where $N = 1.6$ for the pre-test and $N = 1.4$ for the post-test. The difference between the constants N in these calculations reflects the difference in difficulty between the tests and ensures comparability between the pre- and post-test grades. Learning gain was defined as the difference in grade between the post-test and pre-test. Note that negative learning gains were possible.

3.3.1.3 Student experience

Five-point Likert scales were used to measure the students' perceived difficulty of the practice tests (very easy, easy, neutral, difficult, very difficult) and whether the students thought that the practice test concerned topics that they found difficult (completely disagree, disagree, neutral, agree, completely agree). These variables were considered continuous for the purposes of the analyses. The students were also asked how they felt while taking the practice test using an open-ended question. Their answers were categorized independently by two raters into negative, neutral, or positive. The two raters initially agreed 74% (39 out of 53) of the time, and full agreement was reached after discussion.

3.3.1.4 Statistical analyses

One-tailed student's t-tests for independent samples ($\alpha = .05$) were used to test for the effect of experimental condition on learning gain, perceived difficulty of the test, and perceived topics in the test. The hypotheses were that students in the experimental condition would show higher learning gains, perceive the practice test as more difficult, and perceive the practice test to contain more topics that they find difficult compared to students in the control condition. A Chi-square test of independence was used to test for a relationship between experimental condition and how the student felt during practicing.

Students with missing values on variables relevant to the analyses were removed. 29.8% of students (25.0% in the control condition and 34.7% in the experimental condition) who participated in the pre-test did not participate in the post-test and where thus removed from the analysis testing for the effect of experimental condition on learning gain. 3.8% of students who participated in the survey on their experiences with the practice test (2.1% in the control condition and 6.3% in the experimental condition) were removed from the analysis testing for the effect of experimental condition on perceived difficulty of the test, and 7.5% of students (6.3% in the control condition and 9.4% in the experimental condition) were removed from the analysis testing for the effect of experimental condition on perceived topics in the test.

3.3.2 Results

3.3.2.1 Learning gain

Students in the control condition had a slightly higher average learning gain ($M = 0.14$, $N = 57$) than students in the experimental condition ($M = 0.04$, $N = 49$), but the effect was not significant ($t(102) = 0.36$, $p = .64$). There was more variance in learning gain in the control condition ($SD = 1.62$) than in the experimental condition ($SD = 1.19$).

3.3.2.2 Student experience

Students in the experimental condition ($N = 32$) experienced the practice test as significantly more difficult ($t(68) = 2.96$, $p = .002$) than students in the control condition ($N = 48$). The effect size of $d = 0.67$ is large in the context of an educational intervention (Kraft, 2020). The distributions of the students' perceived difficulties can be seen in Figure 3.6.

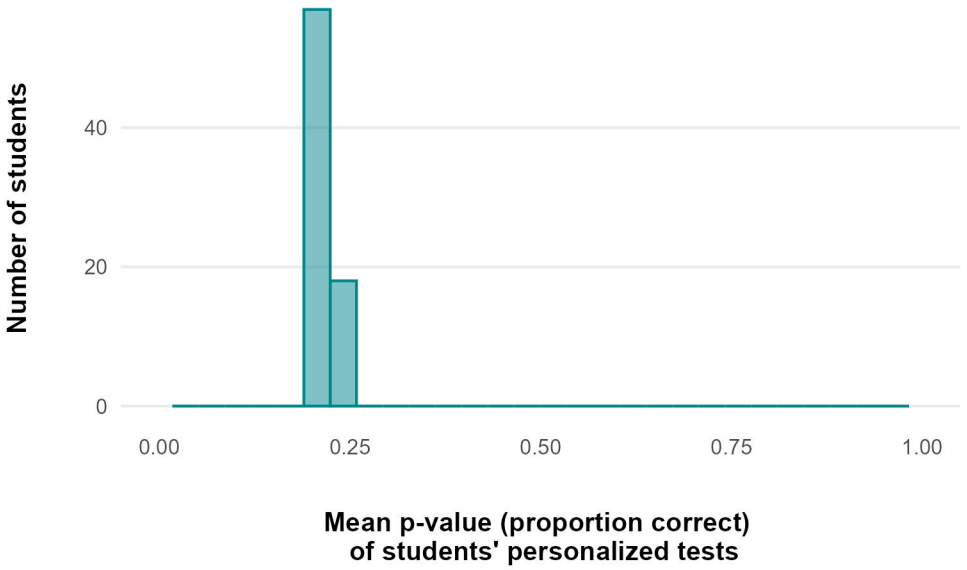
The average p-values of the personalized practice tests for the students in the experimental condition can be seen in Figure 3.7. The average p-value of the standard practice test was 0.47.

There was no significant difference between the control ($N = 45$) and experimental condition ($N = 29$) in whether the students felt like the practice test concerned topics that they found difficult ($t(72) = 0.47$, $p = .68$). There was no significant relationship between the experimental condition and how students felt during the practice test ($\chi^2(2, N = 53) = 4.10$, $p = .13$, $N_{control} = 31$, $N_{experimental} = 22$).

There was a higher percentage of non-response within the experimental condition. We were able to compute learning gain for 65% of the participating students, compared to 75% in the group of students in the control condition. Furthermore, only 35% of students in the experimental condition completed the survey after the practice test, versus 54% of students in the control condition.

Figure 3.6: The perceived difficulty of the practice tests for students in the control and experimental conditions.

Figure 3.7: Average p-values (proportion correct) of the personalized tests for students in the experimental condition.



3.4 Discussion

In this study, we developed a CDM-based recommendation algorithm and evaluated its performance in assembling personalized practice tests to students. The CDM-recommender showed adequate performance and was therefore employed in an experimental intervention in which students preparing for their central examinations were given either a personalized set of practice items generated by the chosen recommender, or a fixed set of items. Contrary to our hypothesis, we did not find a significant difference in learning gain between the experimental groups. In line with our expectations, students who received personalized practice tests reported finding them more difficult than the fixed set.

Our findings with regards to learning gain indicate that while CDM-based personalization of practice tests can successfully identify more challenging content for the student, this does not necessarily translate to immediate improvements in learning outcomes. Since it is known that being faced with very difficult learning materials can hamper students' motivation (e.g., Ryan & Deci, 2000), the benefits of producing practice tests this way would have to be convincing to justify applying them in summative assessment contexts. No validated measure of motivation was collected during the course of this study. However, it is imaginable that the higher non-response in the experimental condition could in part be due to lack of motivation.

Several factors may explain the lack of a difference in learning gain between the experimental and control condition. Since the limited sample size impacts the statistical power of the study, it is possible that there was an effect but we were not able to observe it. There are also reasons to believe that any effect taking place due to our experimental intervention would be a small effect (if only because that is to be expected for educational interventions, see Kraft (2020)). For example, although the students were presumably quite motivated to do well (due to the high-stakes nature of their central examinations), it is likely that the experimental setting in this study led to typical performance for the students as opposed to the maximum performance they are more likely to display during their final examinations (Klehe et al., 2018). Besides that, students may also need time to adjust to personalized practice (Bransford et al., 2000; Shute, 2008), which was not accounted for in the experimental design. Additionally, by the time of the study, students had been long preparing for their exams, so most learning had occurred before the intervention. A longitudinal study with a larger sample size should be considered in order to be able to observe any smaller and more long-term effects. Lastly, it is to be expected that the students in both the experimental and the control condition also learned in ways unrelated to the experimental intervention during the course of the study.

Of course, as the effectiveness of personalized practice heavily depends on the specific nature of the material and the context in which it is applied (Heritage, 2007; Shute, 2008), it is also possible that the type of personalization applied here was not the most effective in achieving learning gain in this context. Additionally, personalized practice tests that are made to deliberately target knowledge gaps, although tailored to individual needs, may inadvertently increase cognitive load, making it more difficult for students to process and learn from the recommended items (Sweller, 1988). The lack of improvement in learning gain may therefore reflect a misalignment between student readiness and the level of the challenge presented. With respect to both cognitive load and motivation, it may be more effective to optimize for a different measure when recommending learning materials. Exploring the balance between cognitive load and question difficulty in personalized practice tests seems essential to ensure that recommendations effectively support student learning without overwhelming them (Van Merrienboer & Sweller, 2005).

Interestingly, in contrast to our intent to save students time by helping them avoid content already mastered, student and teacher feedback indicated a preference for *more* practice materials, not fewer. This suggests that the assumption that personalization should prioritize efficiency may not align with user preferences or needs. In high-stakes exam contexts, students and teachers may value comprehensiveness and reassurance over strictly targeted practice. This highlights the need for more user research in developing personalized learning tools.

We look forward to seeing future studies on this topic in many useful and interesting directions. Including more diverse data sets could also help us better understand the

efficacy of recommender systems across various subjects and educational contexts. For example, studies could examine whether different groups of students (e.g., low vs. high prior achievers) benefit differentially from personalized practice. Additionally, combining recommender systems with adaptive learning technologies may better align recommended items with individual learning trajectories (Brusilovsky & Millán, 2007). Incorporating student feedback into such adaptive systems could further refine the recommendations and enhance the learning experience (Hattie & Timperley, 2007). Another possibility is to allow the student agency in the selection of the practice materials, as suggested by Kulasegaram and Rangachari (2018) and Bernacki et al. (2021). Providing opportunities for choice may help optimize for both learner performance and engagement. Aligning the learning materials with the students' interests has also been known to improve interest in learning, learning experience, motivation and engagement (Bernacki et al., 2021; Walkington, 2013). Lastly, integrating affective and motivational measures in studies could help understand how students perceive and respond to personalization over time.

While this study did not find significant short-term learning gains from CDM-based personalization of practice tests in summative exam preparation, it provided insights into its impact on student learning outcomes and experience. These insights contribute to the growing body of research on adaptive learning and aligning personalization with learners' needs, preferences, and educational goals. While such tools can offer precise targeting of content, their practical value depends on how well they are integrated into broader instructional and motivational frameworks. Educational practitioners should be involved in the design and evaluation of personalization systems in order to enhance both their effectiveness and their acceptance, and to ensure they provide an optimal balance of challenge and support for students (Shute, 2008).

CHAPTER 4

Identifying students' solution strategies in digital mathematics assessment using log data

This chapter has been published as:

De Schipper, E., Feskens, R., Salles, F., Keskpaik, S., Dos Santos, R., Veldkamp, B., & Drijvers, P. (2025) Identifying students' solution strategies in digital mathematics assessment using log data. *Large-scale Assessments in Education, 13*(1), 23. https://doi.org/10.1186/s40536-025-00259-6

Abstract

Background: Students take many tests and exams during their school career, but they usually receive feedback about their test performance based only on an analysis of the item responses. With the increase in digital assessment, other data have become available for analysis as well, such as log data of student actions in online assessment environments. This paper explores how we can use log data to extend performance-related feedback with information related to the applied solution strategy. **Methods:** First, we performed an exploratory model-based cluster analysis in order to identify the solution strategy of 802 students with a modal age of 14 in a pre-algebra item from the French national assessment *CEDRE*. Second, we related the students' solution strategies to their mathematical ability based on the entire assessment. **Results:** Five distinct groups of students with different in-assessment behavior were identified, of which one group had a significantly lower estimated mathematics ability than the other groups. **Conclusion:** These findings can provide a basis for more in-depth feedback and further instruction on the level of an individual student and can inform teaching practices at the class level.

4.1 Introduction

Advancements in computer-based assessment in the past decades have focused on providing a direct measurement of complex concepts or skills (Pellegrino & Quellmalz, 2010). Now that test designers move to take advantage of the wide array of unique possibilities on digital platforms and devices, new technology-enhanced item types are emerging and assessment environments are becoming more than a translation of existing paper-based assessments. Technology-enhanced assessment offers new opportunities for just-in-time and personalized feedback contingent on the student's progress, as well as more adaptive assessments (Timmis et al., 2016).

With the digitization of the educational assessment domain and the rapid developments in technology in general also comes a higher availability of student data, as well as the possibility of storing more meaningful data concerning the test-taking process and situation than just student responses and the accompanying scores. This type of data is referred to as *process data*. Examples of process data include gestures, eye movements, and physiological responses. A specific category of process data is *log data*, which consists of the logged interactions that students have with a digital testing environment, such as mouse movements, mouse clicks, keystrokes and their timing (e.g., Kröhne & Goldhammer, 2018; Lindner & Greiff, 2023; Maddox, 2023). Log data has long been stored and analyzed in the field of web analytics and more recently within digital learning systems as well (e.g., Cetintas et al., 2009a, 2009b; Wang, 2021). In recent decades, studies that make use of log data have begun to emerge in the field of educational assessment. One common application is the use of response times for improving estimates of ability (e.g., Bolsinova & Tijmstra, 2017; Fox & Marianti, 2016; Van der Linden et al., 2010). Studying student behavior (either using log data or other methods such as observational studies) can teach us about test design and student learning processes (Shute et al., 2009).

ILSAs also increasingly make use of digital-based interactive and technology-enhanced items (e.g. PISA (OECD, 2024), TIMSS (von Davier et al., 2024), ICILS (Heldt et al., 2020), and PIAAC (He, Borgonovi, & Paccagnella, 2019)). Process data of these items potentially contain valuable information about interindividual and intraindividual differences in response processes. However, how to effectively incorporate process data into ILSA analyses remains unclear, despite the growing number of studies on this topic in recent years (e.g., Goldhammer et al., 2017; Tang et al., 2020; Ulitzsch et al., 2021).

4.1.1 Log data research in educational assessment

Research into log data in educational assessment has for a large part so far focused on the topics of game-based assessment (Chen et al., 2020; Cui et al., 2019; Kerr & Chung, 2012), response motivation (Nagy & Ulitzsch, 2022; Nagy et al., 2022; Pokropek et al., 2023; Ulitzsch, Yildirim-Erbasli, et al., 2022) or rapid guessing (Deribo et al., 2023; Nagy

et al., 2022), and the assessment of (complex) problem-solving ability (e.g., Chen et al., 2020; Greiff et al., 2016; He & von Davier, 2015; Stadler et al., 2019; Ulitzsch, He, & Pohl, 2022). Studies on the assessment of (complex) problem-solving ability suggest that log data can significantly enhance our understanding of students' cognitive and behavioral processes.

Approaches to log data analysis in the field of (complex) problem-solving can roughly be divided into data-driven or theory-driven methods (or a combination of the two). Log data are very suited to data-driven analysis methods (Chen et al., 2020), and many researchers have consequentially approached the analysis of log data in a bottom-up fashion (Chang et al., 2017; Chen et al., 2019; He, Borgonovi, & Paccagnella, 2019; He, Liao, & Jiao, 2019; He & von Davier, 2015, 2016; Ulitzsch, He, & Pohl, 2022; Xiao et al., 2021). An advantage of data-driven log data analysis methods is that they can generally be generalized over items or tasks (He et al., 2021).

However, for the results of such analyses to be useful in practice, it is important to bridge the gap between the data and didactic theory (Gobert et al., 2013). For this reason, others have opted for a more theory-driven and top-down approach, in which hypotheses about the relations between certain behaviors and outcomes of interest are formulated in advance (Albert & Steinberg, 2011; Chen et al., 2020; Eichmann et al., 2019; Goldhammer et al., 2021; Greiff et al., 2016; Han et al., 2019; He et al., 2023; Jiang et al., 2023; Westera et al., 2014). These theoretically relevant behaviors are often operationalized by (aggregated) events in the log data into *single unit measures*, and subsequently analyzed using standard statistical methods. It is possible to aggregate some sequential information into such single unit measure, but this approach essentially does not take the order of student interactions into account, which poses an important disadvantage. However, there are recent examples of studies using theory-driven approaches that also take the order of student interactions into account. For example, Jiang et al. (2023) and He et al. (2023) compared sequences of log events to expert reference sequences using sequence mining, and Zhang and Andersson (2023) made use of network analysis to represent the order in which students perform operations.

Other recent studies have combined a more top-down, theory-driven approach with data-driven techniques (Eichmann et al., 2020; He et al., 2021; Stadler et al., 2019). For example, Eichmann et al. (2020) categorizes student actions into several didactically meaningful categories before doing a sequence-based analysis. An advantage of this approach is that the analysis allows for students to display behaviors from several different categories (instead of being limited to one category). However, for this approach to be feasible, it is necessary that student actions can be assigned to one (and only one) category, which is not always the case.

4.1.2 Log data research in mathematics assessment

As demonstrated by the studies described above, analyses of log data have led to new insights in the assessment of (complex) problem-solving. In different areas of study, log data could lead to similar advances. In the area of mathematics education, the recent decade has seen a steady increase of studies taking log data into account to advance our understanding of student cognitive and behavioral processes. Some of this work has taken place in the context of online learning or tutoring environments (Derr et al., 2018; Gobert et al., 2015; Hrastinski et al., 2021; Kerr & Chung, 2012; Martin et al., 2015; Olsher et al., 2023), whereas others have analyzed log data coming from collaborative, formative or summative assessment (Araneda et al., 2022; Faber et al., 2017; Jiang et al., 2021, 2023; Mohan et al., 2020; Reis Costa et al., 2021; Salles et al., 2020). Several of these studies have related variables derived from log data to student achievement or success on tasks (Araneda et al., 2022; Derr et al., 2018; Faber et al., 2017; Mohan et al., 2020; Salles et al., 2020). Other interesting applications included using time-on-task to improve the precision of ability estimates (Reis Costa et al., 2021), relating questions that were posed in an online tutoring environment to perceived satisfaction and learning (Hrastinski et al., 2021), using variables derived from log data to identify students at risk of dropping out (Derr et al., 2018) and using sequence mining on keystroke sequences to relate onscreen calculator use to student proficiency (Jiang et al., 2023).

A few studies aimed to identify student solution strategies or error patterns: Kerr and Chung (2012) did so in the context of educational video games and simulations, and both Jiang et al. (2021) and Salles et al. (2020) did so in the context of a large-scale national mathematics assessment. Learning about the different solution strategies that students take in solving items in mathematics assessments, and which strategies are effective in what situations, can provide a basis for feedback towards students and teachers to inform the learning process (Zhang & Andersson, 2023), since it constitutes feedback on possible improvements (Shute et al., 2009).

4.1.3 Theoretical framework and research questions

In this study, we examine different solution strategies that students can use on an interactive mathematics item, both from a theoretical and an empirical viewpoint. We theoretically distinguish the possibility for an algebraic solution strategy as well as for a numerical trial and error solution strategy on the item used in this study. The theoretical framework for student solution strategies used in this study was introduced by Sfard (1991), and defines a distinction between an operational and a structural approach to mathematical concepts.

In the operational approach, the student views a mathematical concept as a process. A numerical trial and error solution strategy is considered an operational approach. In the

structural approach, the concept is viewed as an object with its own characteristics, which can be compared to other objects of the same type. An example of the distinction in these views is the notion of algebraic expression. An algebraic expression can be seen as a series of operations that lead from input to output, or as an object with characteristics such as being quadratic, being equivalent to another one, or being symmetric in its variables. An algebraic solution strategy is considered a structural approach. It is suggested that students start viewing mathematical concepts from a more structural perspective towards the higher secondary grade levels.

Within this study we aim to identify which of these two approaches (if any) students have taken on a digital mathematics assessment item. To this end, we have derived meaningful variables from student log data based on this theoretical framework to answer the following two research questions:

1. Do students use an algebraic (structural), a numerical trial and error (operational), or a different type of solution strategy while solving a digital mathematics assessment item?
2. What is the relationship between student mathematical ability and their solution strategy?

In this study, we report on the results for these research questions and expand on the methodological challenges that we faced along the way.

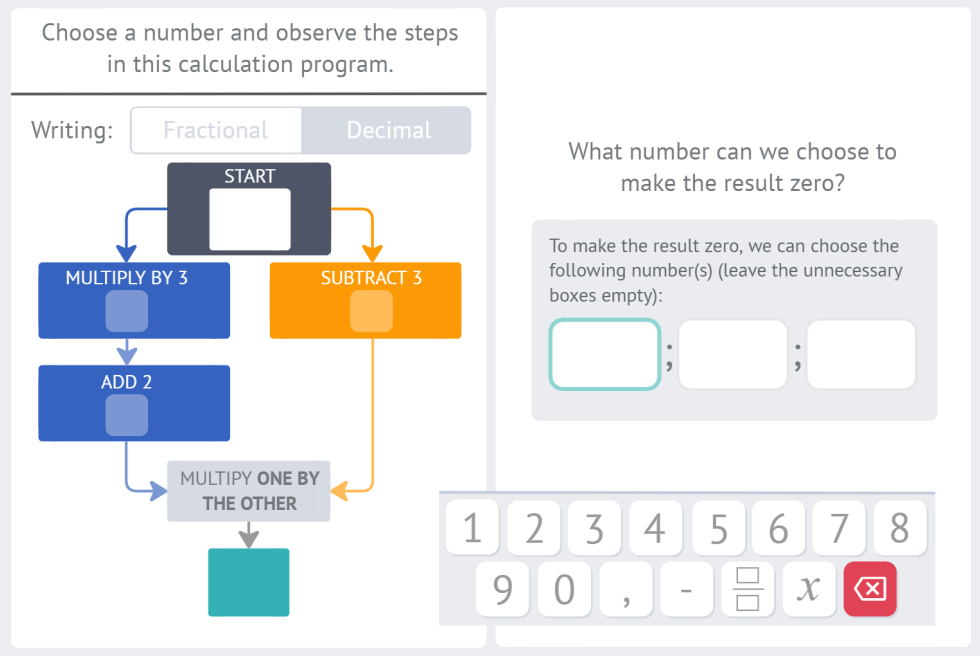
4.2 Methods

In this study, we looked at an item from French national mathematics assessments in grade nine as a test case to develop methods for analyzing mathematical log data in a way that leads to didactically meaningful inferences about a student's solution strategy.

4.2.1 The Product Equation Item

The item that was analyzed in this paper is called 'Product Equation', and is displayed in Figure 4.1. The question that has to be answered by the student is which number they have to choose for the result of the calculation on the left to become zero. The final responses are entered into the input fields on the right. In the input field on the left, the student can try different values and the result of the equation is calculated for them. When the student enters a value into the field, the left side of the calculation multiplies the value by 3 and then adds 2. The right side subtracts 3 from the input value. These intermediate results are then multiplied to produce the final result of the calculation. If the result is 0, the starting value the student has entered is a correct response to the item. The students also have a pencil tool, eraser, measuring tool, graphing tool, and

Figure 4.1: The 'Product Equation' test item.



Note. This test item has been translated from French to English for the benefit of the reader.

calculator at their disposal, although none of these tools are necessary to solve the item. The correct responses to the item are "3" and "$-2/3$".

The scoring for the item was dichotomous and without the possibility of partial credit, meaning that the students could either score zero points or one point on the item in the assessment. The student was awarded the point if they correctly identified one of the two possible correct answers, either in the answer fields or in the input field of the calculation program on the left. Numbers within 0.01 distance of either of the two correct answers were also deemed correct.

The two main strategies that can be followed for solving this item are a trial and error (operational) approach and an algebraic (structural) approach. In a trial and error approach, the student tries out different input values in the calculation program to find the correct one(s) by trial and error. The student may stop after the first correct response has been found, or may continue to search for another correct response. Students who use an algebraic approach realize that the calculation program translates to the expression $(3x+2)(x-3)$. To find its solutions, they should be aware that for a product to be zero, one of the factors should be equal to zero (or both, but that does not apply here). These students may input an "x" as starting value into the program, which then outputs the expression $(3x+2)(x-3)$, and solve the equation either mentally or using pen and paper.

As such, these students may spend more time away from the assessment environment.

4.2.2 Data collection

The item that was analyzed in this paper is part of the assessment cycle called CEDRE (Cycle des Évaluations Disciplinaires Réalisées sur Échantillon). CEDRE is a French national sample-based low-stakes assessment created and organized by the Department for Evaluation, Prospective and Performance (DEPP). DEPP is the entity within the French ministry of education (Le ministère de l'Éducation nationale, de la Jeunesse et des Sports) that is responsible for national educational measurement. CEDRE was created to measure the level of (among other abilities) mathematics among different age groups in the French educational system, as well as to provide a testing ground for innovations in national educational assessment.

In recent years, DEPP has developed technology-enhanced interactive item types for computer-based assessment to measure specific skills within the mathematical domain. Specifically, they developed items which engage higher-order mathematical thinking (as opposed to more rudimentary arithmetic skills) by having the assessment environment do calculations for the student (Salles et al., 2020). It is then up to the student to use the environment and its results to draw conclusions and respond to the item.

In total, the CEDRE administration in 2019 encompassed 348 items, administered to $7,992$ students in 309 schools. The assessment had 30 items in a linked design with 13 test booklets, followed by a second, multistage test and a context questionnaire. The assessment took place in the digital TAO test environment, with which the participating students were familiar beforehand. It was administered to students in ninth grade (*troisième* in the French school system) in both public and private sector schools. The modal age of the participating students was 14 years. Taking part in the assessment was obligatory for students in the participating schools. The test item analyzed in this paper was administered to $1,004$ students. The students were provided with pen and scrap paper.

The students' abilities were estimated with a two-step procedure using the scores on all items in the assessment with a two-parameter logistic (2PLM) model. In the first step, item parameters were estimated using Marginal Maximum Likelihood (MML) with an Expectation-Maximization (EM) algorithm. In the second step, item parameters were kept fixed and student abilities were estimated using Warm's Weighted Likelihood Estimation (WLE). This two-step procedure was used to allow for the comparison of results across different test cycles. More detailed information can be found in the assessment's technical report (Philbert et al., 2022).

The log data resulting from the aforementioned assessment consisted of an identification variable for the test taker, the time and date of the administration, a final state for all the components in the item environments (e.g. a value in the case of an input

field), and all actions the students have taken in the item environment accompanied by a timestamp.

The following steps were undertaken to clean the data. Students who did not interact with the assessment environment at all were removed from the data, as well as students for which no score, estimated ability, or response time was registered. Another five students were removed from the data due to a technical malfunction in the registration of their response times. After cleaning, the data consisted of 802 students.

The interactions derived from students' input into one of the textual input fields in the item interface (the input field of the calculation program, see Section The Product Equation Item) were also cleaned. Specifically, two types of interactions were removed: *construction events*, interactions that were registered while typing a longer sequence of characters, and *deletion events*, interactions that were registered while deleting input from a field. For example, if a student typed "-3" into one of the fields and subsequently deleted this input, four interactions would be registered (respectively): "$-$", "-3", "$-$", and an empty string. The first of these interactions is a construction event and the last two are deletion events, and would thus have been removed while cleaning the data.

4.2.3 Variable construction and conjectured relationships to solution strategy

Variables were constructed from the log data based on the expected student behavior for the algebraic solution strategy, or structural approach (Sfard, 1991), and for the numerical trial and error solution strategy, or operational approach (Sfard, 1991). These variables were used to see if we could identify which approach a student took. Table 4.1 describes how each variable was constructed and reflects conjectures about underlying solving strategies for different student behavior displayed in the variables. The variables are listed in order of the strength of the conjectures.

Table 4.1: Constructed variables and conjectured relationships to solution strategy.

Name	Type	Description	Expectation
Entered the value "x"	binary	Whether the student entered the value "x" into the computational input field. When "x" is entered, the computational program outputs $(3x + 2)(x - 3)$ as the result.	It was expected that students using an algebraic approach more often input the value "x" than students using a numerical trial and error approach.

Answered "$-2/3$"	binary	Whether the student submitted "$-2/3$" as answer to the item.	It was expected that students who took an algebraic approach were able to find the more difficult answer more often than students using a numerical trial and error approach.
Number of interactions	count	The total number of times the student interacted with the item.	It was expected that students who took a numerical trial and error approach interacted more with the assessment environment than students using an algebraic approach.
Longest time without interaction	continuous	The longest interval (in seconds) in which the student has not interacted with the item.	It was expected that students who took an algebraic approach spent a larger amount of time away from the assessment environment than students using a numerical trial and error approach.
Time before interacting	continuous	The interval (in seconds) between when the student was faced with the item and when the student first interacted with the item.	It was expected that students who took a numerical trial and error approach interacted with the item more quickly than students who took an algebraic approach.
Entered the value "$-2/3$"	binary	Whether the student entered the value "$-2/3$" into the computational input field.	It was expected that students using an algebraic approach input the value "$-2/3$" more often than students taking a numerical trial and error approach, as "$-2/3$" is unlikely to appear in a trial-and-improve process, and rather suggests the check of an algebraically found value.
Number of unique values entered into the computational input field	count	The number of unique values the student entered into the computational input field.	It was expected that students who took a numerical trial and error approach entered more unique values into the calculation program than students using an algebraic approach.

Number of values entered into the computational input field	count	The number of values the student entered into the computational input field.	It was expected that students who took a numerical trial and error approach entered more values into the calculation program than students using an algebraic approach.
Answered "3"	binary	Whether the student submitted "3" as answer to the item.	There was no clear expectation of a relationship between this variable and the approach a student took, as the correct answer "3" can be found quickly using either approach.
Response time	continuous	The interval (in seconds) between when the student was faced with the item and when the student moved on to the next item.	There was no clear expectation of a relationship between this variable and the approach that the student took. Both the numerical trial and error approach and the algebraic approach may take a shorter or longer amount of time.
Entered the value "3"	binary	Whether the student entered the value "3" into the computational input field.	There was no clear expectation of a relationship between this variable and the approach a student took, as this can reflect the result of a trial-and-improve process, or a conscious way to check the result of the algebraic strategy.

4.2.4 Analysis

To find meaningfully distinct groups of students based on in-assessment actions, we performed an exploratory model-based cluster analysis using the constructed variables as input variables. The analysis was performed using version 1.5-0 of the depmixS4 package (Visser & Speekenbrink, 2010) in version 4.4.1 of the R programming language (R Core Team, 2025). The models used in the analysis were finite mixture models. Binary variables were modeled using a binomial distribution with a logit link function. Count variables and continuous variables were both modeled using a Gaussian distribution with a log link function.

In Section 4.2.3, we described how theoretical expectations on student behavior informed the construction of a set of variables. In order to determine which of these variables

to include in the finite mixture models, we inspected the correlations between these constructed variables. We used a Pearson correlation for combinations of count or continuous variables, a point-biserial correlation for combinations of binary variables with count or continuous variables, and a tetrachoric correlation for combinations of binary variables. Variables that were very highly correlated ($r > 0.9$) with other variables were excluded from the model.

As a first step in determining the number of latent classes of students to fit in the final model, we compared the fit of models with one to ten classes, simultaneously assessing the robustness of the models to variations in the data by using a bootstrap procedure (e.g. Efron & Tibshirani, 1994). In each of 100 bootstraps, we sampled the total number of students from the original dataset (802) *with* replacement, creating 100 new datasets that differed from the original dataset. For each of these 100 new datasets, we fit ten finite mixture models: one for each number of classes from one to ten. For each model, the BIC (Bayesian Information Criterion (Schwarz, 1978)) model fit measure was computed. We inspected the standard deviations of the BIC values over the 100 bootstraps to assess the robustness of the model fit for variations in the student data. The means of the BIC values over the bootstraps were used to determine which number of estimated classes led to acceptable models in terms of their fit.

Next, each of these models was fitted on the original dataset 100 times using different sets of starting values, to prevent the estimation of the parameters from landing on a local maximum. The expectation-maximization algorithm (Dempster et al., 1977) that was used to estimate the parameters was allowed 100 iterations each time to converge to a solution. A minimal example of the code used for the analysis is included in the appendix (Section 4.6). To choose a final model from the at most ten models that we are left with (one for each number of classes that led to a model with acceptable model fit), descriptive statistics and visualizations of the behavior of students in the different classes on each of the variables in Section 4.2.3 as well as the students' scores were inspected to describe the type of student that was typical in each of the classes. As a rule, we opted for a model with fewer classes (a simpler model) unless adding a class would lead to a solution with more meaningfully distinguishable and describable groups of students (Spurk et al., 2020). The descriptive statistics and visualizations for the selected model are included in the results section. To ensure that the chosen model fitted the data well, we visually inspected the model parameters for the classes with the observed values for students in those classes. For the variables that were included in the selected model, the estimated model parameters are included in the visualizations to provide the reader with a practical sense of the fit of the model. To answer the second research question, two-sided t-tests with $\alpha = 0.05$ and a Bonferroni adjustment were used to determine the statistical significance of the difference in student ability between the classes of students.

Table 4.2: All fitted finite mixture models with model fit statistics.

Number of classes	Log-likelihood	Degrees of freedom	BIC	AIC
2	$-15,737$	23	$31,628$	$31,520$
3	$-15,371$	35	$30,977$	$30,813$
4	$-15,133$	47	$30,581$	$30,361$
5	$-14,911$	59	$30,216$	$29,939$
6	$-14,757$	71	$29,988$	$29,655$

4.3 Results

4.3.1 Exploring student solution strategies

4.3.1.1 Model selection

To identify the different solution strategies students used while solving the digital mathematics item, we performed an exploratory cluster analysis. We first inspected the correlations between the variables, which can be seen in figure 4.2, to determine which variables should be included in the model. Due to their very high correlations with other variables, the following four variables were excluded from the model: whether the student answered "$-2/3$", the number of unique values entered into the computational input field, the number of values entered into the computational input field, and whether the student answered "3". The remainder of the variables described in Section 4.2.3 were included in the finite mixture models.

For each of the one to ten classes of students, models were fit on 100 bootstraps of the data. The means and standard deviations of the BIC fit measures over these bootstraps are shown as an elbow plot in figure 4.3. The standard deviations of the BIC fit measures over the bootstraps were very small for each number of classes, which indicates that the models seem robust to variations in the data. Furthermore, it can be seen that a model with two classes constituted a large improvement in BIC compared to a model with one class. Each model with more than two classes further improved the BIC marginally.

To select a final model, we interpreted the behavior of students in the estimated classes through descriptive statistics and visualizations, starting with a model that estimated two classes, and finally moving up to a model with six classes. The fit measures for the different models that were fitted and interpreted are displayed in Table 4.2.

The model with two estimated classes seemed to distinguish students based on the level of engagement with the item in general. Students in one of the classes spent more time on the item and had a higher performance than students in the other class. Fitting the model with three latent classes constituted an improvement in interpretability over the model with two classes. This model was able to more clearly identify students who

Figure 4.2: Correlations between variables.

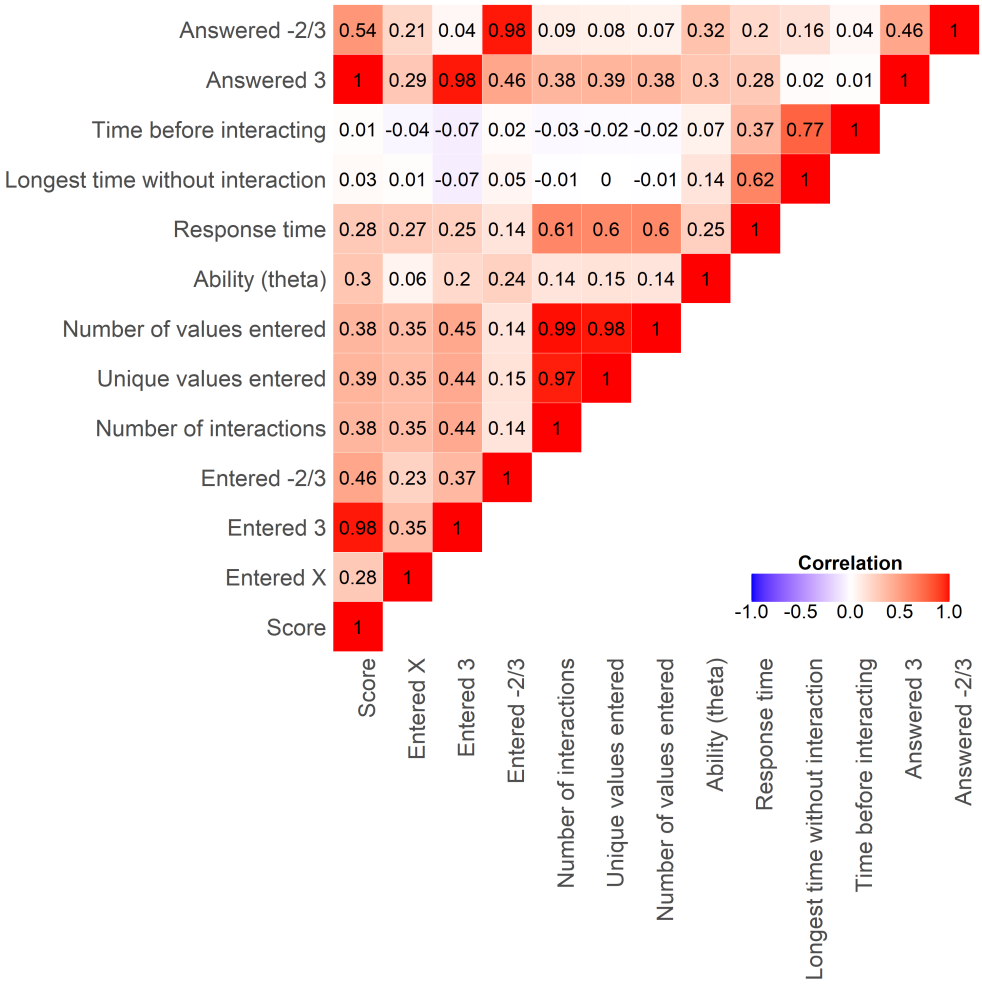
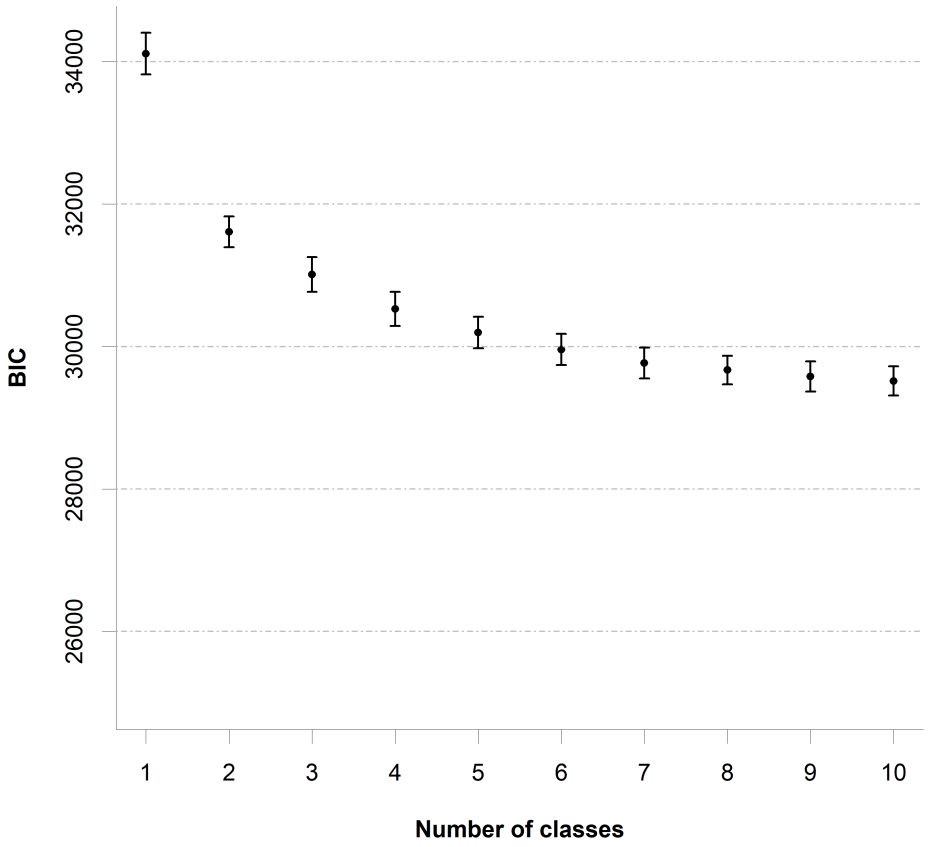


Figure 4.3: Means and standard deviations of BIC fit measures over 100 bootstraps for one to ten latent classes. Note that the y-axis does not start at zero for reasons of legibility.



engaged with the item very little and scored very badly. The other two classes were more difficult to interpret and seemed to contain students that used a mix of solution strategies, with one class spending more time on the item and the other interacting a bit more with the environment.

Estimating a model with four latent classes led to an improvement in interpretability of the classes over a model with three classes. Not only was the model again able to identify a group of students that engaged very little and were usually unsuccessful on the item, the other three classes also showed more internal consistency and interpretable behavior. Two of the classes showed behavior consistent with the expected behavior in the hypothesized algebraic and trial and error solution strategies. The third contained students that spent a reasonable amount of time on the item, but interacted little with the environment and often did not succeed. Reasons for this type of behavior could be that the student did not know how to approach the item, was distracted, or was unsuccessful in solving the item outside of the environment (perhaps algebraically). A model with five estimated classes was more interpretable still: it yielded similar classes of students that were found by the model with four classes, but with increased consistency. It also identified a class of students who spent enough time and effort on the item to enter "3" as the first correct answer but who did not put in much effort to find the second answer. The model with six estimated latent classes did not result in a more interpretable solution than the model with five classes, as it yielded two classes of students with rather similar behavior. Therefore, the final selected model was the model with five estimated latent classes.

4.3.1.2 Selected model

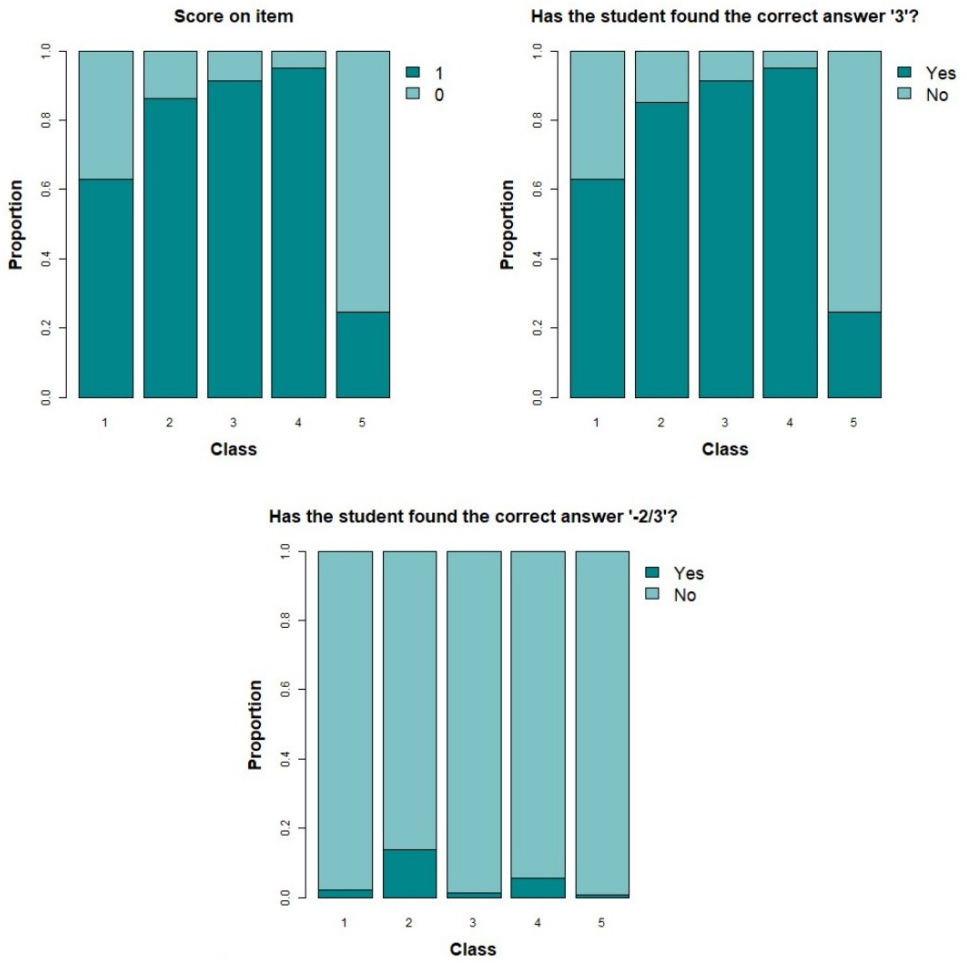
The selected model estimated five latent classes and included seven variables: whether the student entered the value "x" into the computational input field, whether the student entered the value "$-2/3$" into the computational input field, the number of interactions, the longest time without interaction, the time before interacting, the student's response time on the item, and whether the student entered the value "3" into the computational input field. The classes contained 178, 101, 228, 177, and 118 students, respectively. Descriptive statistics on all variables are listed in table 4.3 for all students in the data as well as the students in the estimated classes. The high similarity between the observed and estimated values in Figures 4.5, 4.6, and 4.7 shows that the practical fit of the selected model is adequate.

As can be seen in the top-left corner of Figure 4.4, the vast majority of students in classes two, three, and four successfully completed the item, with students in class four succeeding most often. A smaller percentage, though still a majority, of students in class one successfully completed the item, versus only about a quarter of students in class five. The chart in the top-right corner of the same figure shows a very similar pattern, because students who successfully completed the item almost always answered "3", which was

Table 4.3: Descriptive statistics on all variables for all students in the data as well as in the estimated classes.

Variable name	**Statistic**	**All students**	**Class** 1	**Class** 2	**Class** 3	**Class** 4	**Class** 5
Entered the value "x"	Frequency "Yes" (%)	74 (9.2)	2 (1.1)	18 (17.8)	7 (3.1)	46 (26.0)	1 (0.8)
Answered "$-2/3$"	Frequency "Yes" (%)	32 (4.0)	4 (2.2)	14 (13.9)	3 (1.3)	10 (5.6)	1 (0.8)
Number of interactions	Median (IQR)	12.0 (16.8)	5.0 (7.0)	21.0 (27.0)	13.0 (8.0)	26.0 (12.0)	3.0 (2.0)
Longest time without interaction	Median (IQR)	32.4 (31.1)	53.5 (29.8)	120.7 (69.5)	22.3 (8.0)	33.7 (14.0)	19.2 (16.8)
Time before interacting	Median (IQR)	24.7 (21.6)	42.6 (29.0)	64.4 (91.4)	18.6 (9.8)	27.2 (14.8)	15.5 (14.3)
Entered the value "$-2/3$"	Frequency "Yes" (%)	19 (2.4)	1 (0.6)	5 (5.0)	3 (1.3)	10 (5.6)	0 (0.0)
Number of unique values entered	Median (IQR)	8.0 (13.0)	3.0 (6.0)	14.0 (19.0)	9.0 (6.0)	18.0 (9.0)	1.0 (1.0)
Number of values entered	Median (IQR)	9.0 (16.0)	3.0 (7.0)	18.0 (25.0)	11.0 (9.0)	24.0 (11.0)	1.0 (1.0)
Answered "3"	Frequency "Yes" (%)	603 (75.2)	112 (62.9)	86 (85.1)	208 (91.2)	168 (94.9)	29 (24.6)
Response time	Median (IQR)	118.5 (100.6)	115.0 (51.2)	286.1 (136.7)	95.2 (37.2)	186.4 (76.7)	39.6 (28.6)
Entered the value "3"	Frequency "Yes" (%)	498 (62.1)	77 (43.3)	66 (65.3)	182 (79.8)	159 (89.8)	14 (11.9)
Score	Frequency succeeded (%)	604 (75.3)	112 (62.9)	87 (86.1)	208 (91.2)	168 (94.9)	29 (24.6)
Theta	Mean (SD)	0.08 (1.02)	0.09 (1.02)	0.39 (1.13)	0.19 (0.96)	0.25 (0.94)	−0.69 (0.80)

Figure 4.4: Characteristics of students in different classes with regard to score-related variables.



the easiest to find of the two possible correct answers. Finding the much more difficult correct answer "$-2/3$" was rare: only 32 (4.0%) of students did so. All students who answered "-2/3", with the exception of one student in class two, also answered "3". Class two housed the largest percentage of students who answered "$-2/3$", followed by class four, one, three, and five, respectively.

The charts in Figure 4.5 show how often specific values were typed into the input field of the calculation program (see Figure 4.1) by students in the five classes. A relatively large part of students in classes four and two entered the value "x" into the computational input field. Very few students in classes three, one, and five did so. Most of the students in class two, three, and four tried the value "3", versus a little over half of the students in class one, and much fewer still in class five. Very few students (19, or 2.4%) tried the value "$-2/3$", and most of these were assigned to class four or class two.

The graphs in Figure 4.6 show how often students in the five classes interacted with the item in the digital environment, how many unique values they tried in the input field of the calculation program, and how many interactions they had with this field in general. Students in class five interacted very little with the item, showing both the lowest median and interquartile range of all classes. Following in order of median were students in class one, three, two, and finally four. Students in class four had the highest median, and students in class two showed the most variation in number of interactions. Students with an exceptionally high amount of interactions were most often in class two, and sometimes in class four. The distributions of students for the other two variables showed a very similar pattern, which follows logically from the very high correlations between these three variables.

The distributions of time-related variables for students in the four classes can be seen in Figure 4.7. Class two had a larger interquartile range than the other classes on all three time-related variables (response time, time before interaction with the item, and longest interval with no interaction), indicating that these students portrayed more variation in time-related behavior than students in the other classes. Class two had the highest median on all time-related variables, class five the lowest, and class three the second lowest. Students in class two often had a higher response time than students in class one, but generally waited less long before interacting with the item and their longest intervals in which they did not interact with the system were generally shorter than those of students in class one.

What follows is a description of student behavior per class. Class one contained students who interacted very little with the environment, but spent a reasonable amount of time on it. A large chunk of this time was often spent without any interaction with the environment. A small majority of students in this class successfully completed the item by finding the correct answer "3", but they rarely found the correct answer "$-2/3$". Students in class two spent a lot of time on the item and spent the most time away from

Figure 4.5: Characteristics of students in different classes with regard to value-related variables.

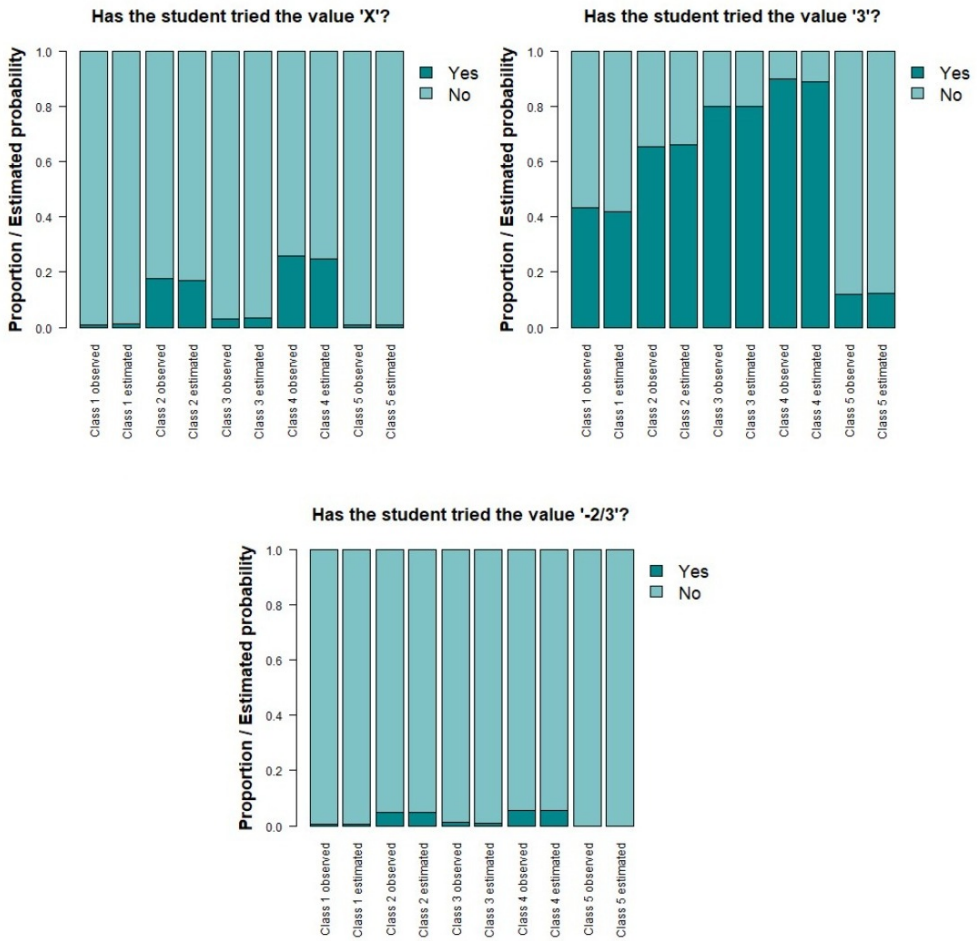


Figure 4.6: Distribution of action-related variables for students in different classes.

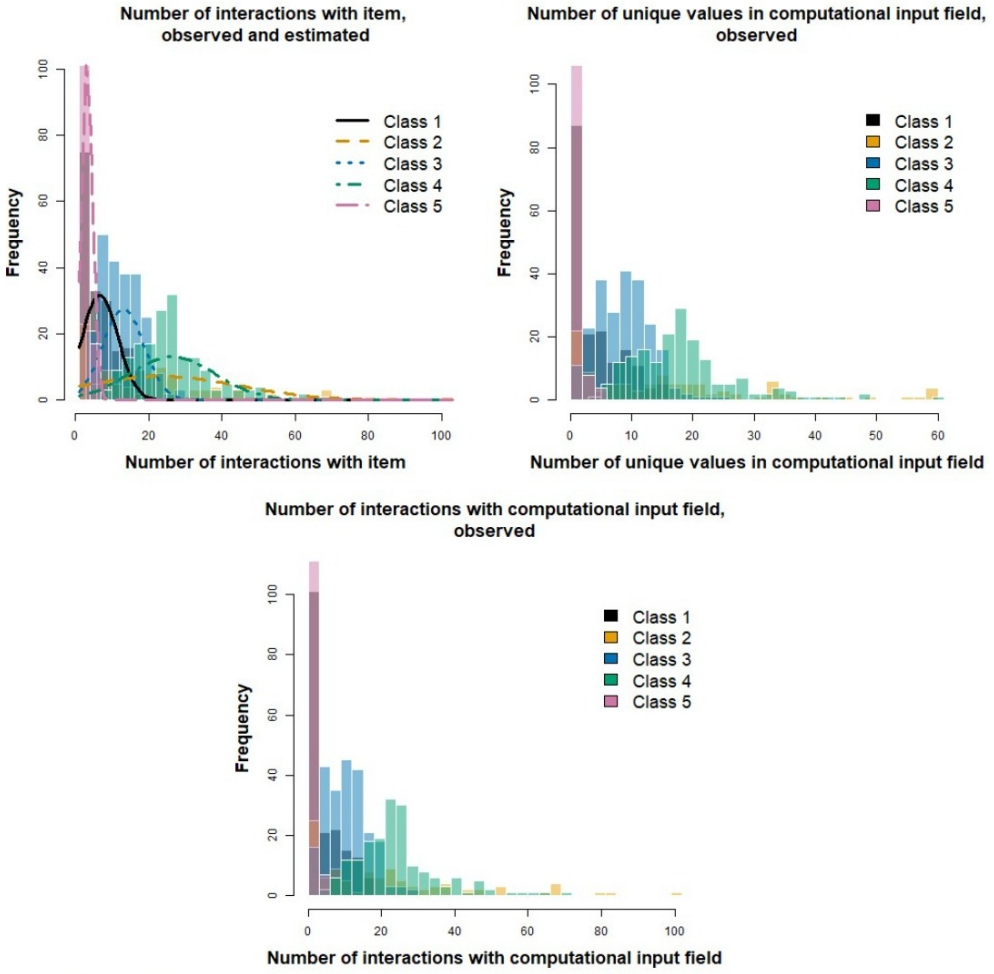


Figure 4.7: Distribution of time-related variables for students in different classes.

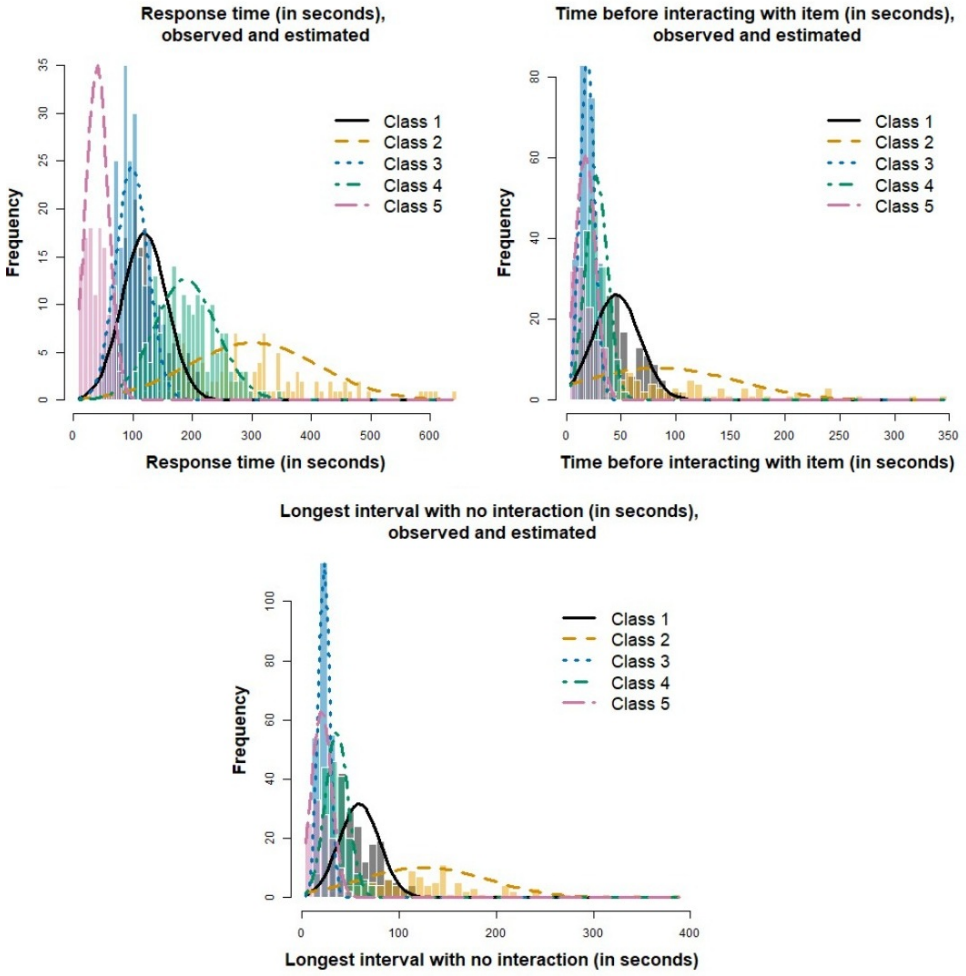
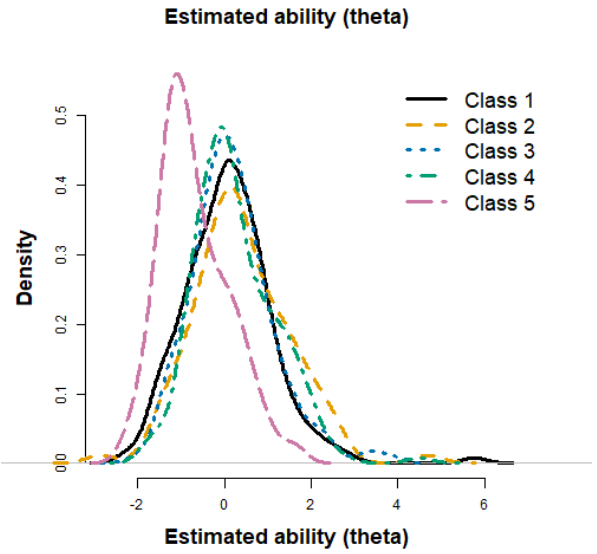


Figure 4.8: Distribution of estimated ability for students in different classes.



the digital environment. They interacted with the environment quite a bit, yet not the most of all classes. Students in class two were most often able to find the more difficult correct answer "$-2/3$". Most students in this class (although fewer than in class three and four) successfully completed the item. Students in class three spent little time away from the digital environment, started interacting with the environment very quickly, and very often successfully completed the item. However, very few of the students in this class found both correct answers to the item. They were unexceptional in the amount of (unique) interactions they performed (not very many nor very few), as well as in the time they spent in the digital environment. Class four contained the students who interacted the most with the item. Almost all of them completed the item successfully, and some of them continued to find the second answer. They spent a long time on the item, although notably, they did not spend much of that time away from the environment, nor did they wait long before interacting with the environment. Furthermore, this class contained the largest percentages of students to try the specific values of interest ("x", "3", and "$-2/3$") in the input field of the calculation program. Class five contained students who interacted little with the item, spent little time on the item, and of whom most did not succeed on the item.

4.3.2 Relationship between mathematical ability and solution strategy

The identified classes (found in Section 4.3.1) differed in their mean estimated ability, which means that more proficient students engaged with the item in a different way than less proficient students. The average estimated ability was the lowest for students in class five. The differences between the mean estimated ability of students in class five and those of the students in the other classes were statistically significant ($p < 0.001$, after Bonferroni adjustments) with large effect sizes ($d = -0.86$ as compared to class one, $d = -1.11$ as compared to class two, $d = -1.00$ as compared to class three, and $d = -1.08$ as compared to class four). The mean ability of students in class one was also significantly lower than that of students in class two ($p = 0.014$) with a small effect size ($d = -0.28$). None of the other differences in mean estimated ability were statistically significant. All means and standard deviations of the estimated abilities can be found in table 4.3, and their distributions are shown in Figure 4.8.

4.4 Discussion

With regard to the first research question, we found five distinct classes of students based on their in-assessment behavior. Class one consisted of students who spent quite some time away from the environment, which for a small majority of the students led to success on the item, but they rarely found both correct answers. These students may have tried (unsuccessfully) to solve the item using pen and paper without having seen the expression resulting from inserting "x" into the computational input field. Other reasons for their time spent away from the environment may have been that they did not know how to approach solving the item, or that they were simply distracted. We will refer to this class as *absent*. The behavior of the students in class two could be consistent with an algebraic strategy on (and a structural approach to) the item, in which the student takes time away from the digital environment to work on the problem using pen and paper. We will refer to this class as *algebraic*. The defining characteristic of students in class three is that they did not spend time and effort on finding the second correct answer to the question. They possibly did not want to try to find another correct answer, or perhaps did not realize that it was possible to find another correct answer. Their in-assessment behavior showed elements of trial and error as well as simple reasoning. We will refer to this class as *pragmatic*. Students in class four showed behavior consistent with the numerical trial and error strategy described earlier and may indicate that the students took an operational approach to solve the item. We will refer to this class as *trial and error*. Contrary to expectations, students who were in this class entered the value "x" more often into the computational input field than students in class two (algebraic approach). A possible

explanation for this is the presence of a button for "x" in the digital interface, which invites students using a trial and error approach to see what happens when it is used. Furthermore, not all students who wanted to use an algebraic approach may have realized it was possible for the program to output an algebraic expression. Students in class five engaged very little with the assessment environment. Possible reasons for this behavior could be a lack of motivation or that the item was too difficult for the student. We will refer to this class as *disengaged*.

With regard to the second research question, students in the disengaged class (five) had a significantly lower mean estimated ability than students in the other four classes. This corresponds with their behavior, in the sense that students in this class scored fewer points and interacted less with the item. It is possible that these students were generally not very motivated to spend time and effort on the assessment, and that their lack of motivation biased the estimates of their abilities. It is also possible that these students were indeed less mathematically able than the other students, in which case the item that was analyzed may have been too difficult for them. The estimated abilities of students in the algebraic class (two) and the trial and error class (four) did not differ significantly from each other. This is a surprising finding, because we would expect students who are capable of using an algebraic approach to be more advanced in their understanding, hence to have a higher estimated ability. A possible explanation for this could be that (part of) the students who did not use an algebraic approach were capable of doing so but opted for a different approach, or that some students who used an algebraic approach do not really master it yet. The small but statistically significant difference between the estimated abilities of students in the absent class (one) and the algebraic class (two) may support the earlier stated possibility that students in class one attempted to solve the item algebraically, but were not able to. Lastly, it is interesting to note that students in the pragmatic class (three) achieved good scores on the item with relatively minimal investments in terms of interactions and time. The estimated abilities of these students did not statistically differ from those of the students in the algebraic class (two) or those in the trial and error class (four), which suggests that these students may be displaying an ability to manage their time and effort effectively during assessment.

The theory-driven analysis approach in this study has advantages as well as disadvantages. An obvious advantage is that any results can more easily be placed and interpreted in context, and have greater didactic relevance. A potential pitfall is that it is possible to miss relevant patterns in the data that were not explicitly searched for. Another disadvantage is that theory-driven analyses are not easily generalized to other test items, since a (partially) new set of variables needs to be developed for the constructs that are relevant to the new item. An analysis such as the one presented in this study is informative due to the richness of the log data from such a technology-enhanced item, but it is also very time-consuming. Using sets of items with an identical structure but different values

can improve the speed and value of the analyses. A limitation of the current study is that no sequence-based variables or sequence-based methods were used in the analysis. Sequence-based analysis of student log data can be performed either in a theory-driven or data-driven manner, and both have offered promising results in the field of complex problem-solving.

A valuable direction of further research would be to validate the findings of the current study (the importance of which was highlighted by Goldhammer et al. in 2021) by finding out whether the results of the current study generalize, for example to different items, tests, contexts, and different ages of students. Various strategies could be used to contribute evidence to the validation, such as using multiple sets of student data on the same test item, using different clustering methods, or asking teachers to classify students into groups based on their perceived solution strategy. Other sources of data, such as think-aloud protocols, may also be considered. Combining different sources of data may lead to a more complete understanding of student processes. Eye tracking data has been of particular interest in this area (Maddox et al., 2018; Zhu & Feng, 2015).

4.5 Conclusion

Students take many tests and exams during their school career, but the feedback they receive about their test performance is usually based only on an analysis of the correctness of the item responses. With the increase in digital assessment, not only the item responses but also other data have become available for analysis. The time spent on solving an item is a well-known example of log data that can be used to enrich feedback, but is — when used at all — often only taken into account to improve ability estimates (e.g., Van der Linden et al., 2010). The purpose of this paper was to assess whether we can use log data in order to extend performance-related feedback with information related to the applied solution strategy. Specifically, we aimed first to make didactically meaningful inferences about students' solution strategies based on their actions in a digital mathematics assessment and second to investigate whether there is a relationship between students' mathematical ability and their solution strategy.

We have found that it is possible to distinguish several classes of student behavior that seem to correspond with the use of different solution strategies. Furthermore, we were able to identify seemingly disengaged and absent classes of students, which is a useful finding, since such students in particular may be in need of further instruction or other didactic intervention. Identifying these students, as well as groups of students who seem to adhere to a specific solution strategy may, on an individual student level, provide a basis for feedback and further instruction by their teacher. The ability to detect such strategies in real time would open the door to providing instantaneous automated feedback on this aspect during formative assessment in online learning systems. At the

level of a class or a teacher, the classification of students into these groups may be used to inform teaching. However, feedback about the application of an operational (viewing a mathematical concept as a process with input and output) versus structural (viewing a mathematical concept as an object with its own properties) solution strategy (Sfard, 1991) is only useful if the information is understood by both students and teachers and the importance of this perspective is acknowledged.

Log data can also reveal points of improvement in the design of the item or the test. Through studying the log data for the item *Product Equation*, it became apparent that very few students found both correct answers. For some students, it may have been the case that they did not realize the question had several possible correct answers. For others, the reason may have been that the item asks them "which number" (singular) as opposed to "which number(s)" (potential plural). Such design choices can influence the behavior of students and thus the log data they leave behind, and log data can in turn help identify improvements to the item design.

In order to be able to provide students and teachers with meaningful information on the applied solution strategy based on log data it is necessary to follow a holistic approach, taking into consideration various aspects related to item construction, assessment platform, data storage, and analysis. Item developers and content experts need to think about which response behavior they wish to elicit from students, and how this behavior should be recorded in the assessment environment. From the item developers, this requires a theoretical understanding of didactic and practical knowledge of the possibilities of digital items and assessment platforms. IT architects have to consider which log data should be stored and how they should be stored. It can often be very cumbersome and time-consuming to extract data from assessment platforms into formats that are feasible for analysis. Research on which log data to store in what format is important and strides are currently being made in this area (e.g., Kröhne & Goldhammer, 2018). Most important, however, is a combined and interdisciplinary effort to fully benefit from the log data and what it can reveal about student learning processes.

4.6 Appendix 1: example of R code used for analysis

```
library(depmixS4)

# nss:          number of estimated classes
# vars:         data.frame with variable information
# model_types:  list of variable distributions

bootstrapBIC <- function(nss = c(1:10),
```

```
                        nBootstrap = 100,
                        data = data,
                        samplesize = nrow(data),
                        vars = vars,
                        model_types = model_types) {

    res <- matrix(NA, nrow = nBootstrap, ncol = length(nss))
    colnames(res) <- nss

    for (i in 1:nBootstrap) {

        sampleData <- data[sample(x = nrow(data),
                                  size = samplesize,
                                  replace = TRUE), ]

        for (ns in seq_along(nss)) {

            mod <- mix(sapply(paste(vars$varname, " ~ 1"), formula),
                       data = sampleData,
                       ns = nss[ns],
                       family = model_types[match(vars$vartype,
                                                  names(model_types))])

             fm <- fit(mod)
             res[i, ns] <- BIC(fm)

        }
    }

    return(res)
}

bsmod <- bootstrapBIC()

mod <- mix(sapply(paste(vars$varname, " ~ 1"), formula),
    data = data, ns = 5,
    family = model_types[match(vars$vartype, names(model_types))])

fm <- multistart(mod, nstart = 100, initIters = 100)
```

CHAPTER 5

What do teachers think of scoring assistance? Teacher experience, speed, and accuracy in NLP-assisted scoring

This chapter has been submitted for publication.

De Schipper, E., Mulder, J., Feskens, R., & Veldkamp, B. P. (2025). *What do teachers think of scoring assistance? Teacher experience, speed, and accuracy in NLP-assisted scoring*.

Abstract

NLP-based scoring assistance has the potential to aid teachers by improving their scoring speed and accuracy and may increase measurement validity by allowing for the inclusion of more open-ended questions in tests. This study investigates teacher experience, speed and accuracy of such assistance using a prototype (*CheckMate*). The prototype's scoring interface uses a similarity-based approach in order to compute score suggestions, highlight important words, and visually cluster similar student answers together. Results of a thematic analysis indicated that teachers appreciate NLP-assisted scoring and believe using the scoring environment would save them time, but also worry about potential harmful influence that may result from displaying score suggestions. Quantitative measures derived from the prototype's log data were used to support the qualitative results. The insights from this study can be used to inform further studies and the development of applications for scoring assistance.

5.1 Introduction

Many teachers find grading a monotonous and tedious task (Kumar et al., 2017). The scoring of open-ended test questions is experienced as especially time-consuming as well as difficult to carry out in a fair and consistent manner (Brennan, 2023). Despite these obvious disadvantages of open-ended questions, there are valid reasons to include them in tests. Open-ended questions, for instance, take less time to construct than closed format questions (Bacon, 2003; Norman et al., 1991; Rademakers et al., 2005), are not subject to guessing and cueing effects (Eggen & Sanders, 1993), and fewer of them are needed to make a reliable test (Rademakers et al., 2005). Perhaps most importantly, it has been argued that closed format question types cannot assess constructs as broadly as open format question types (e.g., Claire Wyatt-Smith, 2009), and that open format question types can sometimes provide more direct measures of the construct of interest (American Educational Research Association et al., 2014).

Ultimately, in order to protect the validity of the measurement, the choice of item type should be dictated by the content and skills to be measured (Brennan, 2023, pg. 318), and not by convenience factors such as the time it takes to score the answers. As such, it is valuable to develop methods that can negate the disadvantages that come with using open-ended test questions. Both teachers and students stand to gain from the development of solutions that promote fair and consistent evaluations of open-ended questions, resulting in a more accurate appraisal of the student's knowledge and abilities and reducing the likelihood of students being disadvantaged by human errors, biases, or lapses in attention.

The ability to automatically and reliably assess student answers is an obvious way to provide potential support to teachers who administer open-ended test questions, as well as to increase scoring consistency and to cut costs (Flor & Cahill, 2025). The body of literature into the automated scoring of open-ended test questions is rich and has been growing for decades. There are various types of open-ended questions, such as fill-in and completion questions, short-answer questions, long-answer questions, and essay or argumentative questions. Scientific literature distinguishes between questions that primarily need to be judged based on writing *quality* and those that require an assessment of writing *content* (e.g. Burrows et al., 2015; Madnani et al., 2017). Evaluating writing quality focuses on form, such as spelling, grammar, sentence structure, or argumentation. Grading the content, on the other hand, involves an evaluation of what the student knows, has learned, or can do within a specific subject matter.

The automatic evaluation of open-ended question answers requires a different approach for these two different forms of open-ended answers (relevance of content or writing quality). This is largely because grading writing quality can use a generic scoring model for each language of interest, while grading content is context-specific and thus requires a more specialized scoring model (Madnani et al., 2017). This could be why

literature on automatically grading essays is relatively older and more advanced. The earliest occurrence appeared in the 1960s and concerned a program called *Project Essay Grade*, designed to score essays of students on the mainframe computers available at that time (Page, 1966). When Valenti et al. published a summary of the literature in this field in 2003, many different automatic grading systems like Intelligent Essay Assessor (IEA), and C-Rater already existed and research in this area continues to this day (Ramesh & Sanampudi, 2022).

Research into scoring assistance for the *content* of open-ended questions started around the turn of the century and has seen explosive growth in recent years. Much of this work focuses on the automatic grading of short open-ended answers (also referred to as *automatic short answer scoring*, *automatic short answer grading* (ASAG), or *automatic short answer marking*) in contrast to longer student answers such as summaries. Most studies within the ASAG field have focused on developing methods that can accurately predict the correct score for a student's answer. Burrows et al. (2015) describe a general trend in these studies moving from rule-based methods to machine learning methods. The currently dominating approach to the automated scoring of short open-ended questions uses large neural network models (Flor & Cahill, 2025). However, the performance of these scoring models is very dependent on the availability of resources available for the target language, as well as the amount and quality of scored student answers available for the model to be trained on (Flor & Cahill, 2025).

Although research on automated scoring shows great potential and utility, there are several related areas in the field of scoring assistance that still require more attention, such as the validity of automated scores. For example, Flor and Cahill (2025) indicate that automated scoring models often unintentionally use superficial signals for scoring, Bejar (2017) highlight possible vulnerabilities that scoring engines can have to construct-irrelevant response strategies, and Ding et al. (2020) show that automated scoring systems can be susceptible to awarding points to adversarial (e.g. nonsensical or unexpected) student answers.

Alternative forms of scoring assistance without automated evaluation have also been developed, with a prominent example being the grouping similar answers using clustering algorithms (e.g. Andersen et al., 2023; Weegar & Idestam-Almquist, 2024). One very relevant recent study by Weegar and Idestam-Almquist (2024) concluded that a combination of clustering and automated scoring of student answers reduced the workload of teachers by 64% to 74%. Earlier studies on the topic of clustering student answers also suggested that such a cluster-based grading method can reduce teacher workload (Basu et al., 2013), and that teachers preferred cluster-based grading (Brooks et al., 2014).

Since its introduction to a more general audience, a new category of scientific work which uses generative AI in order to i.e. score student answers or to give feedback to students using prompts has arisen (e.g. Henkel et al., 2024; Korthals et al., 2025),

with promising results. The popularization of generative AI has also caused an influx of commercial scoring tools, sometimes stand-alone and sometimes integrated with existing testing or learning platforms.

Some existing user-based research suggests that these tools can add practical value, for example by making it possible to scale up the number of students without needing more teachers, shortening the amount of time between assignment submission and receipt of a grade and feedback, and reducing subjectivity in grading (Hahn et al., 2021). However, there are also significant drawbacks, such as the discouragement of creative and innovative student answers and the possibility that it may incentivize students to learn the answer rather than the subject matter. In a few other related user-based studies, methods were explored to make the use of AI in automatically scoring short answers comprehensible (Schlippe et al., 2023) and to research the impact of automatically scoring short answers on the student's learning process and the relationship between the teacher and the student (Siddiqi, 2013).

However, there is as of yet little empirical research into teacher satisfaction and interaction with systems that provide scoring assistance (Del Gobbo et al., 2023), also owing to the fact that the implementation of such systems is rather recent (Ouahrani & Bennouar, 2024). Although scoring assistance has widely acknowledged potential, it is vital that more research is done into the teacher experience of using such systems, since there are also possible pitfalls that should be avoided in their design. For example, some forms of scoring assistance might overlook a teacher's need for agency and the potential benefits of participating in the scoring process, such as gaining insight into students' knowledge gaps.

This study offers a detailed investigation of teacher experiences with a prototype of a system designed to help them score open-ended test questions in Dutch secondary education. Along with a user-friendly interface for horizontal and anonymous scoring, the prototype contains three added NLP-based functionalities for scoring assistance: score suggestions for unscored answers, the clustering of similar student answers, and the highlighting of important words in the scoring model. The analyses in this study are predominantly qualitative and will be supported with quantitative measures when possible.

We aim to study the following research questions:

1. How do teachers experience NLP-assisted scoring?
2. How does NLP assistance impact teachers' scoring accuracy?
3. How does NLP assistance impact teachers' scoring speed?

5.2 Methodology

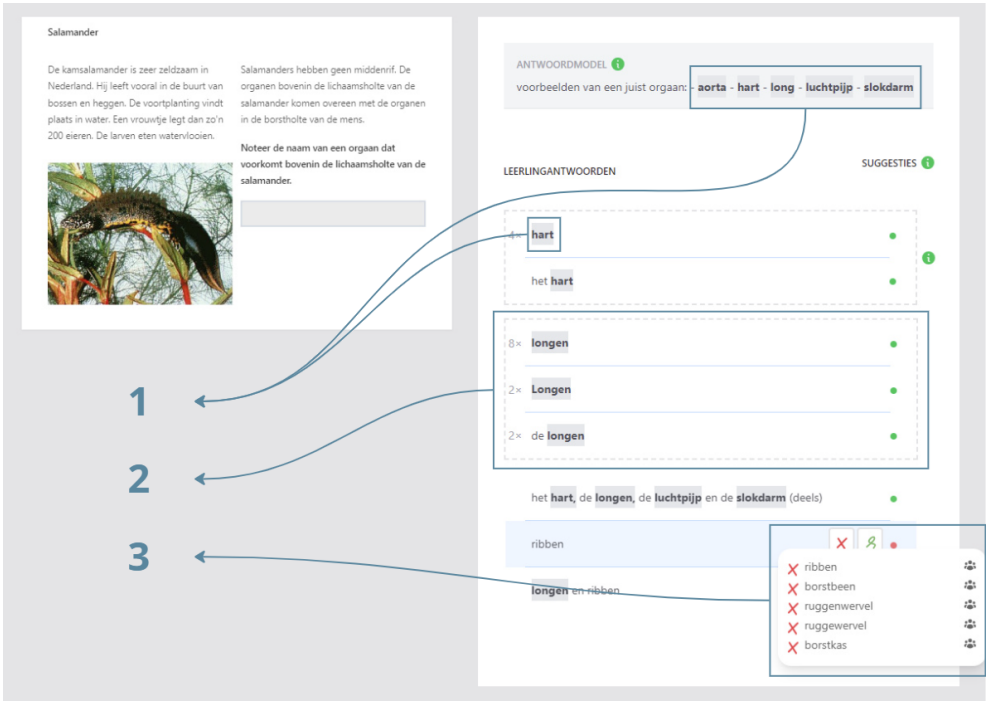
We study the aforementioned research questions with the use of a prototype of a web application for NLP-assisted scoring, named CheckMate. CheckMate was designed and developed by CitoLab (the research and innovation department of Cito Foundation in the Netherlands), in hopes of developing and researching methods to support teachers in the scoring of open-ended test questions. The main goals of the prototype were to reduce the time teachers spend on scoring open-ended test questions, as well as to increase the consistency of their scoring. The developed prototype was then used to answer the posed research questions using both qualitative methods (semi-structured interviews analyzed using thematic analysis) and quantitative methods (log data analysis of teachers' actions in the scoring environment).

Before development, preliminary research was conducted among teachers in secondary education. This was done to gauge whether there was a need for the development of a prototype in this field, and if so, to collect the requirements for such a solution. The preliminary research consisted of interviews with two teachers as well as a short national survey of teachers in the Netherlands via the TeacherTapp app. The interviews included a wide range of questions concerning the teachers' practices around testing and scoring, and were purposefully conducted in a flexible, conversational manner to ensure that topics that were of importance to the teachers got the attention that they deserved. In the national survey, three questions were asked about their perceived burden of scoring tests in the classroom, what types of tests they generally use and what scoring methods they employ. The translated questions and survey results are included in the appendix.

Through the interviews and the national survey, we uncovered that there was a lot of need among teachers: almost all teachers experience a very high workload, and scoring takes them more time than they would like. It was determined that a scoring solution should have a low threshold for use, that the solution should learn from the input that the teacher gives it, and that ownership of the given scores should remain with the teacher. It was decided to develop a prototype that focuses on short open-ended questions consisting of written text ranging from one word to a few sentences.

CheckMate consisted of a scoring interface and a simple reporting interface. The scoring interface is displayed and explained in Figure 5.1, and used techniques from the field of natural language processing (NLP) in order to compute semantic similarity between different student answers and the scoring model. These similarities were used to compute score suggestions, highlight important words, and to visually cluster similar student answers together. The prototype also included an alternative interface for scoring without these added NLP-based functionalities. This alternative interface was used as a digital control condition. The current study will be used to inform further studies and development of the prototype.

Figure 5.1: CheckMate's scoring environment with added functionalities using techniques from NLP: 1) highlighting of important words, 2) clustering of similar student answers, and 3) score suggestions with explanations.



5.2.1 Development of CheckMate

During the course of two three-week periods, many versions of a prototype were designed, built and user tested for intuitiveness, relevance, and usability. Tests of the intuitiveness of the environment were conducted with five non-teacher participants, and more elaborate user tests and interviews for testing relevance and usability of the solution were undertaken with five teachers in secondary education. As a result of the user tests and interviews, functionalities were improved upon or removed altogether. What remained after this iterative process was a version of the prototype that had three main functionalities: clustering semantically similar student answers, giving score suggestions for answers that have not yet been scored, and highlighting important words in the student answers and scoring models. A similarity-based approach (e.g. Bexte et al., 2023; Horbach & Zesch, 2019) was chosen because it can give the teachers insight into how the score suggestions were generated (Suen et al., 2023), and because this method does not need an item-specific model trained on many human-annotated training examples and can thus be more readily used in a classroom context (Attali et al., 2022).

5.2.1.1 Text similarity

At the heart of the clustering and score suggestion functionalities in CheckMate is semantic similarity between each of the known student answers in the system, and between the student answers and the defined scoring model. In order for these similarities to be estimated, the texts should be cleaned (*preprocessed*) and a numeric representation of their meaning should be computed. All of the necessary steps were performed in Python using the *spaCy* library (Honnibal et al., 2020). First, the student answers are split into shorter segments which are often words (*tokenizing*), and reduced to the simplest form of the word (*normalizing* and *lemmatizing*). Then stop words and punctuation are removed and only verbs and (proper) nouns are retained, since we considered them to contain the important information for assessing the correctness of a student answer. For example, a cleaned version of the sentence "We are going to the mall." could be {we, be, go, mall}.

The cleaned student answers and scoring models now consist of a set of tokens that can be looked up in spaCy's pretrained open source language model *nl_core_news_lg*. It is assumed that words that often coexist in the same contexts have a similar meaning or connotation. This is known as the *distributional hypothesis* (Flor & Hao, 2021), and is based on the concept of a distributional structure of language, introduced by Harris (1954). The language model contains a dictionary of words in the Dutch language along with a vector of 300 numerical values for each of the words (*word embeddings*). These vectors represent the meaning of the word in a 300-dimensional space, and words that are near each other in this space coexist more often in Dutch news media than words that are numerically further away from each other. An approximation of the meaning of a student

answer or scoring model is calculated by taking the elementwise average of the vectors for the tokens in the student answer.

To compute the similarities between two student answers, we use the cosine similarity measure, which is the cosine of the angle between the two vectors. Cosine similarity values range between -1 and 1, where semantic vectors with opposite meanings (such as the vectors for antonyms such as 'hot' and 'cold') have values close to -1, and semantic vectors with very similar meanings (such as the vectors for 'good' and 'great') have values close to 1. Vectors for words or phrases that are completely unrelated (e.g. 'frog' and 'coffee') have values close to 0. Similar student answers are expected to have similar semantic vectors, leading them to have high cosine similarity values. This approach to computing similarity between texts has been applied by several other studies in the field of automated scoring (e.g. Putnikovic & Jovanovic, 2023). For student answers u and u', the cosine similarity is:

$$sim(u, u') = cos(\theta) = \frac{\mathbf{r}_u \cdot \mathbf{r}_{u'}}{||\mathbf{r}_u|| \cdot ||\mathbf{r}_{u'}||} = \sum_i \frac{r_{ui} \cdot r_{u'i}}{\sqrt{\sum_i r_{ui}^2} \cdot \sqrt{\sum_i r_{u'i}^2}}, \tag{5.1}$$

where r_{ui} is the score for student answer u on language dimension i.

5.2.1.2 Clustering of student answers

It is a reasonable assumption that semantically identical student answers should be scored identically. In CheckMate, answers that were similar to each other were clustered together into groups (see Figure 5.1). We hypothesized that this functionality might save teachers time and cognitive load by not having to consider similar answers for evaluation several times, and that it could contribute to scoring objectivity by decreasing the chance of inconsistencies in scoring. This hypothesis was strengthened by the decrease in time spent on scoring that was found by both Weegar and Idestam-Almquist (2024) and Basu et al. (2013) after the implementation of student answer clustering. Furthermore, a follow-up study by Brooks et al. (2014) concluded that teachers preferred a cluster-based scoring method.

Student answers were clustered using the DBSCAN algorithm (Ester et al., 1996) in the DBSCAN Python library. This algorithm requires the user to specify two parameters in advance: the maximum distance allowed between answers before they are no longer considered close enough to each other to be in a cluster together (*epsilon*), and the minimum number of similar answers required to form a cluster. After preliminary user tests and some empirical tests on existing data (scored student answers from when the test questions in this study were administered in earlier years), the value of epsilon was set to 0.5 and the minimum number of answers in a cluster was set to 2. The same parameters were used for all items.

5.2.1.3 Estimation of score suggestions

Another functionality that was integrated into CheckMate was the display of score suggestions for unscored answers (see Figure 5.1). These suggestions gave teachers an indication of whether the algorithm estimated a student answer to be correct or incorrect, as well as the certainty of the estimation. The suggestions were based on the semantic similarity of the student's answer to other previously scored answers, which could be viewed by clicking the score suggestion. We hypothesized that displaying these score suggestions would make scoring faster and more consistent, validating a teacher's judgment if in agreement with the score suggestion, and triggering reconsideration of the score if not.

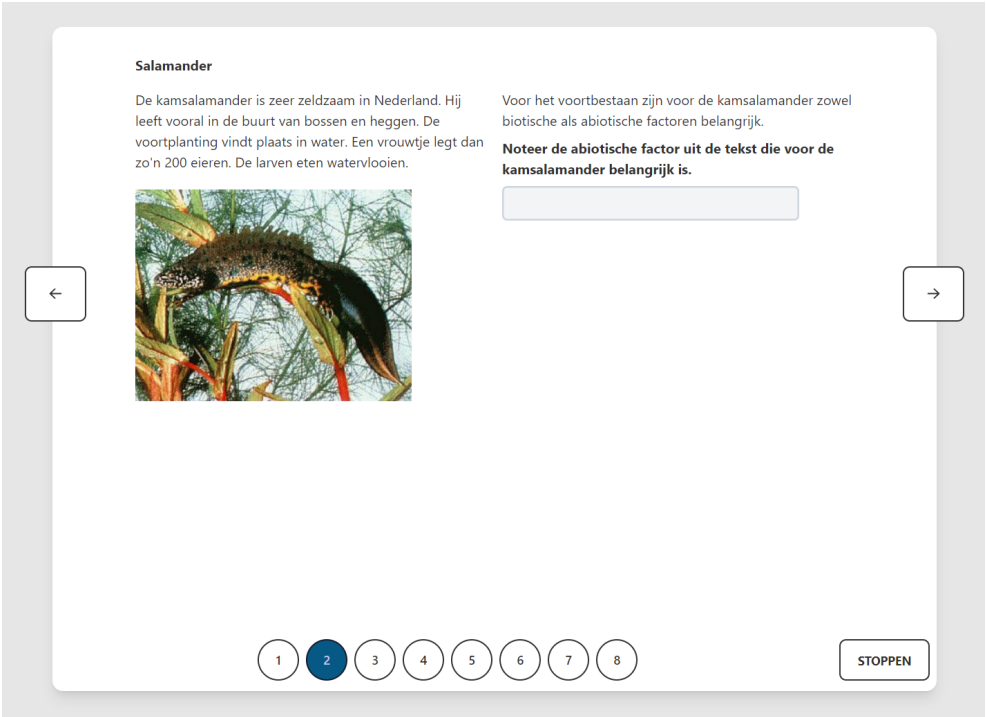
In order to estimate a score suggestion for a student answer, the algorithm used the five most similar student answers that were scored by other teachers, and the five most similar student answers scored by the teacher that is currently scoring (but only if the similarity exceeded 0.5). A score suggestion was computed as the sum of the similarity scores multiplied by the score (1 for correct, -1 for incorrect), divided by the number of answers that were used to compute the score. The final score suggestion was the mean of the score suggestion based on answers scored by other teachers and the score suggestion based on the current teacher. The teacher's own scoring input was taken into account as much as the input of all the other teachers combined in order to make sure that the teacher experienced personal influence during scoring. If the estimated score was lower than -0.1, we suggested to score the answer as incorrect (red indication). If the estimated score was higher than 0.1, we suggested to score the answer as correct (green indication). The certainty of the score suggestion was determined by the deviation from 0 and visualized through the darkness of the color in the interface (darker being more certain). Estimated scores between -0.1 and 0.1 were not displayed in the interface, due to very low certainty. Although it was not necessarily the purpose of this study and this version of the prototype to optimize for prediction performance, the algorithm was improved upon enough to plausibly ensure face validity for the teachers.

5.2.1.4 Highlighting of important words

Lastly, CheckMate highlighted words that were of particular importance (according to the authors) in the scoring models and student answers (see Figure 5.1). We hypothesized that these highlights would make it easier to scan the student answers for correctness. Del Gobbo et al. (2023) also suggested implementing such a functionality for a different reason: to provide explanations to teachers as to how answers are scored.

In order to highlight important words in the scoring model as well as words that are related to them, we used the same Dutch language model to find the 200 closest word tokens in the 300-dimensional space for each important word in the scoring model. These 200 tokens were highlighted if and when they occurred in the students' answers.

Figure 5.2: The online student test environment.



5.2.2 Data collection

Seven biology teachers in the upper grades of pre-vocational secondary education in the Netherlands participated in scoring sessions and interviews for this study. Four of these teachers along with their own classes of students (91 students in total) were randomly assigned to the experimental condition, the other three teachers with their combined 43 students were assigned to the control condition.

For each of the teachers participating in the study, a class of their students took a biology test in the prototype CheckMate consisting of eight open-ended test questions. The length of the answer required for the questions varied between one word and a couple of sentences. The students logged into an online test environment on their own device, in their preferred browser using an identification known to the teacher but not to the researchers as well as a four-letter group code provided by their teacher. A screenshot of the online environment is shown in Figure 5.2. The students were told that they were about to take a biology test and that taking it seriously would serve as good practice for their exams. They were also informed that their answers and scores would be anonymous, and that there would be no personal consequences based on their performance.

Afterwards, on a day and time convenient for the teachers, the scoring session and interview took place. Before the scoring session, each teacher was first shown an instructional video on how to use the scoring interface of CheckMate. The teacher was then asked to score student answers in a demo version, in order to ensure good understanding of the interface and its functionalities. When they felt ready to start, the teacher logged into the scoring environment with their personal group code. Their interactions were recorded using screen captures, and all mouse clicks were logged with their accompanying timestamps. In order to preserve the accuracy of the logged time measurements, the teachers were asked to score the students' answers without discussing their choices with the observer. All teachers participated on a laptop. Some used a mouse and others a trackpad, depending on what they usually used.

Afterwards, semi-structured interviews were conducted with all participating teachers. The semi-structured nature of the interviews allows the interviewer to ask follow-up questions, which offers the opportunity to get a closer understanding of the interviewee's experience (Bloor & Wood, 2006). In the interviews, the teachers were mainly asked about their general impression of the prototypes, as well as their experiences with the three included functionalities (clustering, score suggestions and highlighting). Additionally, they were asked to compare their experience of scoring in CheckMate with their experience of the method or system with which they usually score tests, and whether they had unanswered needs or suggestions for improvement. A translated version of the 20 interview questions that were formulated ahead of time are included in the appendix (Section 5.7).

The interviews were conducted either digitally via video call or live at the teacher's school, and ranged in duration between 20 and 45 minutes. Sound and/or video recordings were made during all the user tests and interviews. All teachers had given explicit consent for the making of such recordings, and for their use in this study. After the interview, teachers that were assigned to the control condition were shown the scoring environment for the experimental condition, and vice versa. They had the opportunity to try the environment as much as they wished, but were not required to do so. Subsequently, they were asked hypothetical questions about how they thought they would experience the other scoring environment (with or without the added NLP-based functionalities).

In order to determine any differences in scoring accuracy between the teachers in either condition, a set of nine representative student answers from students in earlier exam years (*anchor answers*) was included per test question in each scoring session. For these student answers it was known from prior consensus between experts what their true score ought to be. These anchor answers were added to the participating teachers' scoring environments in the prototype in such a way that the teachers did not know which answers belonged to their own students and which were added by the researchers. In total, each teacher scored an extra 72 student answers on top of their own students' answers (nine answers

for each of the eight test questions).

5.2.3 Data Cleaning

The timestamp of the first scoring action (clicking *correct* or *incorrect*) was considered the start of scoring. Time spent on activities other than scoring after this point in time (such as on asking a question to the observing researcher) was estimated using the screen captures and audio recordings and subtracted from scoring time. Time that was spent on reading the test questions, the answer model or the student answers was not filtered from the data, as it was not possible to estimate these durations accurately. It is assumed that this time is approximately equal between the two experimental conditions. The scoring durations ranged between 0.003 seconds and 43.111 seconds, except one duration of 70.07 seconds which we considered an outlier and thus removed. After removal of this outlier, 504 scoring duration values remained for the anchor answers.

5.2.4 Analysis

In order to explore how teachers experienced scoring using the developed prototype, the conducted interviews were transcribed and subsequently coded and analyzed using Atlas.ti version 24.1.0. (ATLAS.ti Scientific Software Development GmbH, 2024). Through open, axial, and selective coding, various overarching themes and relevant concepts were identified, which served as the basis for the coding protocol as shown in the appendix (Section 5.8). The coding protocol was developed by one coder, and discussed with and approved by a second coder. Patterns were derived from the interview data using thematic analysis (Terry et al., 2017). Both inductive and deductive approaches were used: some themes derived from interview questions, others from teachers' contributions. All interviews were coded by these same two coders independently, after which the differing codes were discussed until full coding agreement was reached. Co-occurrence analyses were used to gain insights into what themes were discussed by the teachers when they spoke about the three added functionalities in CheckMate.

Data on scoring duration and scoring accuracy were derived from the teachers' actions through the use of log data. Only the scoring actions that teachers took on anchor answers were used for the analyses on these aspects, in order to increase comparability between teachers. The means and standard deviations of scoring durations (in seconds) were compared descriptively between the two conditions. For each condition, we computed the accuracy of the scores that the participating teachers awarded to the anchor answers (as compared to the true scores), and compared these between the control and experimental conditions. We also computed the accuracy of the predicted scores by the algorithm.

Through the qualitative results of the in-depth interviews and the use of supporting quantitative measures derived from log data, we study the teachers' experiences with

CheckMate and its added NLP-based functionalities, as well as the teachers' speed and accuracy of scoring.

5.3 Results

5.3.1 Teacher Experience

Thematic analysis was used to explore how teachers experienced scoring using the developed prototype. In the end, 53 unique categorical codes in 10 different themes were applied a total of 771 times to interview segments. The ten themes that were found were "advantages", "disadvantages", "CheckMate's functionalities", "preconditions", "daily teaching practice", "trust and distrust", "autonomy", "suggestions", "doubt", and "personal beliefs". We explored these themes and their codes, as well as some of their co-occurrence counts. The occurrence and co-occurrence counts of codes discussed in this section are shown in Table 5.1. For brevity's sake, only codes that are relevant to the research questions are discussed. All quotes have been translated from Dutch to English.

In general, the teachers in both the control and experimental conditions had a positive impression of scoring with the prototype. All teachers indicated that they would find the prototype useful in their scoring practice.

In the seven interviews, there were 125 mentions of advantages to scoring using the prototype's scoring environment. The teachers mentioned advantages in combination with the three extra functionalities in the CheckMate experimental condition 60 times. 32 of these concerned the clustering of student answers, 23 concerned the score suggestions and five concerned the highlighted words. In relation to the clustering of student answers, teachers most often mentioned saving time (*"I appreciate the clustering of student answers, because you get rid of a large bulk of the answers in one go. I'd say around 50% to 60%."*), ease of use (*"They are already grouped together, so that makes it easier to score them."*) and consistency (*"I have previously scored these answers as correct, so now I have to pay attention to do the same to those."*). The most often mentioned advantage in combination with the score suggestions was ease of use, often in relation to being able to see scores that other teachers gave to related answers (*"[...] convenient to see the colored dot, and what other teachers have scored."*). The most often mentioned advantage in combination with the highlighted words was the clarity of the interface (*"They help me by steering the movement of my eyes on the screen."*). The other 65 mentions of advantages of scoring with the prototype were not in combination with specific functionalities, indicating that the scoring interface without added functionalities also held important advantages to the teachers. Teachers most appreciated the interface's ease of use (16 mentions) and clarity (15 mentions), and often mentioned that scoring using the environment would save them time (10 mentions). As most teachers used paper-based testing and scoring in their daily

Table 5.1: Occurrence of codes within themes, and co-occurrence with the three added functionalities in the CheckMate experimental condition.

Type	Description	Clustering	Highlighting	Score suggestions	Other
advantage	clear interface	1	3	2	15
advantage	confirmation of judgment	0	0	7	0
advantage	consistency	6	0	3	0
advantage	ease of use	10	0	8	16
advantage	legibility	0	1	0	4
advantage	objectivity	0	0	1	7
advantage	saving time	12	0	2	10
advantage	speed	3	1	0	7
advantage	student experience	0	0	0	6
disadvantage	confusion	0	2	0	1
disadvantage	dyslexia	0	0	1	0
disadvantage	harmful influence	0	3	10	0
disadvantage	passive attitude	0	2	5	0
disadvantage	subjectivity	0	0	2	0
disadvantage	unclear interface	0	1	0	1
disadvantage	use of color	0	0	1	1
trust	algorithm	2	0	2	1
trust	own judgment	0	0	5	2
trust	teachers in other schools	0	0	7	2
trust	teachers in own school	0	0	5	2
distrust	algorithm	0	1	3	1
distrust	computers	0	0	0	1
distrust	content	0	0	1	4
distrust	own judgment	0	0	0	1
distrust	teachers in other schools	0	1	3	4
autonomy	control	0	0	5	5
autonomy	decision-making authority	0	0	1	3
autonomy	independence	0	0	8	1
autonomy	influence of others	0	0	13	1
autonomy	responsibility	1	0	1	5
autonomy	seeking knowledge	0	0	2	1
doubt	'gray area' answers	1	0	9	6
doubt	unsure	0	0	8	0

practice, using a digital scoring interface saved them time in terms of leafing through student tests (two mentions) and deciphering student handwriting (two mentions).

Disadvantages to scoring with the environment were mentioned 30 times in total, out of which 27 concerned the three added functionalities. None of the mentioned disadvantages concerned the clustering of student answers, eight concerned the highlighting of words and the remaining 19 concerned the score suggestions. For both the highlighting of words and the score suggestions, the most mentioned disadvantage was harmful influence (*"You quickly get tempted to let yourself be persuaded by the colors."*). Another commonly mentioned disadvantage of the score suggestions was that they might lead to a more passive attitude in the teachers (*"[...] it could trigger me to become less exact when I experience high workload."*). The other three mentions of disadvantages of scoring with the prototype were not in combination with specific functionalities.

The participating teachers were also asked about their usual testing and scoring practices. All teachers indicated that they often or always used paper-based testing and scoring, with some teachers using digital assessment in specific contexts. They used a combination of horizontal (per question) and vertical (per student) scoring, preferring horizontal scoring for open-ended and more difficult questions.

In a large majority of the cases in which teachers spoke about trust or distrust, they referred to trust in their own judgments or in the judgment of other teachers. It was much less common for them to speak of trusting or distrusting the scoring algorithm, even when explicitly prompted by the interviewer. These themes were most often discussed in tandem with the score suggestions functionality. Stated reasons for distrust in other teachers were scoring biases that may arise from a system in which the evaluation of teachers is partially dependent on their students' performance, and a belief that some teachers do not pay enough attention while scoring their students' tests. The qualifications that are needed to become a teacher were a stated reason for trust in other teachers. The teachers spoke about distrusting the algorithm when they were certain of their own judgment and the score suggestion indicated the opposite score, whereas in most other cases they displayed a default level of trust in the algorithm on the basis of expertise (*"[...] it has been researched."*).

Topics within the theme of autonomy were mentioned 30 times when discussing the score suggestions functionality, one time when discussing the clustering functionality, and 16 times outside of discussions on the three functionalities. The teachers valued the opportunity to look at the input of other teachers greatly, which is illustrated by the 13 mentions that were related to the influence of others when talking about score suggestions (*"What I really like is that I can see what other teachers would do."*). On the other hand, they also spoke often about the importance of maintaining independence (*"Sometimes I thought: hey, it shows a green dot, but no, this really is incorrect in my opinion, or too vague of an answer."*) and exerting control (*"I am not going to look at it, and will stick*

to my own opinion."). A similar pattern existed around the theme of doubt. This theme was also most often discussed when talking about the score suggestions (17 mentions), one time when talking about the clustering functionality, and six times outside of the three functionalities. Teachers were especially interested in the score suggestions and the scores that other teachers gave when coming across 'gray area' student answers: student answers that they were in doubt about and might usually discuss with a colleague. For example, one teacher said *"Sometimes you are in doubt about a student answer, and then you can look at what others did with such an answer. [...] Then you can learn from each other, and take the knowledge into account in order to score more objectively."*. On the other hand the score suggestions functionality was also the only context in which teachers expressed being unsure about their own opinion (*"Now that I think about it, I'm not sure whether I would like to see the dots."*).

Some of the participating teachers spoke of necessary preconditions for implementing a system such as the developed prototype. These centered around IT, content and logistics (two, four, and two mentions, respectively). Mentioned were the need for students to be digitally skilled, an intuitive interface, access to computers at school at the right moments, a proctoring functionality, being able to easily upload their own tests, as well as a functionality to give their students feedback. The latter precondition was also often mentioned as a suggestion for further development (seven mentions). 19 suggestions concerned changes to the user interface, 18 concerned additional functionalities, and 14 concerned various other topics (such as partial scoring, or ignoring capitalization when stacking identical student answers).

5.3.2 Scoring accuracy

The overall accuracy of the shown predicted scores was 0.82 ($n = 262$). The overall accuracy of the scores given by teachers was 0.91 ($n = 207$) in the control condition and 0.85 ($n = 299$) in the experimental condition. The accuracy per test question can be seen in Table 5.2. They were consistently high ($accuracy >= 0.85$) for teachers in the control condition. For many test questions, the accuracy of the scores was similar between the experimental conditions. The notable exception is question number seven, for which the accuracy of teachers in the experimental condition was considerably lower ($accuracy = 0.56$). The accuracy of the predicted scores for students answers on this test question was even lower ($accuracy = 0.38$), owing to the longer length and more variable phrasing of its longer student answers and scoring model, for which the algorithm used is generally less accurate.

Table 5.2: Accuracy of scores given by teachers in the control condition ($acc_{control}$), accuracy of scores given by teachers in the experimental condition ($acc_{experimental}$), and accuracy of predicted scores ($acc_{predicted}$), along with their corresponding numbers of student answers (n).

Test question	$acc_{control}$	n	$acc_{experimental}$	n	$acc_{predicted}$	n
#1	1.00	18	0.85	46	0.93	46
#2	0.88	26	0.89	38	0.80	30
#3	0.89	27	0.89	37	0.71	31
#4	1.00	27	1.00	36	1.00	35
#5	0.93	27	0.89	37	0.86	29
#6	0.87	31	0.89	36	0.81	36
#7	0.85	27	0.56	36	0.38	29
#8	0.88	24	0.79	33	0.96	26

5.3.3 Scoring Duration

Teachers in the control condition spent an average of 4.4 seconds on scoring one of the anchor answers ($SD = 8.0$, $n = 206$). Teachers in the experimental condition spent an average of 4.1 seconds per anchor answer ($SD = 4.4$, $n = 298$).

5.4 Discussion

This study offered insights into the behavior and wishes of biology teachers in secondary education with regards to scoring assistance. The quantitative results in this study should be interpreted with care due to the small-scale nature and between-teacher design of this study. They are used as supportive or countering evidence for the qualitative results. Furthermore, because we tested the interface of a scoring prototype as a whole, it is impossible for us to draw conclusions about the effects of specific functionalities on the speed and accuracy of scoring.

The results of the qualitative analysis suggest that teachers can save considerable time compared to their usual scoring method by using the provided scoring environment in both the control and experimental condition. Teachers were generally positive about the added functionalities in the experimental condition and believed these would save them time. The quantitative results suggest that there is little difference in scoring speed between the two conditions. There was, however, more variance in scoring speed in the control condition. This may indicate that teachers who are uncertain about how to score an answer might work faster with NLP-assisted scoring.

The clustering of similar student answers was unanimously positively received. The teachers felt that it would save them time, that it made the environment easy to use, and that it would increase the consistency of their scoring. They spoke very little about the

highlighting of words in the scoring model and student answers, even when prompted, and the ones that did so had differing opinions about the functionality. For example, some teachers felt that the highlights made the interface confusing, others felt it improved the clarity of the interface. It is possible that this functionality would have been more useful if the important words had been determined by a content expert as opposed to the authors of this study, or if a highlighting functionality had been implemented differently. The teachers had mixed feelings with regards to the implemented score suggestions. On the one hand, they appreciated getting access to information about what score other teachers might have given to a similar student answer, and felt that they made the interface easy to use. On the other hand, most teachers also had scruples concerning harmful influence of the scoring algorithm or changes in teachers' scoring attitudes. The quantitative data also suggest that teachers were influenced by the score suggestions. Table 5.2 shows that the scoring accuracy of teachers in the experimental condition on test question number seven was considerably lower than the scoring accuracy of teachers in the control condition on that same question. Since this is the question for which the scoring algorithm offered very inaccurate predicted scores, it is plausible that the teachers in this condition were negatively influenced by being exposed to these scores. This should serve as confirmation to parties developing automated scoring algorithms that the scoring accuracy of automated scoring solutions truly matters.

The results of the co-occurrence analysis showed that teachers often mentioned topics in the theme of autonomy when they spoke about the score suggestions, unlike when they spoke about the other implemented functionalities. They spoke most often about the (mostly positive) influence that the opinions of other teachers could have on the outcome of their scoring, and about exerting independence and control during scoring. The co-occurrence analysis also showed that teachers often spoke about the score suggestions in combination with 'gray area' student answers: answers for which it was not very clear to them which score should be given. A final theme that was often touched upon when speaking about the score suggestions was the theme of trust and distrust. Generally, teachers spoke of trusting the judgment of other teachers (both within and outside of their own school), as well as their own judgment. In some cases, but much less often and only when prompted, they spoke of trusting or distrusting the scoring algorithm. Some teachers spoke of distrusting the judgment of teachers from other schools.

The findings show us that the need for autonomy is strongly contextual and dynamic. At certain moments, such as when they were unsure what score to give to a student answer, teachers allowed themselves to be influenced by the score suggestions. At other moments, the teachers did not wish to be influenced, for example when they were strongly convinced of their own judgment or when they had given specific instructions to their own students. Clearly the teachers were often consciously influenced during scoring, but this process may also happen unconsciously — for better or worse, depending on the quality

of the offered scoring assistance. In any case, it is clear that there is a need for teachers to retain the possibility to take their own decisions around the scores of their students.

The interviews in this study also contained other valuable information outside of the scope of the current research questions. For example, we learned that the interface was not suitable for persons with color blindness in its current form, and that there is a need for good functionalities around giving feedback to students and reporting student results.

The most prominent limitations of this study are its small number of participants and between-teacher design, leading to stringent restrictions in drawing conclusions, especially in terms of the quantitative aspects of this study. Although the results of this study can be used to indicate further directions for research and iterative development, a large-scale study is needed to properly evaluate teacher speed and accuracy in such a system for NLP-assisted scoring. A within-teacher design would enable us to control for the baseline speed and accuracy of the participating teachers if the sampled groups were not large enough to assume that the means of these characteristics are equal, as well as for other possible contributing factors such as teachers' number of years of work experience. Another factor that was not controlled for in this study is the hardware that teachers used to click while scoring (trackpad versus mouse). Lastly, there was no paper-based scoring condition in this study, although this was by far the most commonly used method in the daily teaching practice of all participating teachers. The quantitative indicators of speed in this study were only compared to a digital control condition, which the teachers verbally indicated to be much faster than their usual paper-based practice.

Other more minor limitations were a bug in the prototype (the highlighting of words did not work appropriately for one of the teachers), that one teacher accidentally started in the experimental condition (which was quickly corrected), and a lag in the scoring for one class with a particularly large amount of students. This lag caused the teacher to be slightly confused and to talk to the researcher. The conversation between the researcher and the teacher was filtered out of the scoring duration data, but it is possible that not all of the delay caused by the confusion was accounted for. These anomalies were all investigated, and it was judged that they likely had little impact on the results and conclusions of the study.

The results of this study will be used to inform our choices in future research and development on CheckMate. Furthermore, although the techniques used in the prototype to provide NLP-assisted scoring were good enough to be plausible to the teachers in this study, there are many possible improvements and extensions that can and should be made before it can be feasibly used at a larger scale. For example, we have since the execution of this study added a contextual spell-checker and support for partial credit items, and we now use sentence embeddings instead of word embeddings. Furthermore, we continue to do user-based research on interface and functionality requirements and plan to implement and test at a larger scale in the future.

This study represented a small step towards implementation of NLP-assisted scoring for open-ended test questions. If done well, we are confident that such scoring assistance can remove barriers for including open-ended questions in tests by alleviating some of the burden that teachers around the world face while scoring, and that it can contribute to the fair, valid and consistent assessment of students.

5.5 Appendix 1: items

Figure 5.3: Item #1. *Translation:* **Corn.** A corn plant has male and female flowers. The male flowers are located in a panicle at the top of the plant. The female flowers are located halfway up the plant. A corn plant has wind-pollinated flowers. The image shows flower characteristics that can help you determine this. **Name one of these characteristics.**

Figure 5.4: Item #2. *Translation:* **Salamander.** The Great Crested Newt is very rare in the Netherlands. It lives mainly near forests and hedgerows. Reproduction takes place in water. A female lays about 200 eggs. The larvae eat water fleas. Both biotic and abiotic factors are important for the Great Crested Newt's survival. **Note the abiotic factor from the text that is important for the Great Crested Newt.**

Figure 5.5: Item #3. *Translation:* **Salamander.** The Great Crested Newt is very rare in the Netherlands. It lives mainly near forests and hedgerows. Reproduction takes place in water. A female lays about 200 eggs. The larvae eat water fleas. Salamanders do not have a diaphragm. The organs at the top of the salamander's body cavity are similar to the organs in the human thoracic cavity. **Write down the name of an organ found at the top of the salamander's body cavity.**

Figure 5.6: Item #4. *Translation:* **Plants.** Some plants are adapted to a hot and dry climate. They absorb the carbon dioxide they need for photosynthesis in the dark through their stomata. The carbon dioxide is converted into another substance in cells and stored. These plants keep their stomata closed during the day. **Explain how this benefits the plant.**

Planten

Sommige planten zijn aangepast aan een heet en droog klimaat. De koolstofdioxide die ze nodig hebben voor de fotosynthese, nemen ze in het donker op via de huidmondjes. De koolstofdioxide wordt in cellen omgezet in een andere stof en opgeslagen.

Deze planten houden overdag de huidmondjes gesloten.

Leg uit welk voordeel dat de plant oplevert.

Figure 5.7: Item #5. *Translation:* **Plants.** Some plants are adapted to a hot and dry climate. They absorb the carbon dioxide they need for photosynthesis in the dark through their stomata. The carbon dioxide is converted into another substance within the cells and stored. This substance is stored in part P. **What is part P of the plant cell called?**

Planten

Sommige planten zijn aangepast aan een heet en droog klimaat. De koolstofdioxide die ze nodig hebben voor de fotosynthese, nemen ze in het donker op via de huidmondjes. De koolstofdioxide wordt in cellen omgezet in een andere stof en opgeslagen.

Deze stof wordt in deel P opgeslagen.

Hoe heet deel P van de plantencel?

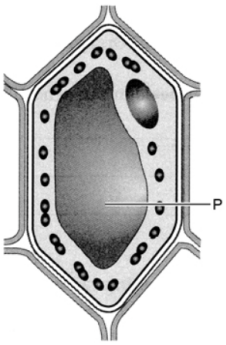


Figure 5.8: Item #6. *Translation:* **Low-fat diet.** "To prevent diabetes later in life, overweight people are better off eating low-fat products than smaller portions of high-fat foods." This conclusion was drawn by a scientist based on the results of a study. He investigated the effect of high-fat and low-fat foods on diabetes in mice. However, he believes that some overweight people will still develop diabetes despite eating low-fat products. **Explain why one person with a low-fat diet has a greater risk of diabetes than another with the same diet.**

Figure 5.9: Item #7. *Translation:* **Organ Systems.** You see parts of two organ systems working together: the circulatory system and the digestive system. **Explain how these two organ systems work together.**

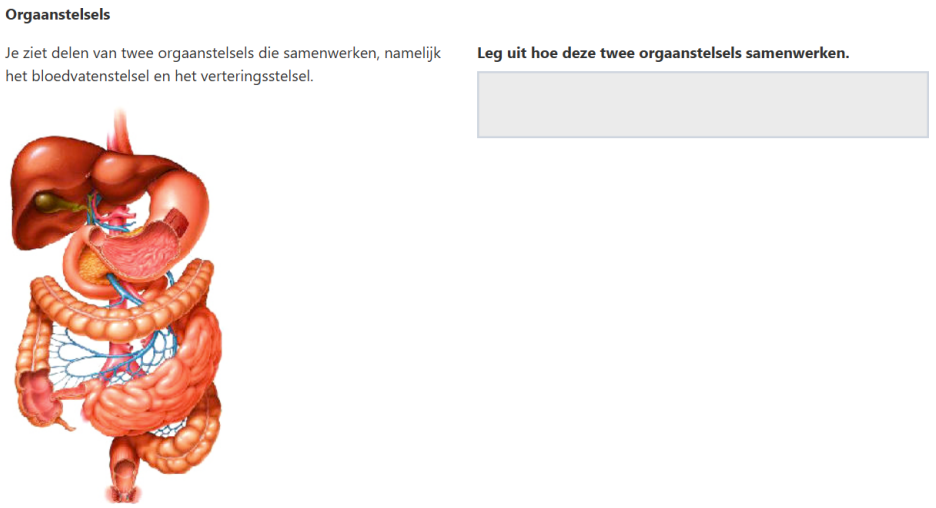
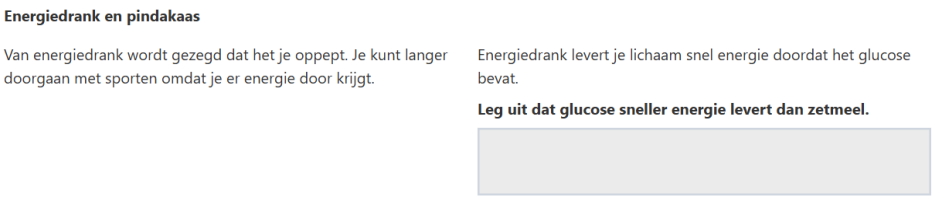


Figure 5.10: Item #8. *Translation:* **Energy drinks and peanut butter.** Energy drinks are said to give you a boost. You can exercise longer because they give you energy. Energy drinks provide your body with energy quickly because they contain glucose. **Explain that glucose provides energy faster than starch.**



5.6 Appendix 2: TeacherTapp survey

Table 5.3: What kind of test questions do you give your students (in summative assessments)?

Answer	% of teachers
Test questions that I constructed	84
Test questions from exams in earlier years	67
Test questions that were constructed by a colleague	65
Exercises from the course method	64
National central tests / final exams	36
Test questions that were constructed by other teachers (e.g. found online)	28
Other	12
Test questions from a student monitoring system	2
I do not test individually	1

Note. The response population consisted of 373 teachers in secondary education in the Netherlands.

Table 5.4: What methods do you use the most often when scoring your students' work?

Answer	% of teachers
Awarding points (without further explanation)	57
Awarding points (with further explanation)	50
Correcting mistakes	45
Noting some points of improvement	37
Feedback sentence (holistic feedback in one or more sentences)	24
Giving one compliment and one point of improvement	19
Detailed feedback	15
Other	7
Tests are scored automatically by software	5
Giving stamps or stickers	5
I do not score individually	1

Note. The response population consisted of 373 teachers in secondary education in the Netherlands.

Table 5.5: What makes scoring taxing for you?

Answer	% of teachers
I do not have enough time / it costs too much time	79
My students are only interested in scores and not in my feedback	53
I do not like scoring	35
There is no room in the schedule to act on the test results	33
Other	6
Disagreement between me and my colleagues about the scores	5
I doubt the correctness of the scores that I give	4
I do not find it useful	3
I do not score individually	1

Note. The response population consisted of 360 teachers in secondary education in the Netherlands.

5.7 Appendix 3: interview questions

First impressions

- What was your first impression of grading using CheckMate?
- How easy or difficult was grading using CheckMate for you? For what reason?
- How would you normally have scored such a test?

Functionalities

Speed

- Do you think CheckMate saved you time compared to grading such a test in your usual way? Why or why not?

Clustering of similar answers

- Did you find it helpful when answers were grouped together? If so, why?
- How does the system determine which answers should be grouped together?
- What could be changed about the grouping of the answers to make the functionality more helpful?

Score suggestions & scoring by other teachers

- How does the system determine a score suggestion?
- Did you use the score suggestions when grading? If so, in what way?
- Score suggestions for other answers sometimes changed while you were grading. What was your experience with this?
- How much do you trust the accuracy of the score suggestions (and therefore the scores that other teachers gave)? Does it matter to you who the teachers are that previously scored the test, in terms of your trust in the accuracy of the score suggestions?
- What could be changed about the score suggestions to make the functionality more helpful?

Highlighted words

- Did you use the highlighted words when grading? If so, in what way?
- What could be changed about the highlighted words to make the functionality more helpful?

Comparison to other systems

- Do you use systems that are similar to CheckMate? What is your experience with these systems? What are the differences between CheckMate and the system(s) you are currently using?
- Can you see yourself using CheckMate in your teaching practice? If so, how do you envision that? If not, why is that? What is needed in order for CheckMate to align with your teaching practice?

Other

- If CheckMate was helpful for you, which aspect was the most helpful?
- What could be added to or removed from CheckMate to increase the speed of your grading?
- How could the instructions for using CheckMate be improved?
- Is there anything else you would like to tell us?

5.8 Appendix 4: coding protocol

Code	Definition
	Theme: advantages
clear interface	The extent to which the program or answers are clearly structured and organized; ease of navigating the system.
confirmation of judgment	Paying attention to information that supports or confirms one's own belief or judgment.
consistency	The ability to review answers in a consistent way, without major variations or contradictions.
ease of use	The simplicity with which something can be used; implies that a product is efficient and understandable for the user, allowing them to perform tasks quickly and smoothly.
legibility	The ease with which answers can be read; student handwriting compared to digital text. Can also relate to accessibility and user interface (e.g., how clear and intuitive a program is).
objectivity	Impartiality; free from personal bias.
saving time	The difference in time between two activities; often linked to optimization and automation.
speed	The amount of time in which a task is performed (measure). Implies the ability to carry out tasks efficiently.
student experience	The perceptions and interactions students have with the system.
	Theme: disadvantages
confusion	A state of uncertainty or lack of understanding, possibly due to contradictory signals, complexity, or lack of clarity.
dyslexia	Not being able to see which student has dyslexia or how this has affected possible misspellings in the answers.
harmful influence	Exerting conscious or unconscious influence on thoughts or decisions; implies possible deception.
passive attitude	Acting in a waiting or unengaged way; lack of initiative or responsibility.
subjectivity	Reality influenced by personal feelings, experiences, beliefs, and opinions.
unclear interface	The absence of clarity or structure, making information hard to understand or the system hard to navigate. Leads to reduced usability.

use of color	Negative impact of color use on readability, accessibility, or interpretation of information.
	Theme: trust & distrust
algorithm	The extent to which teachers believe that outcomes or decisions generated by an algorithm are accurate and reliable.
computers	The extent to which teachers believe in the reliability and security of computers, as well as in the people and processes managing them.
content	The extent to which teachers trust the content of the test.
own judgment	The extent to which teachers believe their own grading is accurate and objective.
teachers in other schools	The extent to which teachers believe that assessments by teachers from other schools are accurate, objective, and of high quality.
teachers in own school	The extent to which teachers believe that assessments by colleagues from their own school are accurate, objective, and of high quality.
	Theme: autonomy
control	The ability to exercise control over a situation, process, or system; includes the ability to guide, decide, and act.
decision-making authority	The ability to make decisions about certain issues or activities (e.g., marking an answer correct or incorrect).
independence	The degree to which an individual acts and decides autonomously.
influence of others	The impact of others on decision-making; both conscious and unconscious influence on behavior, attitudes, and decisions.
responsibility	Making deliberate, thoughtful choices and being accountable for them to oneself or others; awareness of consequences; willingness to accept consequences.
seeking knowledge	The process of gathering and understanding information to make better-informed decisions.
	Theme: doubt
'gray area' answers	Answers that are not clearly right or wrong; their evaluation may depend on the teacher's interpretation.
unsure	A state of hesitation; not (entirely) sure about a belief, decision, or action.

CHAPTER 6

Discussion

A combination of increased data availability, improved computational processing power, and scientific modeling breakthroughs has accelerated the development of AI-based tools in education and offers opportunities to use data from educational assessment in new ways to gain additional insights into learning processes as well as to offer support to various educational stakeholders (Saab, 2025). This dissertation aimed to make use of those opportunities in a responsible and practice-oriented manner, and to add value in the shape of additional support or insights into learning processes to educational practice for students and teachers by doing so. The main research question was:

"How can we responsibly create additional value from educational assessment data?"

6.1 Main findings

The four studies in this dissertation address the main research question by exploring different methods through which different types of educational assessment data can generate added value. Collectively, the chapters suggest that value is created when assessment data are transformed into actionable insights or support that can improve learning, teaching, and decision-making through responsible, user-centered educational technology, but also that the added value is not a given and can come at a cost, e.g. in terms of the learner's experience or the time and expertise required for implementation.

More generally, the use of increasingly advanced technology in the classroom has great potential for offering additional and timely support for both learners and educators, and may even lead to improved fairness when students are treated more consistently. However, methods for using educational assessment data vary immensely in terms of transparency, explainability, financial cost, as well as other relevant criteria. Furthermore, educational technology can require extensive research efforts before implementation.

More advanced technology has greater potential benefits, but can also entail higher costs. An example of this is generative AI, which in recent years has made its way into

the lives of everyone with access to the internet and a device. The potential benefits of using generative AI in educational technology are great, such as its wide range of possible applications. However, the downsides of using generative AI can also be great, for example in terms of financial cost, environmental cost, ethical issues (e.g., copyright of material on which models are trained), replicability, and transparency. Furthermore, generative AI has increased potential of changing the relationship between learner and educator due to its human-like quality. It is necessary to assess separately for each use case and each technology whether the benefits outweigh the costs. We have aimed to do so in each of the four studies in this dissertation.

Chapter 2 demonstrates the technical feasibility and low computational cost of using recommender systems (a form of AI used to recommend products such as songs or movies) for providing automated, individualized practice tests to students. We can predict scores on test questions for students in many different ways, such as through the item response theory models often used in the field of psychometrics. From other fields (e.g., computer science) we can derive different kinds of models that use different types of information to make these predictions, or use similar information in a different way, opening the door to new methods of assembling personalized tests for students. In this study, we applied four different recommendation algorithms that are commonly applied in commercial settings to an educational assessment application: a baseline algorithm that takes into account overall, student, and item averages, a user-based collaborative filtering method, an item-based collaborative filtering method, and a singular value decomposition matrix factorization algorithm. We found that all applied models, including a simple baseline model, performed adequately for the task of assembling personalized practice tests when applied to existing student data.

In **Chapter 3**, we developed a new recommender system that was more grounded in educational theory by using cognitive diagnostic models. The resulting CDM-recommender takes additional item information (in our case: the content domain within the subject Biology) into account while making score predictions for students, yielding recommendations more driven by the content of the construct that is being studied in addition to the students' scores. We compared the performance of this recommender to several other recommendation algorithms, including one based on item response theory, and found their performance to be similar and adequate.

Subsequently, we applied the CDM-recommender in a practical experiment using the web-based prototype Biologie+. Through Biologie+, we examined student experience and learning gains when practicing with a personalized practice test versus a standard and fixed practice test. We found that there was no difference in learning gain between the groups during the course of the experiment, and that the students with the personalized

test found their practice session more difficult. The personalized practice tests were, by definition, more difficult, because they were designed to target gaps in the students' knowledge. Although the participating teachers appreciated the idea of personalization in the practice tests that were offered to their students, the results of this study do not indicate immediate learning gain benefits of this approach, and they do suggest potential disadvantages to learning experience. Therefore, we must conclude that in its current form, the benefits do not outweigh the costs from the perspective of the student.

In contrast, the responses of the teachers to the prototype developed for this study were very positive. They appreciated the user-friendly environment, domain-focused student reports, and large amounts of offered practice materials. Participation in this study was viewed as a privilege, which is a feat in a system where most studies place additional burdens on participants and asks them to endure this voluntarily or in exchange for a small reward. It is clearly feasible to design *some* research in such a way that participation has immediate and clear benefits for the participants, and we should strive to make this a reality as often as possible.

Where Chapters 2 and 3 focused on supplementing the role of the teacher by providing personalization that would be difficult to achieve manually within a teacher's working hours, **Chapter 4** intended to explore groundwork for enriching the information that teachers have about their students. Through analyzing log data from a digital assessment environment, we identified five groups of students that approached solving a mathematics item in different ways, and we related these approaches to their ability in mathematics. Grouping students based on their behavior in a digital assessment environment is a somewhat involved process. On the one hand, fit measures can be used to make data-driven decisions on the number of groups into which students are placed. On the other hand, the resulting groups of behaviors need to be meaningfully interpretable. Domain experts can play an important role in such analyses by using their expertise to guide choices that are difficult to make using data alone. The results of this study as well as other recent research from the field suggest that it is possible to make useful distinctions in the behavioral patterns of students in technology-enhanced digital assessment environments. Such distinctions can help make data-driven pedagogical choices. However, it takes further work in several areas to make the results fully applicable in educational practice, and even then they may only be valid for interpretation at the group level.

In order to validate the results of log data-based studies, it is important to cross-validate with input from other sources, such as think-aloud protocols or observational data. Furthermore, designing and interpreting log data analyses takes a lot of time, and the analyses are generally not easily replicated for other test items, since technology-enhanced items are often unique. Only when items are directly reused or very similar variants are made of them is it feasible to apply earlier data processing and analysis meth-

ods to new data. Otherwise, it is very costly and time-consuming to bring the results of such analyses to educational practitioners to be used in a classroom setting: the items need to be pre-tested, sufficient log data collected, analyses performed and the resulting conclusions validated. Lastly, should it be made possible to report results from log data analyses such as these in (close to) real-time to educational practitioners, there is still a lot of territory to gain with regards to the content and design of such reports.

In **Chapter 5**, we moved from gaining additional insights into student processes to supporting teachers in their duty of grading tests through technological scoring assistance. There are potential benefits to validity, reliability, and efficiency in doing so. First, the burden of scoring open-ended test questions introduces practical limitations for test construction and assembly, narrowing the types of questions that can feasibly be used. If the effort involved in scoring was not an issue, teachers and other test constructors could more freely choose how to measure the construct of interest, improving measurement validity. Second, introducing technological scoring assistance can help score student answers more consistently across the scope in which it is introduced, decreasing measurement error and potentially leading to fairer grading. For example, it may help negate personal biases such as the halo effect, and random personal factors such as distractions, but it may also help decrease variation across graders. It should not matter when a student's answer is being scored and by whom. Introducing scoring assistance may help increase measurement reliability and fairness for learners. Third, and perhaps most obviously, technological scoring assistance may help save teachers valuable time and lighten their administrative burden.

Scoring assistance can take many different forms, including many different functionalities and types of user interfaces. However, scoring assistance in the form of completely automated scoring has by far received the most attention in scientific literature. Far less work has been done on how teachers experience different forms of scoring assistance and what is necessary for the responsible implementation of such technologies. Through Chapter 5, we aimed to contribute to this body of literature, investigating the experience teachers had with a scoring interface as well as three additional functionalities that used techniques from the field of natural language processing (NLP), a branch of AI that is focused on the automatic processing of human language.

In the study, we found that the teachers were generally very positive about the scoring interface, in which they graded horizontally (per item as opposed to per student) and anonymously. According to the participating teachers, scoring the student answers using this interface was faster than their usual method of scoring, even without the additional NLP-based functionalities. The functionality of visually grouping similar student answers was also greatly appreciated. The teachers thought it would make their scoring faster as well as more consistent, which some of them valued perhaps more greatly than saving time. The participants had mixed feelings regarding the implemented score suggestions. On the

one hand, they appreciated getting insights into how other teachers had scored, because it helped them score an answer when in doubt and increased the consistency of scoring. On the other hand, they felt the score suggestions may be a harmful influence. This concern was also supported by quantitative evidence. The score suggestions functionality also led to a large number of discussions on teacher autonomy. If we have learned anything from this study, it is that there are great opportunities for supporting teachers in their grading duties, but we have to take great care to protect teachers' pedagogical autonomy.

6.2 Strengths and limitations

This dissertation has several notable strengths. First, it makes use of a wide array of data types and analytic techniques, ranging from response data and log data to natural language processing and recommender system algorithms. Second, all studies and prototypes were designed with the end-user in mind. By involving teachers and students throughout the design and testing phases, the research maintained a strong connection to educational practice, increasing both its relevance and potential for real-world application. Third, the dissertation demonstrates an innovative integration of educational assessment with techniques from other fields, such as artificial intelligence, data science, and user-centered design. This cross-disciplinary approach reflects current developments in educational measurement and provides a foundation for future collaborations between psychometrics and emerging technological domains.

At the same time, several limitations should be noted. The limited scale of several studies, both in terms of duration and number of participants, constrains the generalizability of the findings. While the results provide valuable exploratory insights, replication in larger and more diverse settings is needed to draw broader conclusions. Furthermore, the work in this dissertation is written mostly from psychometric and user-centered design perspectives with only limited integration with perspectives from other fields. Incorporating theories from various interpretative frameworks could provide valuable new perspectives on the data, methods, and results in this dissertation. Using theory and cognitive-behavioral models from the learning sciences would add valuable information on how learners actually process feedback, self-regulate and construct understanding. Furthermore, data interpretation is socially and culturally situated: cultural, institutional, and relational contexts impact whether data-informed practices translate into actual learning value. It would also be valuable to expand the discourse on data ethics: who benefits from educational technology, who is represented, and who may be marginalized through implementation?

Finally, it is important to recognize that, while interdisciplinary and participatory approaches enrich educational research, they also require substantial time and coordination. Meanwhile, educational technology is being developed at great speed by (oftentimes commercial) organizations that iterate quickly and collect large volumes of user data. Research

environments typically operate at a slower pace due to ethical, methodological, and resource constraints. We should aim to view this difference as an opportunity to combine the best of both worlds in collaboration. Researchers can contribute educational and scientific expertise to ongoing product development processes, ensuring that innovations are both evidence-informed and pedagogically sound. As researchers, we should position ourselves in the right place at the right time, contributing our educational and scientific expertise where it matters.

6.3 Advice for educational researchers and developers

1. Involve stakeholders from the start: co-design or at least consult educational stakeholders when starting research and development. Practitioners' needs and constraints should impact your design choices and further directions of development.
2. Start small and iterate. Run short, well-scoped pilot projects in a few classes, gather feedback from teachers and students, and iterate before wider roll-out. Use pilots to test real classroom fit, workload impact, and learning effects. Us researchers have the tendency to think at our desks for too long (guilty as charged). We can learn about important practical factors that impact research design much faster if we start small and iterate. For developers of educational technology, starting small and iterating can prevent significant amounts of time being spent working in the wrong direction.
3. What educational practitioners need are often simple things that are not necessarily at the forefront of scientific knowledge, or at the forefront of technological development. There is a large gap to bridge between science and educational practice and we should play our part in closing that gap.
4. When introducing educational technology, take particular care to investigate its impact on pedagogical autonomy. Allow for practitioners' personal choice and control when necessary.

6.4 Advice for educational practitioners

1. Treat AI as an augmenting tool, not a replacement. Keep ownership over instructional decisions and use your professional judgment.
2. Look for pedagogical alignment when choosing AI tools. Many are not grounded in learning theory.

3. Be aware of legal constraints around the use of AI tools. Confer with your institution's data protection officer before collecting or sharing student data.
4. Personalization and automation are not better by definition. Look for or gather solid evidence on new educational interventions and the use of technology in the classroom.
5. Involve yourself in research on and development of new educational technology (if at all possible — we know you are busy). Your input is vital, and giving it will benefit future learners and educators, if not yourself.

6.5 Conclusion

The four studies in this dissertation showed that educational assessment data create value when translated – through transparent, user-centered educational technology – into personalized insights, pedagogical feedback, and teacher support that strengthen both learning and teaching. However, such gains do not follow automatically when educational technology is applied. In order to achieve these gains, it is necessary to ground educational technology in theory and take an interdisciplinary and integrated approach, evaluating results from research and development collaboratively and thoroughly.

Summary – Test to tool: creating value from educational assessment data for learners and educators

Data are everywhere. In the field of educational assessment, data from tests are used to estimate the ability of students and characteristics of test questions (*items*), such as their difficulty. Tests are also increasingly often administered digitally, allowing more and different kinds of data to be captured. Besides capturing student answers (*responses*), we can now also capture indicators of a student's answering process, such as the time the student spent answering an item, or how often a student rewrote a response before submitting their answer. This increased availability of richer data, combined with recent improvements in computational processing power and scientific modeling breakthroughs, offers new opportunities for supporting learners and educators (e.g., more insights into learning processes, or higher automation of tasks), but new challenges and dangers simultaneously arise (e.g., model biases, ethical issues around learners' privacy).

These themes were more elaborately discussed in the introduction of this dissertation (**Chapter 1**), which also described the rise of AI in education, as well as the four included studies and the digital prototypes that were developed for two of those studies. The main research question of this dissertation was:

"How can we responsibly create additional value from educational assessment data?"

In the study in **Chapter 2**, we investigated the feasibility of recommending personalized practice tests to students based on scores from an earlier test by using recommender systems. Recommender systems are algorithms that are widely used by commercial companies to recommend products to their users. We applied five such algorithms to existing data from central examinations at the end of Dutch secondary education to predict scores for student responses from earlier years, and then compared these scores to those that the students actually received. The results of the study demonstrated technical feasibility and low computational cost for providing automated, individualized practice tests to students using recommender systems. We found that all applied models, including a simple baseline model, performed adequately for the task of assembling personalized practice tests.

Chapter 3 introduced a new type of recommender system which uses *cognitive diagnostic modeling* (CDM, a class of statistical models that assesses whether a student has achieved mastery on specific topics). With this CDM-recommender, we can recommend personalized practice tests in a way that also takes the topic of the item into account. We compared the CDM-recommender to several other recommendation algorithms and found their performance to be similar and adequate.

Next, we performed a classroom experiment in which students received either a personalized or fixed practice test, and examined the students' learning experience and learning gains. We found no difference in learning gain between the two groups of students during the course of the experiment, but students that received a personalized practice tests found their practice session more difficult. Although the participating teachers appreciated the idea of personalization in the practice tests that were offered to their students, the results of this study do not indicate immediate benefits of this approach, and they do suggest potential disadvantages to learning experience. In contrast, the responses of the teachers to the prototype developed for this study (Biologie+, see Section 1.3) and to participating in the study were very positive, which showed that the study was successful in terms of integrating the experiment with the teachers' daily practice.

In **Chapter 4**, we analyzed the data produced by students' actions in a digital assessment environment to uncover the solution strategy that the students used to solve a mathematics item. In the study, we identified five groups of students that approached solving the item in different ways, and we related these approaches to their ability in mathematics. Knowing what solution strategy a student used and whether they were successful in applying it can help teachers determine pedagogical actions towards their students, and potentially evaluate their own teaching leading up to the assessment. However, making the results of this study applicable to educational practice requires further steps to be taken in several areas.

Scoring open-ended questions is a difficult and time consuming task. **Chapter 5** investigated ways to make scoring student answers to open-ended questions more efficient and consistent. The study was executed using a web-based prototype named CheckMate (see Section 1.3), in which teachers score student answers horizontally (per item as opposed to per student) and anonymously. CheckMate has three additional features: 1) similar student answers are visually grouped, 2) unscored student answers are given score suggestions based on scores given by other teachers to similar student answers, and 3) words that are similar to or the same as important words in the correct answer are highlighted. We found that teachers were generally positive about the scoring interface, finding it easy to use and faster than their usual method of scoring.

The functionality of visually grouping similar student answers was greatly appreciated.

The teachers thought it would make their scoring faster as well as more consistent. However, they had mixed feelings regarding the suggested scores for unscored answers. On the one hand, they appreciated getting insights into how other teachers had scored, because it helped them score an answer when in doubt and increased the consistency of scoring. On the other hand, they felt the score suggestions may be a harmful influence. This concern was also supported by quantitative evidence. The score suggestions functionality also led to a large number of discussions on teacher autonomy. The results of this study showed us that there are opportunities for supporting teachers in their grading duties, but we have to take great care to protect teachers' pedagogical autonomy.

Chapter 6 discussed the results of the four included studies. Collectively, the studies suggest that value is created when assessment data are transformed into actionable insights or support that can improve learning, teaching, and decision-making through responsible, user-centered educational technology, but also that the added value is not a given and can come at a cost, e.g. in terms of the learner's experience or the time and expertise required for implementation. Lastly, the Chapter 6 also offered practical advice for educational researchers, developers and practitioners.

I advise educational researchers and developers to...

1. involve stakeholders in educational research and development from the start through co-design or consultation
2. start with small experiments and iterate
3. keep in mind that what practitioners need is often not the most technologically or scientifically advanced solution
4. pay particular attention to the impact of educational technology on pedagogical autonomy

I advise educational practitioners to...

1. treat AI not as a replacement but as an augmenting tool
2. look for pedagogical alignment when choosing AI tools
3. be aware of legal constraints around the use of AI tools
4. base the implementation of new educational interventions and technology in the classroom on solid evidence
5. involve themselves in research on and development of new educational technology

Samenvatting – Van toets naar tool: waarde creëren uit toetsdata voor leerlingen en docenten

Data zijn overal. In het onderwijs worden data uit toetsen gebruikt om de vaardigheden van leerlingen en de kenmerken van toetsvragen, zoals hun moeilijkheid, te schatten. Toetsen worden ook steeds vaker digitaal afgenomen, waardoor meer en verschillende soorten data kunnen worden verzameld. Naast het verzamelen van antwoorden van leerlingen, kunnen we nu ook indicatoren van het antwoordproces van een leerling vastleggen, zoals de tijd die een leerling aan een vraag besteedde of hoe vaak een leerling een antwoord herschreef voordat zij het antwoord indiende. De toegenomen beschikbaarheid van rijkere data, in combinatie met recente verbeteringen in rekenkracht en doorbraken in wetenschappelijke modellering, biedt nieuwe mogelijkheden om leerlingen en docenten te ondersteunen (met bijvoorbeeld meer inzicht in leerprocessen of een hogere mate van automatisering van taken). Tegelijkertijd ontstaan er echter ook nieuwe uitdagingen en gevaren, zoals het systematisch benadelen van specifieke groepen leerlingen door statistische modellen en ethische kwesties rond de privacy van leerlingen.

Deze thema's werden uitgebreider besproken in de inleiding van dit proefschrift (Hoofdstuk 1), waarin ook de opkomst van AI in het onderwijs werd beschreven, evenals de vier opgenomen onderzoeken en de digitale prototypes die voor twee van deze onderzoeken zijn ontwikkeld. De centrale onderzoeksvraag van dit proefschrift was:

"Hoe kunnen we op een verantwoorde manier toegevoegde waarde creëren uit toetsdata?"

In het onderzoek in **Hoofdstuk 2** onderzochten we de haalbaarheid van het aanbevelen van gepersonaliseerde oefentoetsen aan leerlingen op basis van scores van een eerdere toets met behulp van aanbevelingssystemen. Aanbevelingssystemen zijn algoritmen die door veel commerciële bedrijven worden gebruikt om producten aan hun gebruikers aan te bevelen. We pasten vijf van zulke algoritmen toe op bestaande data van centrale eindexamens in het Nederlandse voortgezet onderwijs om scores te voorspellen voor antwoorden van leerlingen uit eerdere jaren, en vergeleken deze scores vervolgens met de scores die de leerlingen daadwerkelijk behaalden. De resultaten van het onderzoek toonden aan dat het technisch haalbaar is om geautomatiseerde, gepersonaliseerde oefentoetsen aan leerlingen

aan te bieden met behulp van aanbevelingssystemen. We vonden dat alle toegepaste modellen, inclusief een eenvoudig basismodel, adequaat presteerden voor de taak van het samenstellen van gepersonaliseerde oefentoetsen.

In **Hoofdstuk 3** werd een nieuw type aanbevelingssysteem geïntroduceerd dat gebruikmaakt van *cognitief diagnostisch modelleren* (CDM, een klasse van statistische modellen die beoordelen of een leerling specifieke onderwerpen wel of niet beheerst). Met dit CDM-aanbevelingssysteem kunnen we gepersonaliseerde oefentoetsen aanbevelen, waarbij ook rekening wordt gehouden met het onderwerp van de toetsvragen. We vergeleken het CDM-aanbevelingssysteem met verschillende andere aanbevelingsalgoritmen en vonden vergelijkbare en adequate prestaties.

Vervolgens hebben we een experiment in de klas uitgevoerd waarbij leerlingen ofwel een gepersonaliseerde ofwel een standaard oefentoets kregen. We hebben de leerervaring en de leerwinst van de leerlingen onderzocht. We vonden geen verschil in leerwinst tussen de twee groepen leerlingen gedurende het experiment, maar leerlingen die een gepersonaliseerde oefentoets kregen, vonden hun oefensessie moeilijker. Hoewel de deelnemende docenten het idee van gepersonaliseerde oefentoetsen waardeerden, wijzen de resultaten van dit onderzoek niet op directe voordelen van deze aanpak en suggereren ze wel mogelijke nadelen voor de leerervaring. Daarentegen waren de reacties van de docenten op het voor dit onderzoek ontwikkelde prototype (Biologie+, zie Sectie 1.3) en op hun deelname aan het onderzoek zeer positief, wat aantoont dat het onderzoek succesvol was in het integreren van het experiment in de dagelijkse praktijk van de docenten.

In **Hoofdstuk 4** analyseerden we de data die ontstonden door de acties van leerlingen in een digitale toetsomgeving om de oplossingsstrategie te achterhalen die de leerlingen gebruikten om een wiskundige opgave op te lossen. In het onderzoek identificeerden we vijf groepen leerlingen die de opgave op verschillende manieren benaderden en relateerden we deze benaderingen aan hun wiskundige vaardigheden. Weten welke oplossingsstrategie een leerling heeft gebruikt en of deze succesvol is toegepast, kan docenten helpen bij het bepalen van hun pedagogische aanpak en mogelijk bij het evalueren van hun eigen onderwijs in de aanloop naar de toets. Om de resultaten van dit onderzoek echter toepasbaar te maken in de onderwijspraktijk, zijn er op verschillende gebieden nog stappen nodig.

Het beoordelen van open vragen is een lastige en tijdrovende klus. **Hoofdstuk 5** onderzocht manieren om het nakijken van open vragen efficiënter en consistenter te maken. De studie werd uitgevoerd met behulp van een digitaal prototype genaamd CheckMate (zie Sectie 1.3), waarin docenten de antwoorden van leerlingen horizontaal (per vraag in plaats van per leerling) en anoniem nakijken. CheckMate heeft drie extra functionaliteiten:

1) vergelijkbare antwoorden van leerlingen worden visueel gegroepeerd, 2) voor antwoorden van leerlingen die nog niet zijn nagekeken worden suggesties voor scores gegeven op basis van scores die andere docenten aan vergelijkbare antwoorden van leerlingen hebben gegeven, en 3) woorden die lijken op of hetzelfde zijn als belangrijke woorden in het antwoordmodel worden gemarkeerd. We constateerden dat docenten over het algemeen positief waren over de nakijkinterface; ze vonden deze gebruiksvriendelijk en sneller dan hun gebruikelijke nakijkmethode.

De functionaliteit van het visueel groeperen van vergelijkbare antwoorden van leerlingen werd zeer gewaardeerd. De docenten dachten dat dit hun nakijkproces sneller en consistenter zou maken. Ze voelden zich echter gemengd over de scoresuggesties voor nog niet nagekeken antwoorden. Enerzijds waardeerden ze het inzicht in hoe andere docenten hadden gescoord, omdat dit hen hielp bij het nakijken van een antwoord wanneer ze twijfelden en de consistentie van het nakijken verhoogde. Aan de andere kant waren ze van mening dat de scoresuggesties een schadelijke invloed konden hebben. Deze zorg werd ook ondersteund door kwantitatief bewijs. De functionaliteit leidde bovendien tot veel discussie over de autonomie van docenten. De resultaten van dit onderzoek tonen aan dat er zeker mogelijkheden zijn om docenten te ondersteunen bij hun nakijktaken, maar dat we er wel goed op moeten letten de pedagogische autonomie van docenten te beschermen.

Hoofdstuk 6 besprak de resultaten van de vier opgenomen onderzoeken. Gezamenlijk suggereren de onderzoeken dat er waarde wordt gecreëerd wanneer toetsdata worden omgezet in bruikbare inzichten of ondersteuning die het leren, lesgeven en de besluitvorming kunnen verbeteren door middel van verantwoorde, gebruikersgerichte onderwijstechnologie. De toegevoegde waarde is echter niet vanzelfsprekend en kan gepaard gaan met kosten, bijvoorbeeld in de leerervaring of de tijd en expertise die nodig zijn voor implementatie. Tot slot bood Hoofdstuk 6 ook praktisch advies voor onderwijsonderzoekers, -ontwikkelaars en -professionals.

Ik adviseer onderwijsonderzoekers en -ontwikkelaars om...

1. stakeholders vanaf het begin te betrekken bij onderwijskundig onderzoek en ontwikkeling door middel van co-design of consultatie
2. te beginnen met kleine en iteratieve experimenten
3. er rekening mee te houden dat onderwijsprofessionals vaak niet de meest technologisch of wetenschappelijk geavanceerde oplossing nodig hebben
4. bijzondere aandacht te besteden aan de impact van onderwijstechnologie op de pedagogische autonomie

Ik adviseer onderwijsprofessionals om...

1. AI niet te beschouwen als een vervanging, maar als een aanvullend hulpmiddel
2. op te letten op pedagogische afstemming bij de keuze van AI-tools
3. zich bewust te zijn van wettelijke beperkingen rond het gebruik van AI-tools
4. de implementatie van nieuwe onderwijsinterventies en -technologie in de klas te baseren op gedegen bewijs
5. zich te verdiepen in onderzoek naar en ontwikkeling van nieuwe onderwijstechnologie

Dankwoord

Meer dan zeven jaar na de start van dit promotietraject is het nu echt zover - het boekje ligt er. Ik ben blij en trots dat het ondanks tegenslagen toch is gelukt, maar bovenal ben ik dankbaar voor alle mensen die naast me hebben gestaan en dit mogelijk hebben gemaakt.

Op de eerste plaats wil ik mijn begeleidingsteam bedanken. Ik zou de waarde van jullie bijdrage kunnen proberen te vatten in meetbare variabelen zoals het aantal uur dat we hebben overlegd, het aantal mailtjes dat ik van jullie heb ontvangen, of eventueel zelfs in het aantal biertjes dat is gedronken na werktijd, maar dat zou geen valide meting zijn. Jullie betrokkenheid, aandacht en expertise zijn van onmeetbare waarde geweest. **Remco** – dank voor je vertrouwen, alle nuttige en rake feedback, alles wat je voor mij in gang hebt gezet, en bovenal voor het relativeren (*"Het is maar werk."*) wanneer de opgave van het schrijven van dit proefschrift voor mij te groot voelde. Ik waardeer je mentorschap enorm. **Bernard** – het was altijd fijn om met jou te overleggen. Dank voor je vrolijke aanwezigheid, je inspirerende ideeën en je goede adviezen. En nog belangrijker: dank je wel dat je mijn mentale gezondheid altijd boven mijn prestaties hebt gezet. **Jos** – dank je wel voor de rol die je in de eerste jaren in mijn begeleiding hebt gespeeld. Je wist altijd met een verse blik naar mijn werk te kijken en het te verrijken met nieuwe perspectieven.

Naast hen wil ik nog veel meer collega's bedanken die samen voor een fijne werksfeer hebben gezorgd. Aan **CoDE**: dank jullie wel dat jullie mij altijd welkom hebben doen voelen in Twente ondanks dat ik er als externe promovenda niet vaak was. **Cor** en **Marieke**, dank voor het initiatief dat jullie hebben genomen aan de start van mijn promotietraject om het mogelijk te maken. **Romy** – dank je wel voor al je steun en coaching. Je bent de beste leidinggevende ter wereld (vermoed ik, $n \approx 10$) en ik schep tegen iedereen over je op. Het werk dat ik in jouw team mocht doen geeft mijn proefschrift haar waarde. Aan de **Cito PhD's**: wat fijn dat ik jullie had om mee te sparren, leed mee te delen en de grote momenten mee te vieren. In het bijzonder heel veel dank aan mijn paranimf **Aranka** – jouw aanwezigheid en steun heeft me de afgelopen jaren vooruit gedreven. Bij jou kon ik terecht met zowel de grote als de kleinste vragen (*"Aranka, help! Hoe krijg ik dit venster weer terug op mijn andere scherm?"*). Ik ben heel blij dat we dit samen konden doen en kijk uit naar jouw verdediging. **Elise** – ik vond het heel waardevol om met je op te trekken

tijdens de eerdere jaren van dit traject. Dank je wel voor alle steun en het samenwerken op afstand in coronatijd. **Silvia** – met jou kan ik over alles praten. Dank je wel voor je luisterend oor en liefdevolle adviezen. En aan mijn vele andere gezellige en kundige collega's bij **CitoLab**: jullie zijn de reden dat ik met zoveel plezier naar kantoor kom. Dank voor alle kopjes koffie en thee, de warme gesprekken en de vele potjes tafelvoetbal en tafeltennis.

Ik wil ook graag de **docenten** die deelnamen aan de onderzoeken in dit proefschrift en meedachten bij de ontwikkeling van de prototypes CheckMate en Biologie+ bedanken. Jullie liefde voor het vak en passie voor het welzijn en de ontwikkeling van jullie leerlingen inspireren me. Aan de **proefschriftcommissie**: ik waardeer het enorm dat jullie de tijd en moeite hebben genomen om mijn proefschrift te beoordelen en deel te nemen in mijn commissie.

En dan zijn er nog alle mensen buiten mijn werkomgeving die mij gezond, gelukkig en scherp hielden. **Peter-Jan** – ik dank een groot gedeelte van mijn fysieke én mentale gezondheid aan onze talloze klimafspraken. Bij jou ga ik altijd vrolijker, energieker en zorgelozer weg dan toen ik kwam. Dat geldt overigens ook voor de wekelijkse repetities met **Vocal Group Contagious Collective**, met dank aan al die leuke, getalenteerde vrouwen. Aan mijn **familie**: dank jullie wel voor jullie ondersteuning, begrip, interesse en aanmoedigingen over de jaren heen. Ik heb altijd gevoeld dat jullie in mij geloven. Aan **Janne**: dank je wel voor je geduldige hulp bij alles wat met Python te maken had (in het bijzonder voor de momenten dat ik zelf niet bepaald geduldig was). **Felicia** – dank dat ik bij jou thuis de laatste hand aan de inhoud van mijn proefschrift mocht komen leggen, en dat je me toen zo hebt ontzorgd. **Victor** – dank voor alle liedjes die mijn werkdagen zoveel leuker hebben gemaakt (*DJ Turn It Up!*). Aan mijn paranimf **Yana**: er zijn vast niet veel mensen die gezegend zijn met zo'n vriendin als jij bent voor mij. Dank je wel voor je eindeloze steun (soms zelfs midden in de nacht) en dat je me op de kritieke momenten hielp beseffen waarom ik dit proefschrift wilde schrijven. Ook veel dank aan **Hendri**, **Oğuzhan**, **Jasper**, **Lotte**, **Abel**, **Lisette**, **Matthijs**, **Jesse**, **Nicole**, **Leon** en alle andere fijne mensen in mijn leven – het leven is leuker met jullie erin.

Eva de Schipper

Bibliography

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Adomavicius, G., & Zhang, J. (2012). Impact of data characteristics on recommender systems performance. *ACM Transactions on Management Information Systems (TMIS)*, *3*(1). https://doi.org/10.1145/2151163.2151166

Aher, S. B., & Lobo, L. (2013). Combination of machine learning algorithms for recommendation of courses in e-learning system based on historical data. *Knowledge-Based Systems*, *51*, 1–14. https://doi.org/https://doi.org/10.1016/j.knosys.2013.04.015

Alamri, H., Lowell, V., Watson, W., & Watson, S. L. (2020). Using personalized learning as an instructional approach to motivate learners in online higher education: Learner self-determination and intrinsic motivation. *Journal of Research on Technology in Education*, *52*(3), 322–352. https://doi.org/10.1080/15391523.2020.1728449

Albert, D., & Steinberg, L. (2011). Age differences in strategic planning as indexed by the Tower of London. *Child Development*, *82*(5), 1501–1517. https://doi.org/10.1111/j.1467-8624.2011.01613.x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education and others. (2014). *Standards for educational and psychological testing*.

Andersen, N., Zehner, F., & Goldhammer, F. (2023). Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning*, *39*(3), 841–854. https://doi.org/10.1111/jcal.12717

Araneda, S., Lee, D., Lewis, J., Sireci, S. G., Moon, J. A., Lehman, B., Arslan, B., & Keehner, M. (2022). Exploring relationships among test takers' behaviors and performance using response process data. *Education Sciences*, *12*(2), 104. https://doi.org/10.3390/educsci12020104

ATLAS.ti Scientific Software Development GmbH. (2024). Atlas.ti for windows (version 24.1.0).

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.903077

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31–36. https://doi.org/10.1177/0273475302250570

Balabanović, M., & Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Commun. ACM*, *40*(3), 66–72. https://doi.org/10.1145/245108.245124

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391–402. https://doi.org/10.1162/tacl_a_00236

Bejar, I. I. (2017). Validation of score meaning for the next generation of assessments: The use of response processes. In K. Ercikan & J. W. Pellegrino (Eds.). Taylor & Francis. https://doi.org/10.4324/9781315708591

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)? *Educational Psychology Review*, *33*(4), 1675–1715. https://doi.org/10.1007/s10648-021-09615-8

Bexte, M., Horbach, A., & Zesch, T. (2023, July). Similarity-based content scoring - a more classroom-suitable alternative to instance-based scoring? In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 1892–1903). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.119

Billsus, D., Pazzani, M. J., et al. (1998). Learning collaborative information filters. *Icml*, *98*, 46–54. https://cdn.aaai.org/Workshops/1998/WS-98-08/WS98-08-005.pdf

Bjork, E. L., Bjork, R. A., et al. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68), 56–64.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bloor, M., & Wood, F. (2006). *Keywords in qualitative methods: A vocabulary of research concepts*. Sage.

Bobadilla, J., Serradilla, F., & Hernando, A. (2009). Collaborative filtering adapted to recommender systems of e-learning [Artificial Intelligence (AI) in Blended Learning]. *Knowledge-Based Systems*, *22*(4), 261–265. https://doi.org/10.1016/j.knosys.2009.01.008

Bokde, D. K., Girase, S., & Mukhopadhyay, D. (2015). An approach to a university recommendation by multi-criteria collaborative filtering and dimensionality reduction techniques. *2015 IEEE International Symposium on Nanoelectronic and Information Systems*, 231–236. https://doi.org/10.1109/iNIS.2015.36

Bolsinova, M., & Tijmstra, J. (2017). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13–38. https://doi.org/10.1111/bmsp.12104

Bransford, J., Brown, A., Donovan, M. S., & National Research Council Commission on Behavioral and Social Sciences and Education. (2000). *How people learn: Brain, mind, experience, and school*. National Academy Press.

Breese, J. S., Heckerman, D., & Kadie, C. (2013). Empirical analysis of predictive algorithms for collaborative filtering. https://doi.org/10.48550/arXiv.1301.7363

Brennan, R. L. (2023). *Educational measurement*. Rowman & Littlefield.

Brooks, M., Basu, S., Jacobs, C., & Vanderwende, L. (2014). Divide and correct: Using clusters to grade short answers at scale. *Proceedings of the First ACM Conference on Learning @ Scale Conference*, 89–98. https://doi.org/10.1145/2556325.2566243

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 3–53). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_1

Bulut, O., Cormier, D. C., & Shin, J. (2020). An intelligent recommender system for personalized test administration scheduling with computerized formative assessments. *Frontiers in Education*, *Volume 5 - 2020*. https://doi.org/10.3389/feduc.2020.572612

Burleigh, E. (2026). Despite promises that ai will create more jobs, 1.2 million jobs were actually slashed last year – a grim throwback to losses from the 2008 financial crisis [https://fortune.com/2026/01/22/despite-promises-ai-create-more-jobs-1-2-million-jobs-actually-slashed-last-year/, Last accessed on 2026-02-19].

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Cacheda, F., Carneiro, V., Fernández, D., & Formoso, V. (2011). Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Trans. Web*, *5*(1). https://doi.org/10.1145/1921591.1921593

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*(1), 30–41. https://doi.org/https://doi.org/10.1016/j.cedpsych.2008.07.001

Cassady, J. C., & Gridley, B. E. (2005). The effects of online formative and summative assessment on test anxiety and performance. *The Journal of Technology, Learning and Assessment*, *4*(1).

Cetintas, S., Si, L., Xin, Y. P., & Hord, C. (2009a). Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Transactions on Learning Technologies*, *3*(3), 228–236. https://doi.org/10.1109/TLT.2009.44

Cetintas, S., Si, L., Xin, Y. P., & Hord, C. (2009b). Predicting correctness of problem solving from low-level log data in intelligent tutoring systems. *Proceedings of the 2nd International Conference on Educational Data Mining*. https://eric.ed.gov/?id=ED539076

Chang, C.-J., Chang, M.-H., Chiu, B.-C., Liu, C.-C., Chiang, S.-H. F., Wen, C.-T., Hwang, F.-K., Wu, Y.-T., Chao, P.-Y., Lai, C.-H., Wu, S.-W., Chang, C.-K., & Chen, W. (2017). An analysis of student collaborative problem solving activities mediated by collaborative simulations. *Computers & Education*, *114*, 222–235. https://doi.org/10.1016/j.compedu.2017.07.008

Chen, F., Cui, Y., & Chu, M.-W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education*, *30*(3), 481–503. https://doi.org/10.1007/s40593-020-00202-6

Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Recommendation system for adaptive learning [PMID: 29335659]. *Applied Psychological Measurement*, *42*(1), 24–41. https://doi.org/10.1177/0146621617697959

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, *10*, 486. https://doi.org/10.3389/fpsyg.2019.00486

Claire Wyatt-Smith, J. J. C. (Ed.). (2009). *Educational assessment in the 21st century*. Springer. https://doi.org/10.1007/978-1-4020-9964-9

Cui, Y., Chu, M.-W., & Chen, F. (2019). Analyzing student process data in game-based assessments with bayesian knowledge tracing and dynamic bayesian networks. *Journal of Educational Data Mining*, *11*(1), 80–100. https://doi.org/10.5281/zenodo.3554751

Cukurova, M., Kent, C., & Luckin, R. (2019). Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring. *British Journal of Educational Technology*, *50*(6), 3032–3046. https://doi.org/https://doi.org/10.1111/bjet.12829

De Schipper, E., Feskens, R., & Keuning, J. (2021). Personalized and automated feedback in summative assessment using recommender systems. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.652070

Del Gobbo, E., Guarino, A., Cafarelli, B., & Grilli, L. (2023). Gradeaid: A framework for automatic short answers grading in educational contexts–design, implementation and evaluation. *Knowledge and Information Systems*, *65*(10), 4295–4334. https://doi.org/10.1007/s10115-023-01892-9

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Deribo, T., Goldhammer, F., & Kroehne, U. (2023). Changes in the speed–ability relation through different treatments of rapid guessing. *Educational and Psychological Measurement*, *83*(3), 473–494. https://doi.org/10.1177/00131644221109490

Derr, K., Hübl, R., & Ahmed, M. Z. (2018). Prior knowledge in mathematics and study success in engineering: Informational value of learner data collected from a web-based pre-course. *European Journal of Engineering Education*, *43*(6), 911–926. https://doi.org/10.1080/03043797.2018.1462765

Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, *22*(1), 143–177. https://doi.org/10.1145/963770.963776

Desrosiers, C., & Karypis, G. (2011). A comprehensive survey of neighborhood-based recommendation methods. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 107–144). Springer US. https://doi.org/10.1007/978-0-387-85820-3_4

Ding, Y., Riordan, B., Horbach, A., Cahill, A., & Zesch, T. (2020, December). Don't take "nswvtnvakgxpm" for an answer –the surprising vulnerability of automatic content scoring systems to adversarial input. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 882–892). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.76

Dixson, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory Into Practice*, *55*(2), 153–159. https://doi.org/10.1080/00405841.2016.1148989

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dupriez, V., Delvaux, B., & Lothaire, S. (2016). Teacher shortage and attrition: Why do they leave? *British Educational Research Journal*, *42*(1), 21–39. https://doi.org/10.1002/berj.3193

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218. https://doi.org/10.1007/BF02288367

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429246593

Eggen, T., & Sanders, P. (1993). Psychometrie in de praktijk. *Arnhem: Cito Instituut voor Toetsontwikkeling*.

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, *128*, 1–12. https://doi.org/10.1016/j.compedu.2018.08.004

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning*, *36*(6), 933–956. https://doi.org/10.1111/jcal.12451

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410605269

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, *96*(34), 226–231.

Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education*, *106*, 83–96. https://doi.org/10.1016/j.compedu.2016.12.001

Flor, M., & Cahill, A. (2025). Automated scoring of open-ended written responses: Possibilities and challenges. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative digital-based international large-scale assessments : Foundations, methodologies, and quality assurance* (pp. 265–298). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-90951-1_11

Flor, M., & Hao, J. (2021). Text mining and automated scoring. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in r and python* (pp. 245–262). Springer International Publishing. https://doi.org/10.1007/978-3-030-74394-9_14

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553. https://doi.org/10.1080/00273171.2016.1171128

Funk, S. (2006). *Netflix update: Try this at home*. Retrieved December 11, 2006, from https://sifter.org/~simon/journal/20061211.html

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The r package cdm for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2), 1–24. https://doi.org/10.18637/jss.v074.i02

Ghauth, K. I., & Abdullah, N. A. (2010). Learning materials recommendation using good learners' ratings and content-based filtering. *Educational technology research and development*, *58*(6), 711–727. https://doi.org/10.1007/s11423-010-9155-4

Gobert, J. D., Kim, Y. J., Sao Pedro, M. A., Kennedy, M., & Betts, C. G. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. *Thinking Skills and Creativity*, *18*, 81–90. https://doi.org/10.1016/j.tsc.2015.04.008

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: On validating the interpretation of process indicators based on log data. *Large-Scale Assessments in Education*, *9*(1), 20. https://doi.org/10.1186/s40536-021-00113-5

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments* (pp. 407–425). Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0_24

Golub, G. H., & Reinsch, C. (1971). Singular value decomposition and least squares solutions. In F. L. Bauer (Ed.), *Linear algebra* (pp. 134–151). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-39778-7_10

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095

Hagen, K. (2025, July). Ouder spiekt mee in leerlingvolgsysteem. https://www.aob.nl/actueel/artikelen/ouder-spiekt-mee-in-leerlingvolgsysteem/

Hahn, M. G., Navarro, S. M. B., De La Fuente Valentín, L., & Burgos, D. (2021). A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access*, *9*, 108190–108198. https://doi.org/10.1109/ACCESS.2021.3100890

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Han, Z., He, Q., & Von Davier, M. (2019). Predictive feature generation and selection using process data from pisa interactive problem-solving items: An application of random forests. *Frontiers in Psychology*, *10*, 2461. https://doi.org/10.3389/fpsyg.2019.02461

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

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

He, Q., Borgonovi, F., & Paccagnella, M. (2019). Using process data to understand adults' problem-solving behaviour in the programme for the international assessment of adult competencies (piaac): Identifying generalised patterns across multiple tasks with sequence mining. *OECD Education Working Papers*, *205*, 1–50. https://doi.org/10.1787/650918f2-en

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, *166*, 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., Liao, D., & Jiao, H. (2019). Clustering behavioral patterns using process data in piaac problem-solving items. In *Theoretical and practical advances in computer-based educational measurement* (pp. 189–212). Springer, Cham. https://doi.org/10.1007/978-3-030-18480-3_10

He, Q., Shi, Q., & Tighe, E. L. (2023). Predicting problem-solving proficiency with multiclass hierarchical classification on process data: A machine learning approach. *Psychological Test and Assessment Modeling*, *65*(1), 145–177.

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In *Quantitative Psychology Research* (pp. 173–190). Springer. https://doi.org/10.1007/978-3-319-19977-1_13

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 750–777). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-4666-9441-5.ch029

Heldt, M., Massek, C., Drossel, K., & Eickelmann, B. (2020). The relationship between differences in students' computer and information literacy and response times: An analysis of iea-icils data. *Large-scale Assessments in Education*, *8*(1), 12. https://doi.org/10.1186/s40536-020-00090-1

Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 300–304. https://doi.org/10.1145/3657604.3664693

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, *89*(2), 140–145. https://doi.org/10.1177/003172170708900210

Honnibal, M., Montani, I., van Landeghem, S., & Boyd, A. (2020). SpaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303

Horbach, A., & Zesch, T. (2019). The influence of variance in learner answers on automatic content scoring. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00028

Hrastinski, S., Stenbom, S., Benjaminsson, S., & Jansson, M. (2021). Identifying and exploring the effects of different types of tutor questions in individual online synchronous tutoring in mathematics. *Interactive Learning Environments*, *29*(3), 510–522. https://doi.org/10.1080/10494820.2019.1583674

Huang, Z., Liu, Q., Zhai, C., Yin, Y., Chen, E., Gao, W., & Hu, G. (2019). Exploring multi-objective exercise recommendations in online education systems. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1261–1270. https://doi.org/10.1145/3357384.3357995

Hug, N. (2020). Surprise: A python library for recommender systems. *Journal of Open Source Software*, *5*(52), 2174. https://doi.org/10.21105/joss.02174

Jiang, Y., Cayton-Hodges, G. A., Oláh, L. N., & Minchuk, I. (2023). Using sequence mining to study students' calculator use, problem solving, and mathematics achievement in the national assessment of educational progress (naep). *Computers & Education*, *193*, 104680. https://doi.org/10.1016/j.compedu.2022.104680

Jiang, Y., Gong, T., Saldivia, L. E., Cayton-Hodges, G., & Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment. *Large-scale Assessments in Education*, *9*(1), 1–31. https://doi.org/10.1186/s40536-021-00095-4

Karpicke, J. D., Butler, A. C., & III, H. L. R. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? [PMID: 19358016]. *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009

Kerr, D., & Chung, G. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining*, *4*(1), 144–182. https://doi.org/10.5281/zenodo.3554647

Khribi, M. K., Jemni, M., & Nasraoui, O. (2008). Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, 241–245. https://doi.org/10.1109/ICALT.2008.198

Klehe, U.-C., Grazi, J., Ones, D., Anderson, N., Viswesvaran, C., & Sinangil, H. (2018). Conceptualization and measurement of typical and maximum performance. In *The sage handbook of industrial, work and organizational psychology* (pp. 73–87). SAGE Publications Ltd. https://doi.org/10.4135/9781473914940.n5

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention the-

ory. *Psychological bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Koren, Y. (2008). Factorization meets the neighborhood: A multifaceted collaborative filtering model. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 426–434. https://doi.org/10.1145/1401890.1401944

Koren, Y. (2010). Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Trans. Knowl. Discov. Data*, *4*(1). https://doi.org/10.1145/1644873.1644874

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37. https://doi.org/10.1109/MC.2009.263

Koren, Y., Rendle, S., & Bell, R. (2022). Advances in collaborative filtering. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 91–142). Springer US. https://doi.org/10.1007/978-1-0716-2197-4_3

Korthals, L., Rosenbusch, H., Grasman, R., & Visser, I. (2025). Grading university students with llms: Performance and acceptance of a canvas-based automation. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education. posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium, blue sky, and wideaied* (pp. 36–43). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-99264-3_5

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kröhne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*(2), 527–563. https://doi.org/10.1007/s41237-018-0063-y

Kulasegaram, K., & Rangachari, P. K. (2018). Beyond "formative": Assessments to enrich student learning [PMID: 29341810]. *Advances in Physiology Education*, *42*(1), 5–14. https://doi.org/10.1152/advan.00122.2017

Kumar, S., Chakrabarti, S., & Roy, S. (2017). Earth mover's distance pooling over siamese lstms for automatic short answer grading. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2046–2052. https://doi.org/10.24963/ijcai.2017/284

Lang, K. (1995). Newsweeder: Learning to filter netnews. In A. Prieditis & S. Russell (Eds.), *Machine learning proceedings 1995* (pp. 331–339). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-377-6.50048-7

Liang, G., Weining, K., & Junzhou, L. (2006). Courseware recommendation in e-learning system. In W. Liu, Q. Li, & R. W.H. Lau (Eds.), *Advances in web based learning – icwl 2006* (pp. 10–24). Springer Berlin Heidelberg.

Lindner, M. A., & Greiff, S. (2023). Process data in computer-based assessment: Challenges and opportunities in opening the black box. *European Journal of Psychological Assessment*, (4), 241–251. https://doi.org/10.1027/1015-5759/a000790

Luo, J., Dong, F., Cao, J., & Song, A. (2010). A context-aware personalized resource recommendation for pervasive learning. *Cluster Computing*, *13*(2), 213–239. https://doi.org/10.1007/s10586-009-0113-z

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 253–275. https://doi.org/10.1111/bmsp.12070

Maddox, B. (2023). The uses of process data in large-scale educational assessments. *OECD Education Working Papers*, *286*, 1–23. https://doi.org/10.1787/5d9009ff-en

Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G., & Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: Evidence from the oecd piaac assessment. *European Journal of Psychology of Education*, *33*(3), 543–558. https://doi.org/10.1007/s10212-018-0380-2

Madnani, N., Loukina, A., & Cahill, A. (2017, September). A large scale quantitative exploration of modeling strategies for content scoring. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 12th workshop on innovative use of NLP for building educational applications* (pp. 457–467). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5052

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., & Koper, R. (2011). Recommender systems in technology enhanced learning. In P. B. Kantor, F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 387–415). Springer. https://doi.org/10.1007/978-0-387-85820-3_12

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2024). *Dexter: Data management and analysis of tests* [R package version 1.5.0]. https://CRAN.R-project.org/package=dexter

Martin, T., Petrick Smith, C., Forsgren, N., Aghababyan, A., Janisiewicz, P., & Baker, S. (2015). Learning fractions by splitting: Using learning analytics to illuminate the development of mathematical understanding. *Journal of the Learning Sciences*, *24*(4), 593–637. https://doi.org/10.1080/10508406.2015.1078244

Melville, P., & Sindhwani, V. (2017). Recommender systems. In C. Sammu & G. I. Webb (Eds.). Springer.

Mohan, K., Bergner, Y., & Halpin, P. (2020). Predicting group performance using process data in a collaborative assessment. *Technology, Knowledge and Learning*, *25*(2), 367–388. https://doi.org/10.1007/s10758-020-09439-5

Molenaar, P. D. I. (2024). *Mens-ai-samenwerking in het onderwijs: De hybride toekomst [oration]*. https://hdl.handle.net/2066/310574

Mooney, R. J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. *Proceedings of the Fifth ACM Conference on Digital Libraries*, 195–204. https://doi.org/10.1145/336597.336662

Nagy, G., & Ulitzsch, E. (2022). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*, *82*(5), 845–879. https://doi.org/10.1177/00131644211045351

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2022). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*, *39*(3), 751–766. https://doi.org/10.1111/jcal.12719

Norman, G. R., Van der Vleuten, C. P. M., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education*, *25*(2), 119–126. https://doi.org/10.1111/j.1365-2923.1991.tb00037.x

OECD. (2024). *Pisa 2022 technical report*. https://doi.org/10.1787/01820d6d-en

OECD. (2025). *Results from talis 2024: The state of teaching*. https://doi.org/10.1787/90df6235-en

Olsher, S., Chazan, D., Drijvers, P., Sangwin, C., & Yerushalmy, M. (2023). Digital assessment and the "machine". In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of digital resources in mathematics education* (pp. 1–27). Springer International Publishing. https://doi.org/10.1007/978-3-030-95060-6_44-1

O'Mahony, M. P., & Smyth, B. (2007). A recommender system for on-line course enrolment: An initial study. *Proceedings of the 2007 ACM Conference on Recommender Systems*, 133–136. https://doi.org/10.1145/1297231.1297254

Ouahrani, L., & Bennouar, D. (2024). Paraphrase generation and supervised learning for improved automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *34*, 1627–1670. https://doi.org/10.1007/s40593-023-00391-w

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238–243.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, *43*(2), 119–134. https://doi.org/10.1080/15391523.2010.10782565

Philbert, L., Bernigole, V., Ninnin, L.-M., Santos, R. D., le Cam, M., Salles, F., & Rocher, T. (2022). *Cedre rapport technique, mathématiques 2019* (tech. rep.). Direction

de l'évaluation, de la prospective et de la performance. https://www.education.gouv.fr/media/133181/download

Pokropek, A., Żółtak, T., & Muszyński, M. (2023). Mouse chase: Detecting careless and unmotivated responders using cursor movements in web-based surveys. *European Journal of Psychological Assessment*, *39*(4), 299–306. https://doi.org/10.1027/1015-5759/a000758

Putnikovic, M., & Jovanovic, J. (2023). Embeddings for automatic short answer grading: A scoping review. *IEEE Transactions on Learning Technologies*, *16*(2), 219–231. https://doi.org/10.1109/TLT.2023.3253071

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rademakers, J., ten Cate, T. J., & Bär, P. (2005). Progress testing with short answer questions. *Medical Teacher*, *27*(7), 578–582. https://doi.org/10.1080/01421590500062749

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Reis Costa, D., Bolsinova, M., Tijmstra, J., & Andersson, B. (2021). Improving the precision of ability estimates using time-on-task variables: Insights from the pisa 2012 computer-based assessment of mathematics. *Frontiers in Psychology*, *12*, 579128. https://doi.org/10.3389/fpsyg.2021.579128

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 1–35). Springer US. https://doi.org/10.1007/978-0-387-85820-3_1

Ricci, F., Rokach, L., & Shapira, B. (Eds.). (2015). *Recommender systems handbook* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4899-7637-6

Rivera, A. C., Tapia-Leon, M., & Lujan-Mora, S. (2018). Recommendation systems in education: A systematic mapping study. In Á. Rocha & T. Guarda (Eds.), *Proceedings of the international conference on information technology & systems (icits 2018)* (pp. 937–947). Springer International Publishing.

Roediger III, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger III, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention [PMID: 16507066]. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Saab, N. (2025). *Digital agency in onderwijs en samenleving: Op naar kritisch en bewust technologiegebruik [oration]*. ICLON, Leiden University. https://hdl.handle.net/1887/4270715

Salles, F., Dos Santos, R., & Keskpaik, S. (2020). When didactics meet data science: Process data analysis in large-scale mathematics assessment in france. *Large-scale Assessments in Education*, *8*, 1–20. https://doi.org/10.1186/s40536-020-00085-y

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web*, 285–295. https://doi.org/10.1145/371920.372071

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2002). Incremental singular value decomposition algorithms for highly scalable recommender systems. *Fifth international conference on computer and information science*, *1*(012002), 27–28.

Schenke, W., & Meijer, J. (2018). *Datagebruik in het onderwijs. problematiek uiteengezet.* Kohnstamm Instituut, Amsterdam. https://kohnstamminstituut.nl/wp-content/uploads/2018/11/1011-Datagebruik-in-het-onderwijs.pdf

Schildkamp, K., & Poortman, C. (2022, March). Met data je onderwijs verbeteren. https://www.onderwijskennis.nl/kennisbank/met-data-je-onderwijs-verbeteren

Schlippe, T., Stierstorfer, Q., Koppel, M. t., & Libbrecht, P. (2023). Explainability in automatic short answer grading. In E. C. K. Cheng, T. Wang, T. Schlippe, & G. N. Beligiannis (Eds.), *Artificial intelligence in education technologies: New development and innovative practices* (pp. 69–87). Springer Nature Singapore.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics*, *22*(1), 1–36. https://doi.org/10.1007/BF00302715

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Shute, V. J., Ventura, M., Bauer, M., et al. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In *Serious games* (pp. 317–343). Routledge. https://doi.org/10.4324/9780203891650

Siddiqi, R. (2013). Impact of automated short-answer marking on students' learning: IndusMarker, a case study. *2013 5th International Conference on Information and Communication Technologies*, 1–7. https://doi.org/10.1109/ICICT.2013.6732782

Spruyt, B., Droogenbroeck, F. V., Siongers, J., & Kavadias, D. (2023). Het lerarentekort kritisch bekeken vanuit internationaal vergelijkend perspectief. *Tijdschrift voor Onderwijsrecht en Onderwijsbeleid*, *1*, 19–27. https://cris.vub.be/ws/portalfiles/portal/93235824/Spruyt_et_al_Lerarentekort_talis_TORB.pdf

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*, *10*, 777. https://doi.org/10.3389/fpsyg.2019.00777

Suen, K. Y., Yaneva, V., Ha, L. A., Mee, J., Zhou, Y., & Harik, P. (2023, July). ACTA: Short-answer grading in high-stakes medical exams. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of nlp for building educational applications (bea 2023)* (pp. 443–447). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.36

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Tang, T., & McCalla, G. (2005). Smart recommendation for an evolving e-learning system: Architecture and experiment. *International Journal on E-Learning*, *4*(1), 105–129. https://www.learntechlib.org/p/5822

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, *85*(2), 378–397. https://doi.org/10.1007/s11336-020-09708-3

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). The sage handbook of qualitative research in psychology. SAGE.

Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance [Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010)]. *Procedia Computer Science*, *1*(2), 2811–2819. https://doi.org/10.1016/j.procs.2010.08.006

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: opportunities, challenges and risks. *British Educational Research Journal*, *42*(3), 454–476. https://doi.org/10.1002/berj.3215

Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, *47*(1), 3–35. https://doi.org/10.3102/10769986211010467

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika*, *86*(1), 190–214. https://doi.org/10.1007/s11336-020-09743-0

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, *2*(1), 319–330. https://www.learntechlib.org/p/111481

Van der Linden, W. J., Glas, C. A., et al. (2000). *Computerized adaptive testing: Theory and practice* (Vol. 13). Springer. https://doi.org/10.1007/0-306-47531-6

Van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347. https://doi.org/10.1177/0146621609349800

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review*, *17*, 147–177. https://doi.org/10.1007/s10648-005-3951-0

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Vervaart, J. (2025, November). Han onderzoekt mogelijk datalek na ongeoorloofd ai-gebruik door docent. https://www.gld.nl/nieuws/8392671/han-onderzoekt-mogelijk-datalek-na-ongeoorloofd-ai-gebruik-door-docent

Vialardi, C., Bravo, J., Shafti, L., & Ortigosa, A. (2009). Recommendation in higher education using data mining techniques. *International Working Group on Educational Data Mining*. https://files.eric.ed.gov/fulltext/ED539088.pdf

Visser, I., & Speekenbrink, M. (2010). depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, *36*(7), 1–21. https://www.jstatsoft.org/v36/i07/

von Davier, M., Fishbein, B., & Kennedy, A. (Eds.). (2024). *Timss 2023 technical report (methods and procedures)*. Boston College, TIMSS & PIRLS International Study Center. https://timss2023.org/methods

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of educational psychology*, *105*(4), 932–945. https://doi.org/10.1037/a0031882

Wang, F. H. (2021). Interpreting log data through the lens of learning design: second-order predictors and their relations with learning outcomes in flipped classrooms. *Computers & Education*, *168*, 104209. https://doi.org/10.1016/j.compedu.2021.104209

Weegar, R., & Idestam-Almquist, P. (2024). Reducing workload in short answer grading using machine learning. *International Journal of Artificial Intelligence in Education*, *34*(2), 247–273. https://doi.org/10.1007/s40593-022-00322-1

Westera, W., Nadolski, R., & Hummel, H. (2014). Serious gaming analytics: what students log files tell us about gaming and learning. *International Journal of Serious Games*, *1*(2), 35–50. https://doi.org/10.17083/ijsg.v1i2.9

Wetenschappelijke Raad voor het Regeringsbeleid [WRR]. (2021). *Opgave AI. de nieuwe systeemtechnologie*. https://www.wrr.nl/publicaties/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie

Xiao, Y., He, Q., Veldkamp, B., & Liu, H. (2021). Exploring latent states of problem-solving competence using hidden markov model on process data. *Journal of Computer Assisted Learning*, *37*(5), 1232–1247. https://doi.org/10.1111/jcal.12559

Zaíane, O. (2002). Building a recommender agent for e-learning systems. *International Conference on Computers in Education, 2002. Proceedings.*, *1*, 55–59. https://doi.org/10.1109/CIE.2002.1185862

Zhang, M., & Andersson, B. (2023). Identifying problem-solving solution patterns using network analysis of operation sequences and response times. *Educational Assessment*, *28*(3), 172–189. https://doi.org/10.1080/10627197.2023.2222585

Zhang, Z. (2024). Personalized resource recommendation method of student online learning platform based on lstm and collaborative filtering. *Journal of Intelligent Systems*, *33*(1), 20240017. https://doi.org/10.1515/jisys-2024-0017

Zhu, M., & Feng, G. (2015). An exploratory study using social network analysis to model eye movements in mathematics problem solving. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 383–387. https://doi.org/10.1145/2723576.2723591