KU LEUVEN

FACULTY OF SCIENCE

Bringing a New Perspective to the Classroom

Detecting and Explaining Student Outliers

Anissa FAIK

Supervisor: Prof. Katrien Verbert
Mentor: *Jeroen Ooge*
Assessors:
Dr. Jingyuan Sun
Aristotelis Chrysakis

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Toegepaste Informatica

Academic year 2022-2023

Preface

Starting this journey, a single question resonated in my mind, posed by my mentor Jeroen Ooge: “How ambitious are you?” Little did I know that this year would test me in ways I never imagined, pushing me to the edge and lighting my passion for education. Those closest to me witnessed the immense challenge I embraced, not just in completing this thesis, but also in pursuing my educational bachelor’s and juggling my part-time teaching job. Stressful? That word falls short of capturing the whirlwind that consumed me. However, amid the chaos, I found myself enriched by the invaluable lessons imparted by Jeroen, my students, my family, and my partner. Their unwavering support is often unspoken but deeply felt. Allow me to express my gratitude to each of them. To Jeroen, thank you for your unwavering support and for constantly pushing me to exceed my limits.

Yemma, your steadfast support has molded me into the strong woman I am today. Your unwavering belief in me has been my guiding light. You taught me that knowledge knows no bounds. *Hakima*, your unwavering presence carried me through the darkest moments. Know that your support never went unnoticed, and I am forever grateful for your unwavering love. *Abdelmomin*, our endless discussions bring me immeasurable joy. I hope all your dreams come true, you have it in you. *Hana*, my little girl, your late-night interruptions are but a small price to pay for the immeasurable love and affection you shower upon me. My partner, *Mustapha*, thank you for being my rock through every challenge. You were there to navigate the storm alongside me, providing comfort and laughter during the most trying times.

And last but certainly not least, my students. *Isa, Haci, Narek, Karol, Yasser, Sebastian, Ilias, Daria, Fayssal, Ali S, Ali E, Atalay, Rania, Zeynep, Oliwier, Ayoub, Efecan, and Ceylan.* You brightened my days! I hope that I have inspired you to embrace an unwavering determination and to dream without limits. For in this journey of life, it is knowledge that propels us forward, shaping us into better versions of ourselves.

Contents

Abstract

In a world driven by artificial intelligence, the demand for transparency and understanding has given rise to explainable artificial intelligence (XAI). Particularly in education, where unfairness issues pose challenges, there is a pressing need for tailored XAI solutions. This calls for research to bridge the gap, empower educators, and ensure the effective integration of AI in education.

This thesis explores the need for explainable artificial intelligence (XAI) in education, addressing instructors' lack of understanding and trust in AI technologies. It focuses on three main research objectives: (1) investigating the need for XAI in education to enhance instructors' perception, trust, and support in integrating AI into teaching practices, (2) exploring the complex nature of trust and its impact on AI utilization in education, and (3) addressing the gap in providing explanations for detected anomalies or outliers in educational settings. To achieve these objectives, an existing e-learning platform, *Wiski*, is extended to incorporate an explainable outlier detection system. Additionally trust, perceived accuracy, effectiveness, and satisfaction of the explanations provided to teachers are assessed.

A switching replications study was conducted with 11 teachers, followed by semi-structured interviews. Our findings led us to the following conclusion (1) explanations and alignment with teachers' perceptions impact perceived accuracy, which in turn influences trust. (2) Teachers expressed overall satisfaction with both model- and data-centric explanations, utilizing them effectively to understand the system. However, due to low intervention rates and various influencing factors, determining the precise effectiveness of data-centric explanations was challenging. (3) Some teachers found value in sharing the dashboard with students, and teachers desired added control of the parameters and customization options.

Abstract (Nederlands)

In een wereld gedreven door kunstmatige intelligentie is de vraag naar transparantie en begrip geleid tot de opkomst van uitlegbare kunstmatige intelligentie (XAI). Met name in het onderwijs, waar uitdagingen op het gebied van oneerlijkheid spelen, is er een dringende behoefte aan op maat gemaakte XAI-oplossingen. Dit vraagt om onderzoek dat de kloof overbrugt, leerkrachten in staat stelt en zorgt voor een effectieve integratie van AI in het onderwijs.

Deze masterproef onderzoekt de behoefte aan uitlegbare kunstmatige intelligentie (XAI) in het onderwijs, waarbij wordt ingegaan op het gebrek aan begrip en vertrouwen van docenten in AI-technologieën. Het richt zich op drie belangrijke onderzoeksdoelstellingen: (1) het onderzoeken van de behoefte aan XAI in het onderwijs om de perceptie, vertrouwen en ondersteuning van docenten bij de integratie van AI in het lesgeven te verbeteren, (2) het verkennen van de complexe aard van vertrouwen en de impact ervan op het gebruik van AI in het onderwijs, en (3) het aanpakken van de kloof bij het bieden van uitleg voor gedetecteerde afwijkingen of uitschieters in onderwijsomgevingen. Om deze doelstellingen te bereiken, is een bestaand e-learning platform, genaamd *Wiski*, uitgebreid om een uitlegbaar uitschieterdetectiesysteem te bevatten. Daarnaast worden vertrouwen, waargenomen nauwkeurigheid, effectiviteit en tevredenheid van de aan leerkrachten verstrekte uitleg beoordeeld.

Er werd een switching replications-studie uitgevoerd met 11 leerkrachten, gevolgd door semi-gestructureerde interviews. Onze bevindingen leidden tot de volgende conclusie: (1) uitleg en afstemming met de perceptie van leerkrachten hebben invloed op de waargenomen nauwkeurigheid, die op haar beurt het vertrouwen beïnvloedt. (2) Leerkrachten waren over het algemeen tevreden met zowel model- als data-gecentreerde uitleg, waarbij ze deze effectief gebruikten om het systeem te begrijpen. Het bepalen van de precieze effectiviteit van data-gecentreerde uitleg was echter uitdagend vanwege het lage aantal interventies en verschillende beïnvloedende factoren. (3) Sommige leerkrachten vonden waarde in het delen van het dashboard met studenten, en verlangden meer controle over de parameters en aanpassingsopties.

Chapter 1

Introduction

In today's world, artificial intelligence (AI) has permeated various domains, including entertainment, health, finance, and education, intending to offer benefits and improve decision-making processes [1]. AI systems, such as recommender systems and social media algorithms, have become integral parts of our daily lives, providing personalized recommendations and connections based on our preferences and behaviors [1]. However, the lack of transparency in AI systems, often referred to as "black-boxes," raises concerns about the level of trust we can place in these opaque systems [1, 4, 7].

To address this challenge, there is a growing demand for AI systems to be interpretable and explainable, allowing users to understand the reasoning behind their predictions and decisions[1]. For this, the concept of explainable artificial intelligence (XAI) has emerged, aiming to provide explanations for the outcomes produced by opaque AI systems [1]. In the educational domain, the demand for XAI is particularly crucial. Studies, such as the work by Khosravi et al. [65], have highlighted the presence of unfairness issues that have hindered the widespread use of AI in education. This emphasizes the necessity for further research on explainable AI specifically tailored to the educational context. Additionally, it has been reported that instructors often lack sufficient understanding of AI, leading to concerns and hesitations in utilizing AI technologies in their teaching practices [64].

In this thesis, we focus on the following gaps and research objectives:

1. The need for XAI in education: Addressing instructors' lack of understanding of AI, enhancing their perception, and providing support in integrating AI into teaching practices, addressing critical concerns related to trust and effectiveness [64, 65].

2. Complex nature of trust: Exploring the multi-faceted nature of trust and its impact on the utilization of AI in education.

3. Explainable outlier detection in education: Addressing the gap in providing explanations for detected anomalies or outliers in educational settings, where existing literature primarily focuses on detection accuracy without considering the importance of explainability.

To address the identified gaps in the literature, this thesis takes a comprehensive approach. An existing e-learning platform, *Wiski*, is extended to include an explainable outlier detection system for teachers, identifying “weak” and “strong” students for educators. Explanations for these outlier detections are provided and their impact on trust and perceived accuracy is considered. Trust is evaluated using a quantitative approach that considers multiple factors influencing it, moving beyond a single-dimensional construct. By assessing explanations, this thesis aims to understand their effectiveness and satisfaction among teachers. Furthermore, the research explores how teachers intend to utilize a dashboard with outlier detection in a classroom setting.

In Chapter 2, we delve into the literature, providing background context and exploring related work. Chapter 3 outlines the research method and materials employed. The development process of the extended *Wiski* platform is discussed in Chapter 4. Moving forward, Chapter 5 presents the results obtained from the research, followed by a comprehensive discussion of these findings, the limitations of our work, and future work in Chapter 6. Finally, in Chapter 7, we conclude the thesis by summarizing our key findings.

Chapter 2

Background and Related Work

This literature study aims to explore the current state of research in the field of *explainable artificial intelligence (XAI)*, techniques to determine the proficiency level of students, and various outlier detection algorithms. The first section will provide an overview of XAI and its importance in developing transparent and trustworthy AI systems. The second section will focus on assessing student proficiency and determining exercise difficulty. The third section will delve into outlier detection algorithms. Overall, this literature study aims to provide a comprehensive understanding of these three topics and their related work.

2.1 Explainable Artificial Intelligence

Artificial intelligence (AI) is widely used and can be found in different fields such as entertainment, health, finance, education, military, etc. [1]. It has the intention to offer benefits for society by efficiently learning and making decisions for us. For instance, *recommender systems* (RS) are used to provide personalized recommendations to online shoppers, suggesting products based on their preferences and browsing history. Similarly, AI is employed by entertainment platforms to propose exciting new movies and series to viewers, and by social media platforms to suggest new connections and friends.

2.1.1 Lack of Transparency

As the use of AI is affecting our daily lives, a concern is surging around its transparency. For instance, an incident in Arizona highlights the potential risks associated with autonomous vehicles. In this case, a self-driving Uber failed to detect a pedestrian crossing the street, resulting in a tragic fatality [1, 2]. The adaptation of artificial intelligence in healthcare should also be integrated with caution as it can impact the lives of patients. An illustrative example is the case of an AI model that had learned pneumonia patients with asthma should not be admitted to a hospital and had overseen their urgent admission to the ICU [1]. A transparent system is essential to clarify ambiguous circumstances and prevent such situations from occurring in the future.

As these two examples emphasize the need for transparency from a technical and diagnostic standpoint, other important desiderata such as fairness are stressed as well. For instance, a screening system used by St. George's Hospital Medical School was found to discriminate against applicants based on ethnicity [17]. These concerns are not limited to transportation and healthcare; industries such as legal, finance, and the military also rely on decisions made by AI and urgently require transparency [1].

AI systems that do not disclose their internal workings are often referred to as "black-boxes" [1, 4]. This raises questions about the level of trust that can be placed in these opaque systems [7]. To address this issue, there is a growing need for AI systems to be both *interpretable* and *explainable*. Doshi-Velez et al. [3] define interpretability as the ability to explain or to present in understandable terms to a human. Arrieta et al. [4] emphasize the importance of explainability and define it as the details and reasons a model gives to make its functioning clear or easy to understand, given a certain audience.

Considering these definitions, *explainable artificial intelligence (XAI)* aims to provide explanations needed for the predictions made by opaque AI systems. Although XAI pivots around explainability, the term interpretability is more commonly used in the *machine learning (ML)* community. These two terms are closely related and are often used interchangeably. Interpretable AI systems are explainable if their decisions can be understood by humans [1]. XAI aims to understand and explain these opaque systems.

However, a clear definition or standard for XAI has yet to be established. Gunning et al. [4, 5, 6] from Darpa's Explainable Artificial Intelligence Program describe XAI as "a suite of machine learning techniques that enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners." Arrieta et al. [4] take on a different approach and define XAI based on the definition of "explanation" from the Cambridge Dictionary of English Language:"The details or reasons that someone gives to make something clear or easy to understand."[15]. Alternatively, Adadi et al. [1] refer to XAI as "the movement initiatives, and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept." Deviating from a formal definition, but rather explicating its research field and efforts resulting from it. Despite all the research, there is still no consensus on a formal definition for explainable artificial intelligence.

Additionally, different goals for XAI have been set by various researchers. These goals widely vary for numerous papers. Arrieta et al. [4] mention trustworthiness, causality, transferability, informativeness, confidence, fairness, accessibility, interactivity, and privacy awareness as goals for explainable artificial intelligence. Contrasting this approach, Adida et al. [1] mention reasons for XAI such as explain to justify, control, improve, and discover.

2.1.2 Explanations

Explainable artificial intelligence aims to provide more transparency with explanations. How these explanations are provided or what techniques are applied can vary widely, resulting in different researchers in the XAI field proposing a taxonomy or framework for this issue. This section discusses the key aspects of the classification of explanations and their capabilities in depth.

Scale

Mohensi et al. [16] explain that global explanations provide more insight into the internal functioning of an entire machine learning model, whereas local explanations refer to explaining the result of only one instance. This is similar to the distinction made by Adadi et al. [1] between understanding the model on a global or local level. The interpretability of a model differs according to its scope. Global interpretability refers to the understanding of the whole model and its internal functioning, while local interpretability aims for and is limited to explaining only one decision made by that model. The scope varies for each context depending on the target audience of the explainable artificial intelligence.

Post-hoc Explanations vs Interpretable Models

Models can either be interpretable by design or require ad-hoc explainers, as proposed by Mohensi et al. [16]. Models such as decision trees are already transparent by nature and do not need much explanation techniques whereas other models such as complex neural networks do require explanations. Ad-hoc explainers provide reasoning on why the decision is made by a machine learning algorithm. Similarly, Arrieta et al. [4] distinguish between (i) post-hoc explainability for models that are not interpretable by design, and (ii) models that are transparent and understandable by default.

What to Explain?

Furthermore, Mohensi et al. distinguish between explanations based on the question they answer. The following are described:

- **How Explanations:** Explain how the model works.
- **Why Explanations:** Explain why the model made a prediction for an instance. Arrieta et al. and Hohman et al. call this a local (instance) explanation. Here Mohseni et al. made a difference in model-agnostic and model-specific explanation techniques. Similar to the classification of Arrieta et al. [4].
- **Why-not Explanations:** Explain why a certain output was not the predicted value for an instance. Also called contrastive explanations.
- **What-if Explanations:** Explain the difference in predictions if the input or model parameters would be adjusted.
- **How-to Explanations:** Explain the certain adjustment to the input or model that would affect the output. Similar to the counterfactuals mentioned by Hohman et al.
- **What-else Explanations:** Explain by presenting the user with similar instances that generated the same (or similar) output from the model. Also called explanation by example by Arrieta et al. and nearest neighbors by Hohman et al.

In addition to the aforementioned techniques, Arrieta et al. [4] also discuss the method of explanations by simplification, which involves building and explaining a new simplified model based on the trained model. They also mention feature relevance explanation methods, which calculate a relevance score for each feature to quantify its impact on the model's output. On the other hand, Hohman et al. [18] introduce local instance explanations as a distinct method for quantifying the contribution of each feature to the model's prediction. Additionally, Hohman et al. [18] highlight regions of error as a technique that provides insights into areas where the model exhibits high uncertainty for specific predictions.

Anik et al. [70] also discuss the notion of data-centric explanations, which involve providing descriptions of the data utilized for training a model. In contrast, model-centric approaches, similar to Mohseni et al.'s global explanations, aim to explain the overall model rather than its performance in specific cases [71].

Furthermore, Arrieta et al. [4] classify these post-hoc explainability methods according to model-agnostic and model-specific techniques. Model-agnostic techniques are methods that can be plugged into any model such as explanation by simplification and feature importance explanation methods, whereas model-specific explainability techniques refer to the adaptation of these techniques for models that require specific tailoring for explaining their decisions.

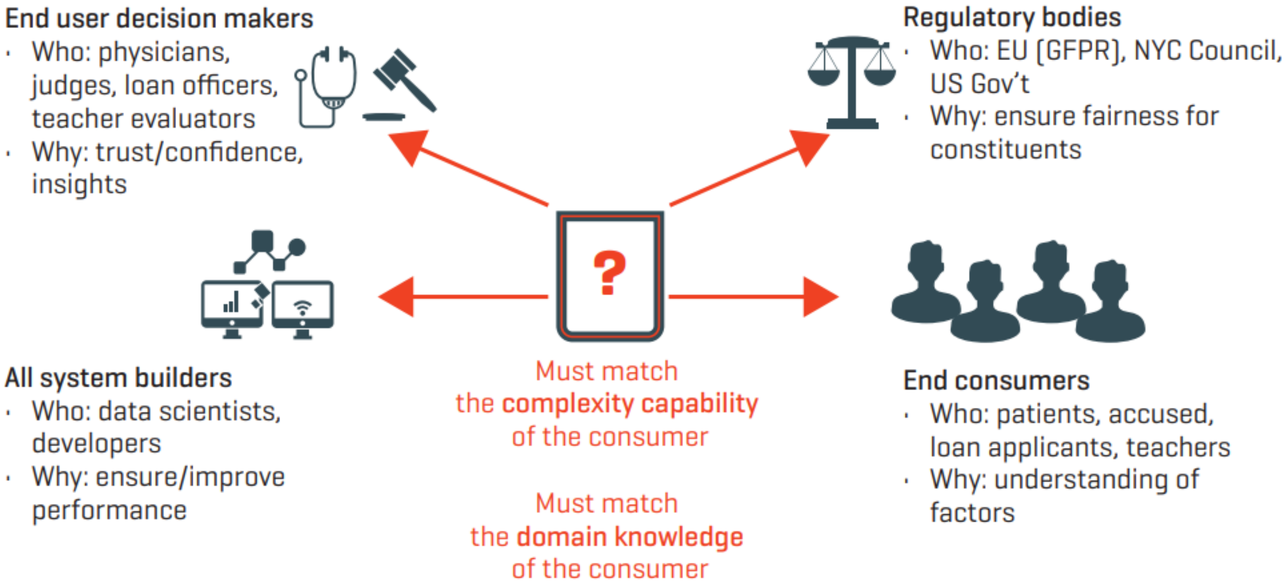
How to Explain?

Explanations can be presented in a visual, verbal, or analytical format [16]. Visual explanations use visual elements to provide insight into the internal reasoning of a machine learning algorithm. Verbal explanations use words, phrases, and natural language. Additionally, a combination of visual and textual explanations can also be implemented. Analytical explanations and visual explanations share similarities in terms of their visual component. However, analytical explanations primarily focus on exploring the data and the model's representations through the use of numerical metrics and data visualization techniques [72, 73, 74, 75]. This approach taken by Mohseni et al. [16] is similar to that of Arrieta et al. who also discuss textual and visual explanations.

Different Explanations for Different People

As explanations are provided for a certain group of users, different considerations need to be made. Mohseni et al. [16] distinguish between AI novices, data experts, and AI experts. Focusing on the difference in XAI design goals for each group, Mohseni et al. [16] describe AI novices as users of AI systems that have no expertise on machine learning systems. These users require XAI design goals such as algorithmic transparency, user trust, reliance, bias mitigation (ensuring fairness), and privacy awareness. Secondly, data experts such as data scientists and domain experts use AI systems for analysis and essentially decision-making. This user group has its own XAI design goals such as (i) model visualization and inspection, and (ii) model tuning and selection. The final user group is the AI experts who build and develop the AI system and provide either local or global interpretability of their model. XAI should provide local or global interpretability of the model for its developers as well as model debugging to diagnose and optimize it. Ribera et al. [63] take on the same user-centered XAI approach by dividing users into (i) developers and AI researchers, (ii) domain experts, and (iii) lay users.

Hind et al. [7] take a similar approach by making a detailed distinction between AI system builders, end-user decision-makers, regulatory bodies, and end consumers. Figure 2.1 outlines the diverse groups and their distinct needs. They introduce regulatory bodies as an additional group. Regulatory bodies such as government agencies strive to protect citizens and their rights and thus want to ensure that AI systems are fair for everyone. Explanations can provide insight into how decisions are made and if they are fair. Essentially, Hind et al. [7] and Mohseni et al. [16] both make a distinction based on the domain knowledge and complexity capability of the user. Bearing this in mind, explanations for the decisions of artificial intelligence should always consider the target audience and these two key aspects.



End user decision makers
- Who: physicians, judges, loan officers, teacher evaluators
- Why: trust/confidence, insights

Regulatory bodies
- Who: EU [GFPR], NYC Council, US Gov't
- Why: ensure fairness for constituents

All system builders
- Who: data scientists, developers
- Why: ensure/improve performance

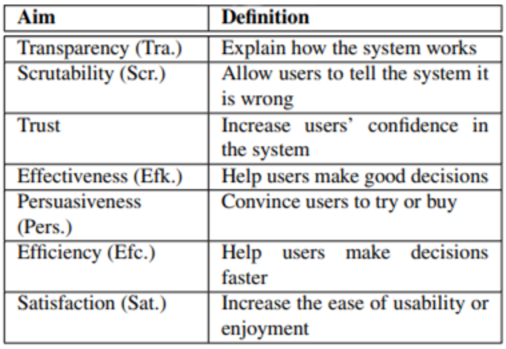
End consumers
- Who: patients, accused, loan applicants, teachers
- Why: understanding of factors

Figure 2.1: The different user groups (AI system builders, end-user decision makers, regulatory bodies, and end consumers) and their explanation needs from [7].

Aims

Tintarev and Masthoff [20] propose seven different aims that an explanation can have as shown in Figure 2.2.

Designing an explanation that incorporates all seven aims is exceedingly difficult because there exist various trade-offs between these explanation objectives. Tintarev and Masthoff [20] mention the following trade-offs. An increase in transparency might negatively



Aim	Definition
Transparency (Tra.)	Explain how the system works
Scrutability (Scr.)	Allow users to tell the system it is wrong
Trust	Increase users' confidence in the system
Effectiveness (Efk.)	Help users make good decisions
Persuasiveness (Pers.)	Convince users to try or buy
Efficiency (Efc.)	Help users make decisions faster
Satisfaction (Sat.)	Increase the ease of usability or enjoyment

Figure 2.2: Explanation aims proposed by Tintarev and Masthoff from [20].

impact the efficiency of the explanation [19]: as the interface provides an extensive explanation for a recommendation or prediction, this can increase the amount of time needed by users to process all the provided information. Another trade-off exists between the persuasiveness and effectiveness aim. An explanation can aim to persuade the user to take the recommendation or prediction into consideration which might result in the user making a spontaneous decision rather than thinking it through. Hoffman et al. [53] defined several key attributes of explanation satisfaction. This includes understandability, a feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness.

2.2 XAI in Education

Like many other fields, education is increasingly adopting artificial intelligence. *Educational AI* or *recommendation systems* differ from general-purpose AI or RS in many ways. Garcia-Martinez et al. [9] describe the following differences and factors for educational recommendation systems [9]:

- **Goal**: Educational RS has a distinct goal of facilitating learning objectives by recommending educational resources or activities.
- **Context**: The context of educational RS is more specific and takes into account factors such as pre- and post-requisites, timeframe, instructional design, pedagogical scenarios, and social network.
- **Pedagogical factors**: Educational RS also considers pedagogical factors such as learning history, knowledge, preferences, processes, strategies, styles, patterns, activities, feedback, misconceptions, weaknesses, progress, and expertise, which may not be as relevant in general-purpose recommendation systems.
- **User classification**: Users in educational recommendation systems are often classified based on their function (student, teacher, developer) or knowledge level (beginner, intermediate, advanced) to provide more tailored recommendations.

These factors for educational recommender systems can apply to artificial intelligence in an educational context as well. Not only emphasizing the obvious learning objectives of these AI systems but other factors such as the context, pedagogical factors, and predictions used for learning purposes is a common trend. However, concerns are emerging as well around the transparency of educational interventions supported by these artificial intelligence systems [8]. This led to a need for explainable artificial intelligence that sheds more light on how educational decisions are made by these opaque systems.

Open Learner Models

Open Learner Models (OLMs) are a prime example of providing transparency used to make pedagogical decisions. OLM model the pedagogical characteristics of the learner such as their knowledge and competencies to help facilitate the achievement of educational objectives [8, 10]. The beliefs inferred from the model are then used to reason and make predictions. These models are open and transparent, enabling the learner to gain insight into the model's understanding of their knowledge. By providing access to its learned beliefs, the model empowers the learner to assess their understanding and identify areas for improvement. One example of how transparency is provided is through a skill meter displaying the learner's proficiency in a particular skill or by comparing the learner's misconceptions to correct domain concepts [36]. Overall, an OLM aims to support reflection and self-regulation for learners [11]. Intelligent tutoring systems usually integrate learner models to assist the learning process [11, 13]. Conati et al. [13] establish a connection between OLMs and XAI, underscoring the significance of interpretability in AI-based educational systems. By leveraging the bridge between OLMs, educational AI systems can provide transparent and understandable insights into the recommendations and feedback offered by the system.

2.3 Determining the Proficiency Level of Students

Artificial intelligence and recommender systems are increasingly being used in education to personalize the learning experience of students. One way they do this is by determining the proficiency level of students. Applications such as adaptive learning systems build upon abstract representations of students and exercises. For instance, in adaptive learning, exercises or learning resources are recommended to students according to their needs. However, to do so, the difficulty of exercises needs to be determined and certain characteristics such as how knowledgeable students are about the various learning concepts must be modeled. Furthermore, the determined proficiency levels (or other characteristics) of students can then be used to recommend exercises according to their needs. This section discusses various methods to model a student's mastery level of a learning concept.

2.3.1 Item Response Theory

Item Response Theory originates from the psychometrics field and proposes that the probability of a (discrete) response to an item is in function of the characteristics of the item and the characteristics of the person responding to it [21]. IRT is described by Torkamaan et al. [22] as "the set of mathematical models that have been designed and used to explain the relationship between latent attributes and their outcomes." For instance, in an educational context, the latent attributes would emphasize the student's ability to answer questions and their relationship to the outcome (i.e., the answers to these questions). The Rasch Model is the simplest IRT model [21] and is widely used in the education domain [22]. It only has one latent parameter, namely the probability of the student u successfully answering an item i. The Rasch Model uses the following logistic function:

$$P(X_{ui} = 1 \mid \beta_u, \delta_i) = \frac{e^{\beta_u - \delta_i}}{1 + e^{\beta_u - \delta_i}} \tag{2.1}$$

X_{ui} denotes a binary variable that takes the value 1 if student u answers item i correctly. Additionally, β_u represents the ability of student u, and δ_i represents the difficulty of item i.

Using this logistic function results in an increasing probability of successfully resolving an exercise, as the difference between the student's ability and the exercise's difficulty decreases and vice versa. However, IRT models come with some disadvantages as well: they use a maximum likelihood estimation procedure which results in a computationally demanding algorithm and the initial calibration of the items' difficulty estimations. This model needs a vast number of test people for it to be reliable [20]. If this is not calibrated beforehand and estimated in the learning environment, it has the risk of having non-converging estimations as only a few students make a certain exercise in an immense range of items.

2.3.2 Elo Rating System

The Elo Rating System was originally designed for rating chess players. The reasoning behind the rating is quite simple which makes this system easily adaptable. Each player gets an Elo rating. The rating of a player is updated based on how surprising the outcome is or the expected probability of it [23]. If a strong player wins as expected, the update to their rating will be minor, but if a weak player beats a strong player, the update to their rating will be large. The Elo Rating System estimates a rating θ_i for a player i. $R_{ij} \in \{0, 1\}$ represents the probability of player i winning a match against player j. The "surprisingness" of a win is calculated by the expected probability that player i wins. A shifted logistic distribution is used with respect to the difference in Elo ratings:

$$P(R_{ij} = 1) = \frac{1}{1 + e^{-(\theta_i - \theta_j)}} \tag{2.2}$$

To update the Elo ratings of the players, the following formula is used, with K being a constant that indicates the weight held by the latest match:

$$\theta_i := \theta_i + K(R_{ij} - P(R_{ij} = 1)) \tag{2.3}$$

The Elo Rating System can be used in education by assigning ratings to students and exercises. A student solving an exercise is viewed as a match. When a "weak" student solves a hard exercise, it leads to a larger rating update and vice versa. Adapting the Elo Rating System results in estimating two parameters θ_s and d_i, respectively for student s and item i, with K being a constant that indicates the weight held by the latest attempt:

$$\theta_s := \theta_s + K(\text{correct}_{si} - P(\text{correct}_{si} = 1)) \tag{2.4}$$

$$d_i := d_i + K(P(\text{correct}_{si} = 1) - \text{correct}_{si}) \tag{2.5}$$

The constant K needs to be set carefully, as it dictates how fast the estimations converge. If K is too small, the estimations converge slowly. If K is too large, it causes instabilities as too much weight is given to the last attempt. An alternative solution is to use an uncertainty function that gives more weight to attempts from new players and decreases the influence of new attempts as more data is gained.

Multivariate Elo Rating System

Abdi et al. [25] introduced a multivariate extension of the basic Elo Rating System that can model students' skills and items' difficulties with multiple learning concepts. This provides a more detailed insight into the skills of a student by modeling their knowledge for each component. The extended multivariate model estimates a skill parameter for each concept for a student. Thus, modeling the mastery level of a student for each topic.

As the basic Elo Rating System, only one parameter is estimated for the difficulty of an item.

2.4 Outlier Detection

Outlier detection systems use AI techniques such as machine learning algorithms, both supervised and unsupervised. These algorithms allow the identification of data points that deviate from the typical patterns within a dataset. Outlier detection systems are widely used in various fields. The most common example is fraud detection where these outlier detection algorithms are used to determine if bank transactions seem unusual [31]. Other fields, such as healthcare, use outlier detection as well [32]. For instance, patient safety care where unusual patterns are detected in patient data.

Different unsupervised ML algorithms exist to identify outliers in a dataset. There are distance-based, neighbor-based, and isolation methods.

Distance-based methods use distance measures to detect outliers by calculating the distance between a point and the rest of the data points. A data point is considered an outlier if a certain fraction of the other data points in the dataset is farther away from it than a specific distance [37, 38]. Different distance measures, such as the Euclidean distance and the Mahalanobis distance, can be used.

Neighbor-based methods analyze the neighboring data points for each instance of the dataset to identify outliers. Neighbor-based or density-based methods use distance metrics, such as the Euclidean distance, to compute the density of the region around a data point. Outliers are usually farther away from their nearest neighbors, thus forming low-density regions [37]. One example of a density-based method is Local Outlier Factor (LOF). It ranks data points according to their computed "outlierness" score, which compares the local density of a datapoint to the local density of its k nearest neighbours [39, 40]. If a point has a significantly lower density than its neighbors, it's more likely to be an outlier.

The local density of a point x is computed using the reachability distance. The reachability distance between two points x and x_k is the maximum of the k-distance of x_k, which is the largest distance to one of its k nearest neighbours, and the distance from x to x_k [41]:

$$\text{Reachability}(x, x_k) = \max(\text{k-distance}(x_k), d(x, x_k)). \tag{2.6}$$

The local density of x is the inverse of the average reachability distance between the nearest neighbor x_k to x [41]:

$$\text{LDensity}(x) = \frac{k}{\sum_{x_k \in N_k(x)} \text{reachability}(x, x_k)} \tag{2.7}$$

The LOF score is computed as the average of the ratios of the local density of x to its k nearest neighbors and the kth nearest neighbor's local density [41]:

$$LOF(x) = \frac{1}{k} \sum_{x_k \in N_k(p)} \frac{\text{LDensity}(x_k)}{\text{LDensity}(x)} \tag{2.8}$$

The higher the LOF-score, the more likely the point is an outlier. According to Breunig et al. [40, 45], data points with LOF-scores of 1.0 or lower have similar or higher densities than their neighbors. Any data point with a LOF-score above 1.0 is deemed to have a significantly lower density than its neighbors and is thus identified as an outlier.

Other density-based methods include Angle-based Outlier Detection (ABOD) and Subspace Outlier Detection (SOD). ABOD measures the angles between data points and their neighboring points. Larger angles indicate higher outlierness, based on the assumption that outliers have larger spatial deviations from their neighbors [26, 29]. SOD focuses on a subspace of the most relevant features. This subspace is used to determine how likely the point is an outlier by computing the deviation from its neighbors. The standard deviation of the point from the mean of the subspace is calculated, if this value is above a certain threshold, it is considered an outlier [26, 28].

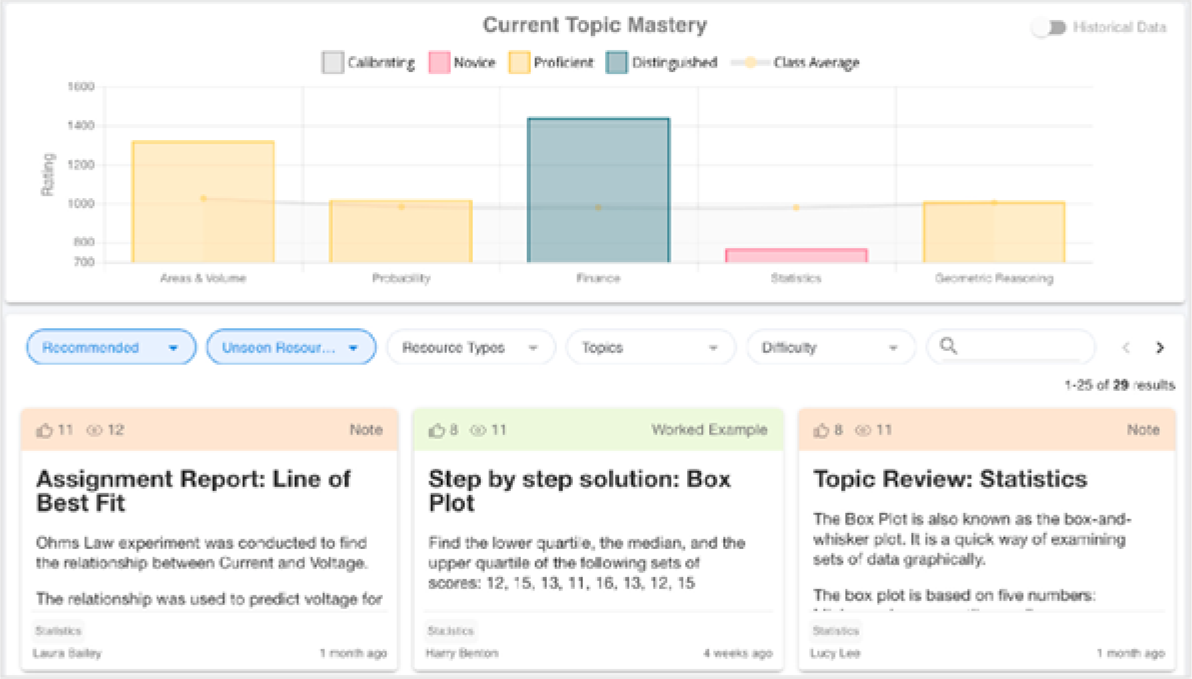
Isolation method such as Isolation Forest by Lui et al. [30] compute an isolation score for each instance of the data set [26]. Random recursive splits are performed on the different attributes and their values. This results in a tree that isolates certain data points based on a condition on an attribute. The isolation score for an instance is then computed by the path length to the data point. This means that smaller path lengths indicate that the data point is easily isolated from other instances, which suggests that the data point is an outlier.

2.5 Examples of (Explainable) AI-based Systems in Education

In the realm of adaptive learning systems, various interfaces, and visualizations have been developed to enhance the educational experience and support instructors in online classrooms. This introduction explores four distinct systems: RiPPLE, Groupnamics, iMoodle, and Mastery Grids. RiPPLE leverages open learner models and visualizations to adapt learning resources and activities based on student needs. Groupnamics provides visual interfaces for overseeing group discussions, and facilitating instructor engagement. iMoodle utilizes association rules to predict at-risk students and offers real-time interventions for teachers. Mastery Grids employs visualizations to represent learners' knowledge levels, supporting personalized learning and facilitating understanding of individual progress. Together, these systems contribute to the advancement of adaptive learning and instructional support in diverse educational contexts.

2.5.1 RiPPLE

RiPPLE is a crowdsourced adaptive learning system. The main objective of RiPPLE is to adapt the available learning resources and activities to the needs of the student. It also uses crowdsourcing, enabling collaboration with students and instructors by letting them evaluate these learning resources adapted to them [14]. These adaptations are made using an open learner model (see section 2.2). RiPPLE implements a multivariate Elo-rating system to assign a score rating the proficiency level of a student on a certain learning concept (see section 2.3.2). These ratings are used to adapt resources and activities according to the student's needs. As one of the characteristics of an OLM, RiPPLE makes this representation of the learner available to them. Visualisations are used to display the student's current mastery level (e.g., novice, proficient, distinguished) and its progress over time. A line chart is also used as an example-based explanation approach to show the average competency level of each knowledge concept. Finally, RiPPLE uses local explanations to explain the recommended resources to the individual students. The recommended resources and the learner model are displayed on the same page using the same mastery level colour scheme, to increase transparency and show how the recommendations align with the student's estimated competency level. This allows the student to see how the resources are tailored to their specific needs, and to understand how the recommendations are generated. This leads to students' trusting the recommendations made to them [14].



Assignment Report: Line of Best Fit

Ohms Law experiment was conducted to find the relationship between Current and Voltage.

The relationship was used to predict voltage for

Laura Bailey 1 month ago

Step by step solution: Box Plot

Find the lower quartile, the median, and the upper quartile of the following sets of scores: 12, 15, 13, 11, 16, 13, 12, 15

Harry Benton 4 weeks ago

Topic Review: Statistics

The Box Plot is also known as the box-and-whisker plot. It is a quick way of examining sets of data graphically.

The box plot is based on five numbers:

Lucy Lee 1 month ago

Figure 2.3: RiPPLE's explanation for recommendations and the visualised OLM from [12].

2.5.2 Groupnamics

Groupnamics is a visual interface designed to support instructors in overseeing parallel group discussions in online classrooms [54]. It offers a view of student groups, showcasing their recent vocal activities, discussion statuses and direct messages. The interface incorporates various visualisations (see Figure 2.4), including gray circles representing students' speaking status, which provides instructors with an overview of conversation frequency and balance within each group. Additionally, the background color of each group's name box transitions from gray to orange, signaling longer periods of silence and drawing attention to inactivity. Students can express their status as "HELP" or "DONE," leading to a corresponding change in the group box's color to red or green, respectively. Moreover, students can send messages directly to the instructor at the group level, with the messages displayed within the respective group's message window.

Groupnamics primarily focuses on assisting instructors in effectively overseeing groups and determining the necessary level of granularity for monitoring and intervention. Its visualisation interface provides a way to manage parallel group discussions in online classrooms and enhancing instructors' ability to engage with students.

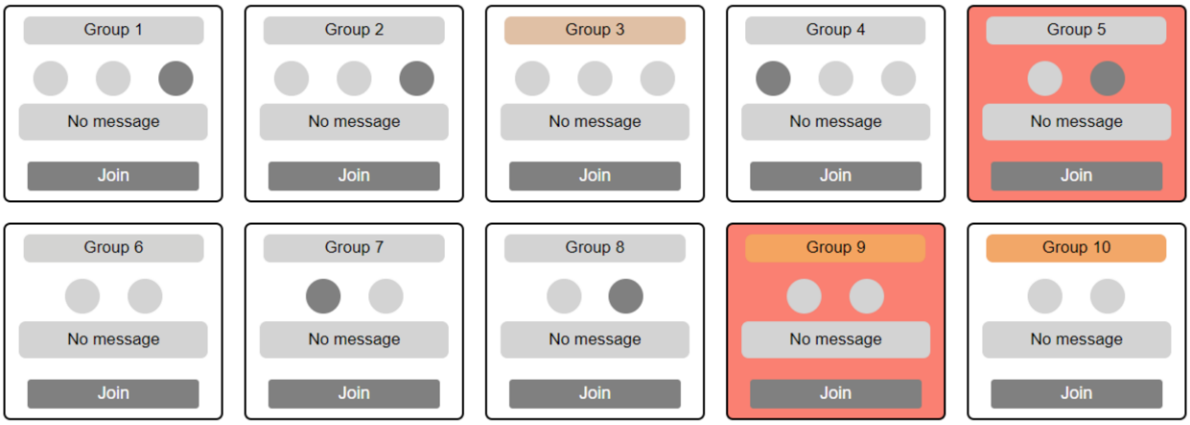


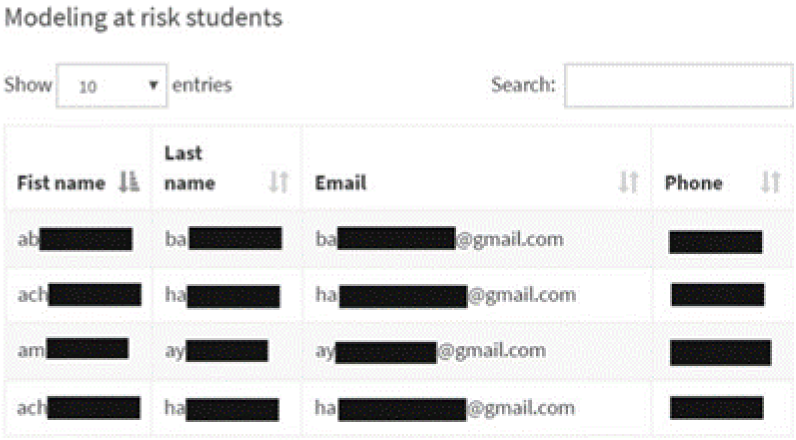
Figure 2.4: Groupnamics incorporating various visualisations to oversee parallel groups.

2.5.3 iMoodle

iMoodle is a gamified e-learning platform that provides a teacher-oriented dashboard visualising the learning processes of students. The main objective of iMoodle is to increase the students' success rating by predicting the students who are at risk of failing the semester course [35] as well as supporting teachers through notification alerts to stimulate real-time interventions. iMoodle predicts at-risk students using association rules based on the Apriori algorithm. This algorithm discovers relationships among attributes and provides if-then statements regarding the value of an attribute in the form of $X \Rightarrow Y$. This algorithm was first applied to another dataset from a university of Tunisia to obtain the needed association rules to make predictions based on student performance. From these rules, five factors were selected to help identify students who are at-risk of failing: (1) the number of acquired badges, (2) grades, (3) student's rank on the leaderboard, (4) course progress (number of finished activities), and (5) form and chat interactions. The at-risk students are predicted based on the associated rules and then displayed to encourage teachers to intervene (see Figure 2.5). The way these students were identified is not explained to the teacher, though a dashboard is provided that displays the metrics related to the students' performances and engagements (e.g., the completion rate of each learning activity).

2.5.4 Mastery Grids

Guerra-Hollstein et al. [83] developed the Mastery Grids (MG) interface as a fine-grained approach to learner modeling (OLM). It visualizes learners' knowledge levels using an



Modeling at risk students

Show 10 entries Search:

Fist name	Last name	Email	Phone
ab[redacted]	ba[redacted]	ba[redacted]@gmail.com	[redacted]
ach[redacted]	ha[redacted]	ha[redacted]@gmail.com	[redacted]
am[redacted]	ay[redacted]	ay[redacted]@gmail.com	[redacted]
ach[redacted]	ha[redacted]	ha[redacted]@gmail.com	[redacted]

Figure 2.5: iMoodle displaying the at-risk students from [35].

OLM. The visualizations in the Mastery Grids (MG) interface provide insights into the learner's knowledge levels by representing their mastery of knowledge components (KCs) and topics. By using different intensities of green to indicate the learner's knowledge and blue to represent the peer group's average knowledge, the interface enables easy comparisons and facilitates understanding of individual progress and relative performance. The interactive nature of the visualizations allows learners to explore specific topics and activities, enhancing their engagement and facilitating personalized learning.

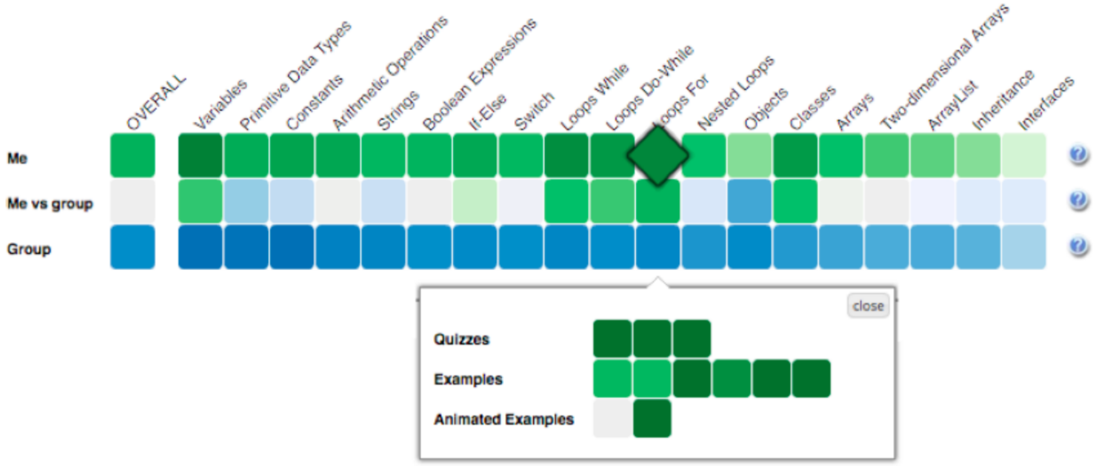


Figure 2.6: Mastery Grids interface for Java programming from [83].

2.6 Research Gaps

XAI for Educators

Khosravi et al. [65] highlight how the presence of unfairness issues has hindered the widespread use of AI in education, emphasizing the necessity for further research on Explainable AI in the educational domain. Additionally, it has been reported that instructors often possess an insufficient understanding of AI, leading to concerns regarding its utilization [64]. Consequently, there is a need for XAI in education to effectively communicate prediction results to instructors, enhance their perception of AI, and provide support in addressing the challenges they encounter while integrating AI into teaching practices. Trust to carry out without error and effectiveness, as identified by Chounta et al. [64], emerges as a critical concern for teachers.

Trust

Trust is a complex and abstract concept that has been defined in many different ways [49]. It is difficult to measure and can change over time [50, 51]. Some researchers have attempted to measure trust as a one-dimensional construct [50, 51, 52], but this approach does not capture the complexity and multi-faceted nature of trust as highlighted by Ooge et al. [46, 47, 48].

Explainable Outlier Detection

The literature highlights two gaps in the field of explainable outlier or anomaly detection (XAD). Firstly, while there is extensive literature on anomaly detection methods, the emphasis has been primarily on detection accuracy, neglecting the importance of providing explanations for the detected anomalies or outliers [55, 56, 57, 58]. Furthermore, the literature lacks any research on the application of outlier or anomaly detection algorithms in an educational context. As a result, there exists a significant gap in the area of explainable outlier detection, particularly within educational settings.

Chapter 3

Methods and Materials

The objective of this research is to address the identified gaps in the literature. Firstly, the Wiski platform will be extended to incorporate an outlier detection system that identifies “weak” and “strong” students for educators. Subsequently, explanations for these detections will be provided, considering their impact on trust and perceived accuracy. Trust will not be evaluated through a single-dimensional construct but rather through a quantitative approach that uncovers the multiple factors that influence it. Furthermore, the effectiveness of the explanations and teachers’ satisfaction will be analyzed. Finally, this research contributes to the field of explainable outlier detection in education.

Based on the research goals outlined, the following research questions are posed:

Research Question 1

Which factors affect teachers’ trust in an explainable outlier detection system? For example, how do perceived accuracy and understanding of the system affect trust?

Research Question 2

How do teachers assess model-centric and data-centric explanations in terms of effectiveness and satisfaction?

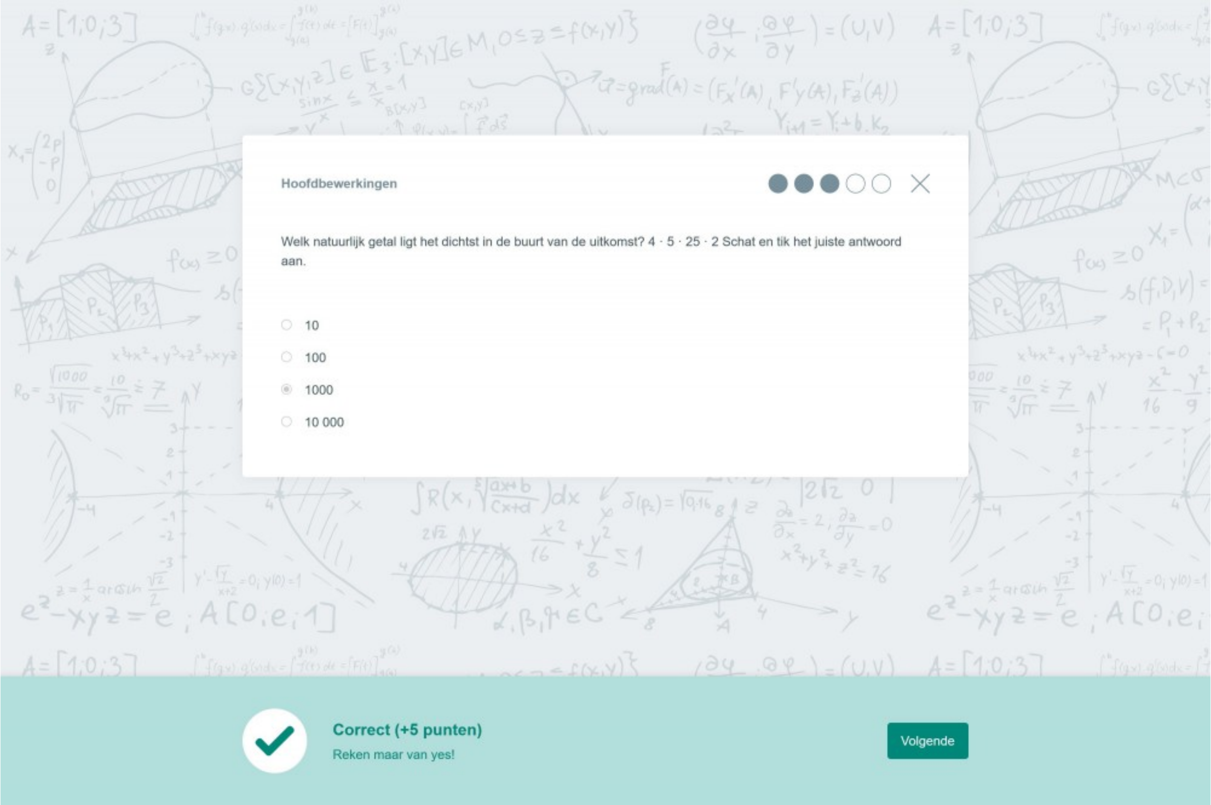
Research Question 3

How do teachers (intend to) utilize a dashboard with outlier detection in a classroom setting?

In this chapter, the methodology employed to investigate the impact of the dashboard with visual explanations on the attitudes and behavior of teachers is presented. The chapter begins by discussing the pilot study conducted to gain insights into the needs and expectations of teachers in the context of the research. The extension of Wiski is introduced and its technical implementation. A switching replications study was subsequently conducted to examine the effects of the interface and its explanations. Finally, the evaluation method employed to assess the interface is discussed. The conducted research has been approved by the Sociaal-Maatschappelijke Ethische Commissie (SMEC) (file: G-2022-6129).

3.1 Wiski: An E-learning Platform that Detects Outlying Students

The Wiski e-learning platform, developed by Ooge [43], offers students a range of multiple-choice math questions provided by publisher die Keure (see **??**–3.2). The Wiski platform was extended and modified to include new functionalities for teachers. During the study, this could be accessed at www.anissa.wiski.be.



Hoofdbewerkingen ●●●○○ ×

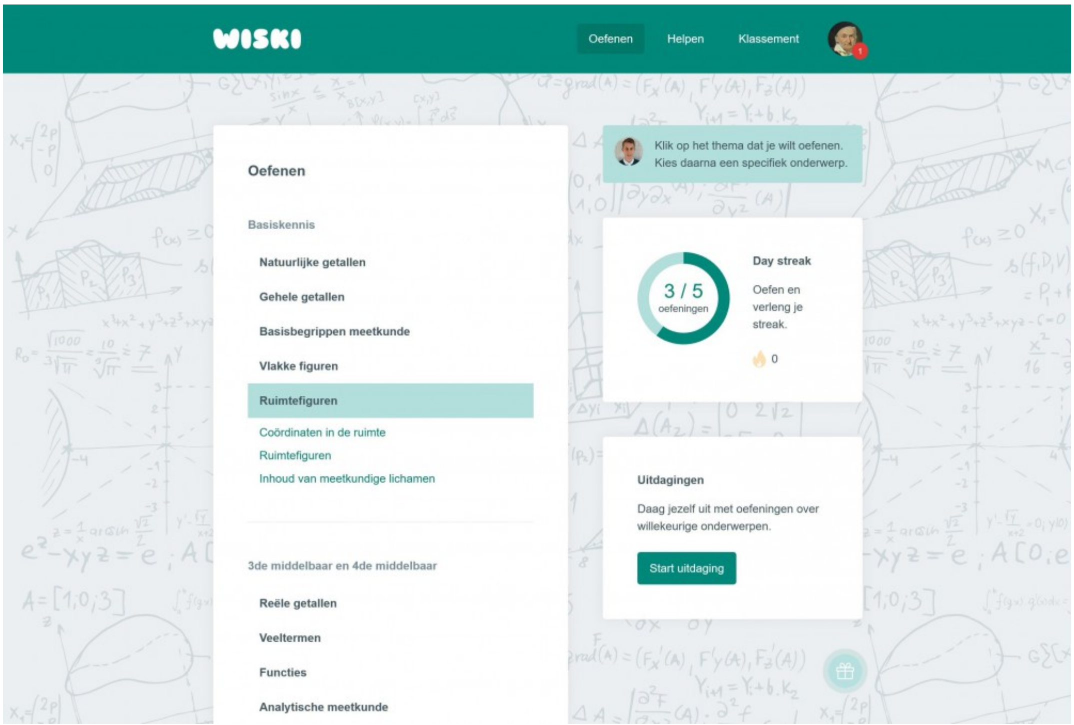
Welk natuurlijk getal ligt het dichtst in de buurt van de uitkomst? 4 · 5 · 25 · 2 Schat en tik het juiste antwoord aan.

○ 10
○ 100
● 1000
○ 10 000

✓ Correct (+5 punten)
Reken maar van yes!

Figure 3.1: A correctly solved exercise on Wiski [43].

A first new functionality enabled teachers to put together a set of exercises for a particular subject of their choice. Teachers had the option of either randomly generating a selection of exercises from a particular subject (Figure 3.3) or handpicking specific exercises to add to their set (Figures 3.4 to 3.5). A second functionality was an overview of their students (Section 3.1). An explainable outlier detection system was integrated that can detect students based on their performance. The overview indicates which students had been detected by the outlier detection algorithm for a particular set of exercises. Teachers could also filter based on specific exercises within the set and display only those students who have been detected by the algorithm. Think-aloud sessions were conducted for each prototype, where participants performed tasks while verbalizing their thoughts, exposing their thinking process, and highlighting encountered difficulties. This user-centered and



Klik op het thema dat je wilt oefenen. Kies daarna een specifiek onderwerp.

Oefen en verleng je streak.

Daag jezelf uit met oefeningen over willekeurige onderwerpen.

Figure 3.2: An overview of all subjects for which students can solve exercises on Wiski [43].

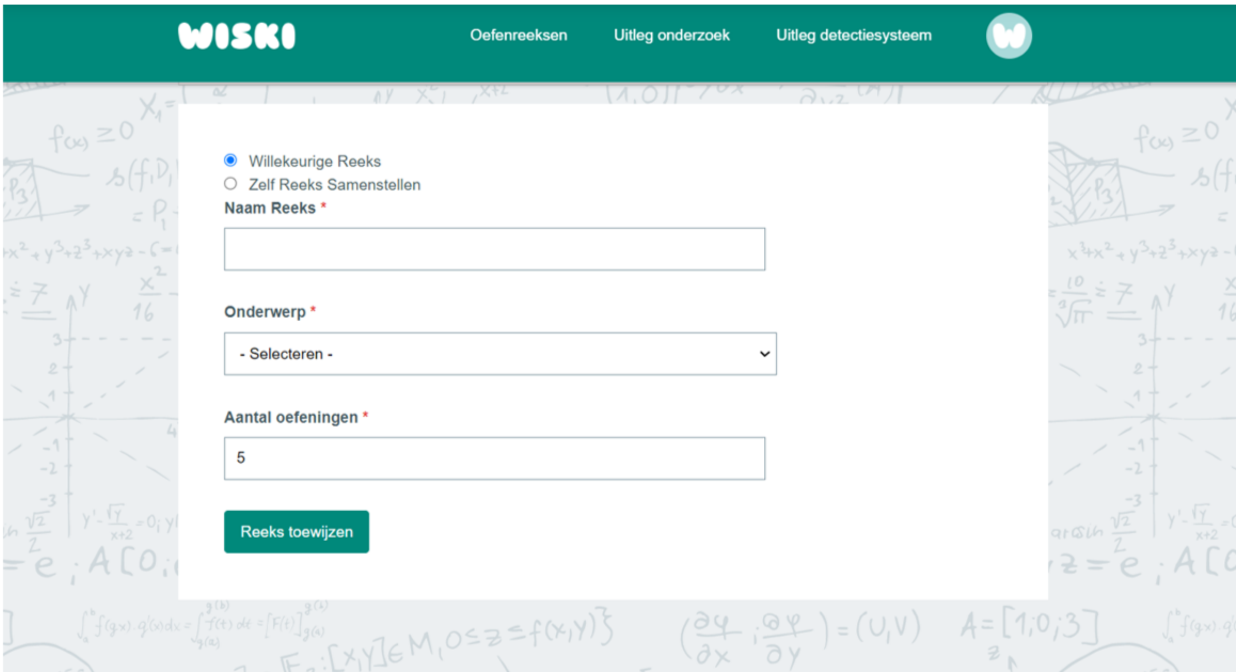
iterative approach is detailed in chapter 4. Additionally, different interventions were also provided for teachers to utilize when a student was detected by the outlier algorithm. The interventions available for students who were falling behind included being paired with a study buddy, receiving assistance from a teacher or simply receiving no intervention for the time being. In contrast, for students who were excelling, teachers have the option to assign more challenging exercises or to provide no intervention at all. Teachers actively engaged in the design process of this interface, leading to the development of low-fidelity and high-fidelity prototypes. Finally, a third functionality showcases questions that are detected by the algorithm.

3.1.1 Explanations

The extended Wiski platform equiped teachers with explanations that provided valuable insights into both the system's detections and the underlying model.

How explanations Teachers were provided with a model-centric or global explanation that delved into the workings of the outlier detection model, including the calculation of parameters such as speed and number of attempts scores. These explanations were presented in textual form, supplemented by illustrations (refer to Figures 3.8 to 3.10).

Why explanations In addition, local explanations were provided to teachers through a data-centric approach (see Section 3.1). The goal was to clarify why the model made a certain prediction for a specific student. This was done by visualizing the distribution of students' scores for speed and attempts, and where students' scores fell in the distribution



● Willekeurige Reeks
○ Zelf Reeks Samenstellen

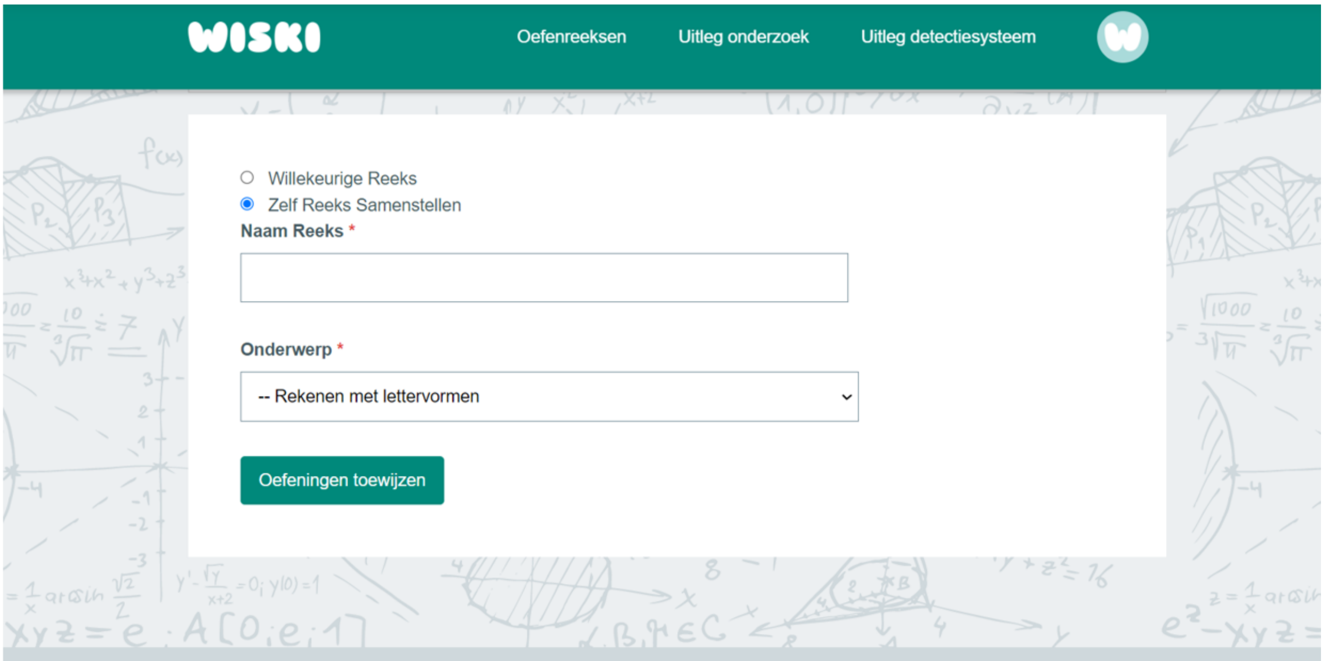
Naam Reeks *

Onderwerp *
- Selecteren -

Aantal oefeningen *
5

Reeks toewijzen

Figure 3.3: Form to generate a randomized set of exercises for a specific subject.



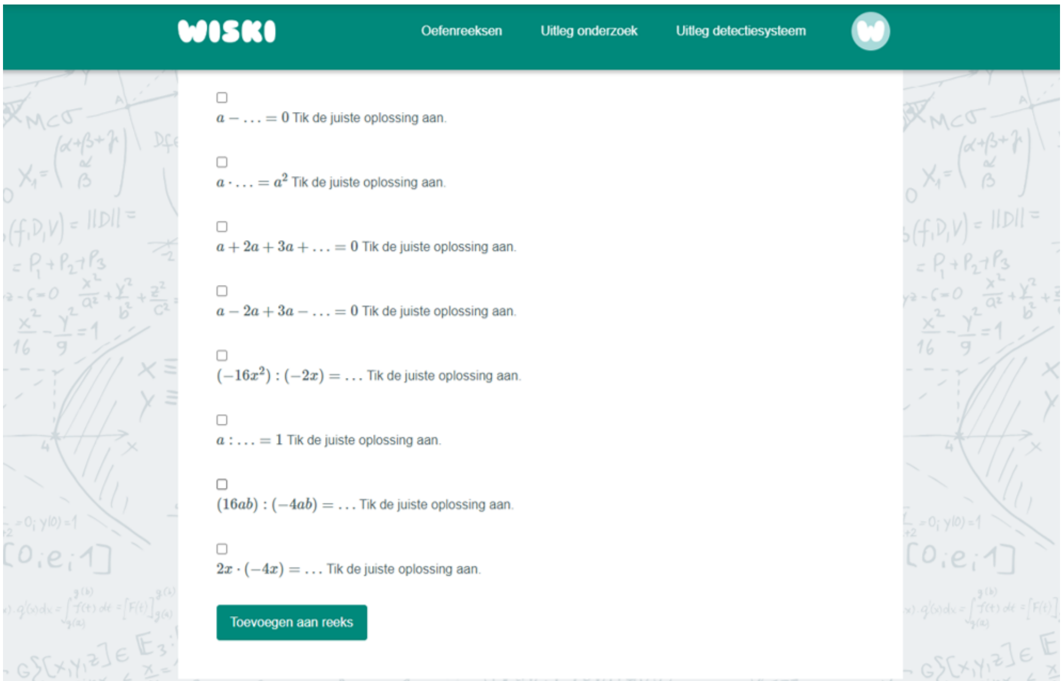
○ Willekeurige Reeks
● Zelf Reeks Samenstellen

Naam Reeks *

Onderwerp *
-- Rekenen met lettervormen

Oefeningen toewijzen

Figure 3.4: Form to manually select specific exercises to create a set for a particular subject.



□
$a - \ldots = 0$ Tik de juiste oplossing aan.

□
$a \cdot \ldots = a^2$ Tik de juiste oplossing aan.

□
$a + 2a + 3a + \ldots = 0$ Tik de juiste oplossing aan.

□
$a - 2a + 3a - \ldots = 0$ Tik de juiste oplossing aan.

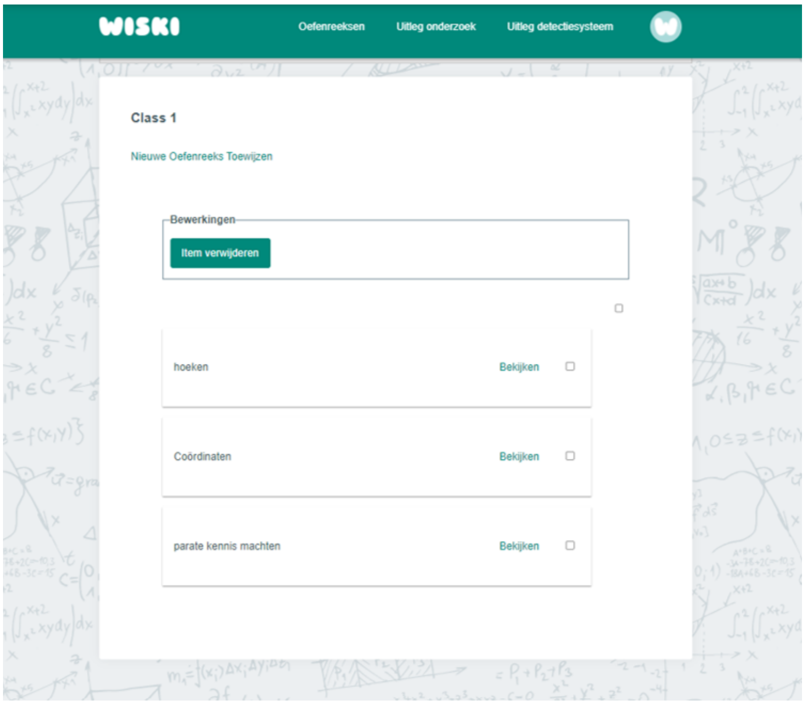
□
$(-16x^2) : (-2x) = \ldots$ Tik de juiste oplossing aan.

□
$a : \ldots = 1$ Tik de juiste oplossing aan.

□
$(16ab) : (-4ab) = \ldots$ Tik de juiste oplossing aan.

□
$2x \cdot (-4x) = \ldots$ Tik de juiste oplossing aan.

Figure 3.5: Overview of all exercises for a particular subject.



Class 1

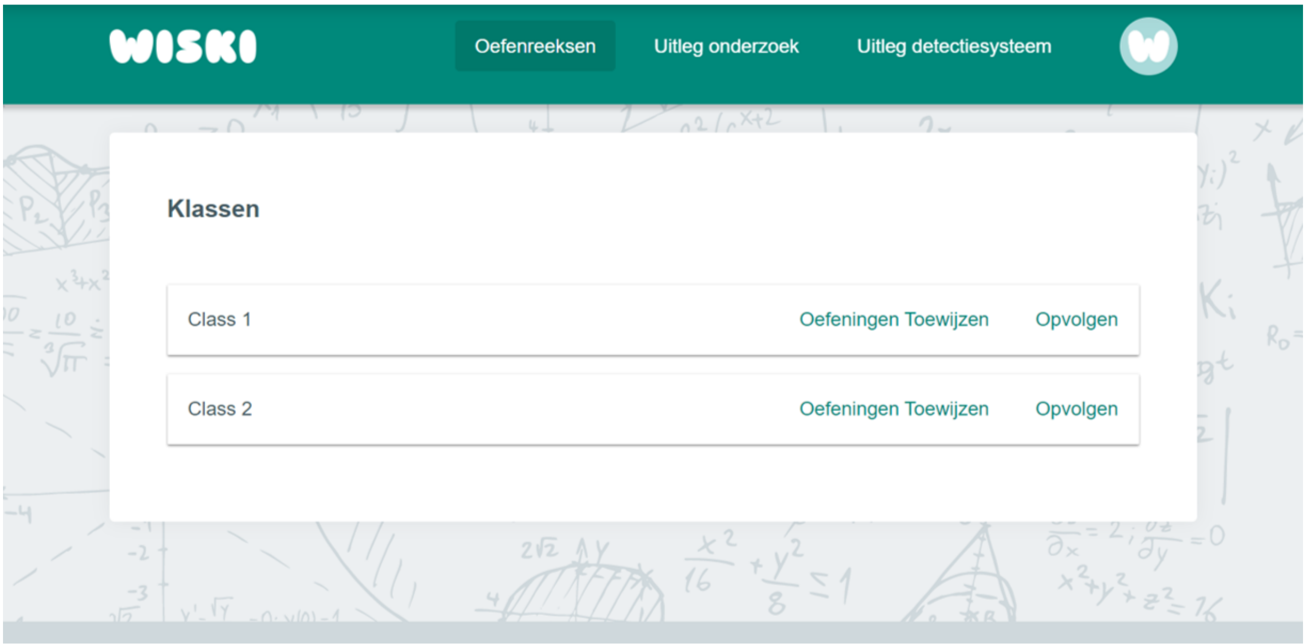
□

hoeken Bekijken □

Coördinaten Bekijken □

parate kennis machten Bekijken □

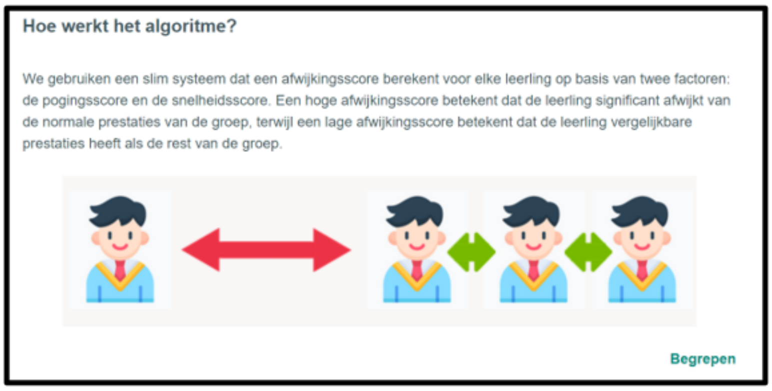
Figure 3.6: Overview of exercise sets that a teacher has prepared for a class.



Klassen

Class 1	Oefeningen Toewijzen	Opvolgen
Class 2	Oefeningen Toewijzen	Opvolgen

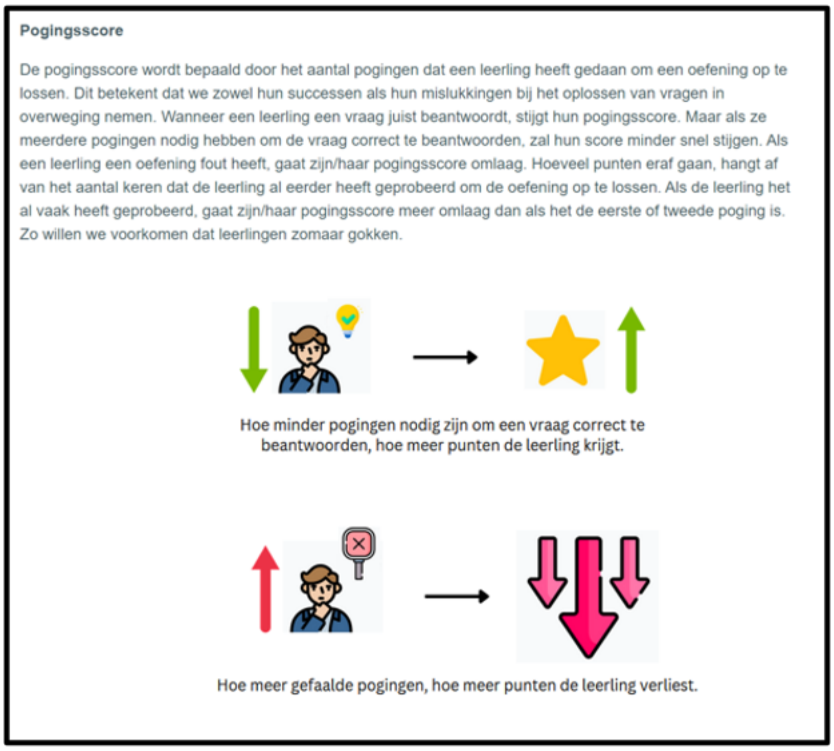
Figure 3.7: Overview of all classes of a teacher.



Hoe werkt het algoritme?

We gebruiken een slim systeem dat een afwijkingsscore berekent voor elke leerling op basis van twee factoren: de pogingsscore en de snelheidsscore. Een hoge afwijkingsscore betekent dat de leerling significant afwijkt van de normale prestaties van de groep, terwijl een lage afwijkingsscore betekent dat de leerling vergelijkbare prestaties heeft als de rest van de groep.

Figure 3.8: Explanation on outlier detection algorithm (English translation in D.1).



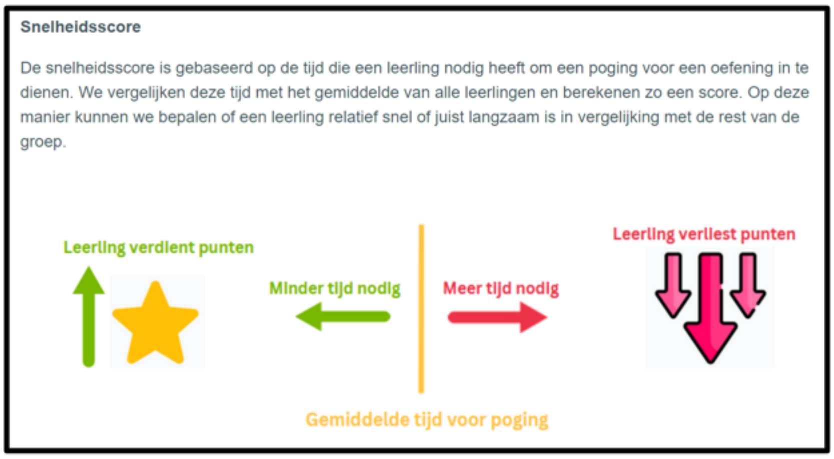
Pogingsscore

De pogingsscore wordt bepaald door het aantal pogingen dat een leerling heeft gedaan om een oefening op te lossen. Dit betekent dat we zowel hun successen als hun mislukkingen bij het oplossen van vragen in overweging nemen. Wanneer een leerling een vraag juist beantwoordt, stijgt hun pogingsscore. Maar als ze meerdere pogingen nodig hebben om de vraag correct te beantwoorden, zal hun score minder snel stijgen. Als een leerling een oefening fout heeft, gaat zijn/haar pogingsscore omlaag. Hoeveel punten eraf gaan, hangt af van het aantal keren dat de leerling al eerder heeft geprobeerd om de oefening op te lossen. Als de leerling het al vaak heeft geprobeerd, gaat zijn/haar pogingsscore meer omlaag dan als het de eerste of tweede poging is. Zo willen we voorkomen dat leerlingen zomaar gokken.

Hoe minder pogingen nodig zijn om een vraag correct te beantwoorden, hoe meer punten de leerling krijgt.

Hoe meer gefaalde pogingen, hoe meer punten de leerling verliest.

Figure 3.9: Explanation on attempt score (English translation in D.1).



Snelheidsscore

De snelheidsscore is gebaseerd op de tijd die een leerling nodig heeft om een poging voor een oefening in te dienen. We vergelijken deze tijd met het gemiddelde van alle leerlingen en berekenen zo een score. Op deze manier kunnen we bepalen of een leerling relatief snel of juist langzaam is in vergelijking met de rest van de groep.

Figure 3.10: Explanation on speed score (English translation in D.1).

of all students. In addition, the successful and failed attempts of students for the exercises in the sequence were also visualized.

3.1.2 Detecting Outlying Students with LOF

To detect outlying students, we used the LOF algorithm with k set to half of the total number of students in the class and a threshold value of 1.0 for the LOF-score to classify data points as outliers [40, 45].

Two scores were kept for each student to analyze their performance. The first score was based on the student's speed in completing exercises, while the second score was based on the number of attempts. These two scores were used as data for the LOF outlier detection algorithm. Students who were detected and had both scores above the average were classified as "strong" students, while those with scores below the average were classified as "weak" students.

Speed Score
The speed score was based on the time it takes for a student to submit an attempt for an exercise. This time was compared to the average time needed by all students and calculated a score. This way, it was determined if a student is relatively fast or slow compared to the rest of the group. If their speed for an attempt was equal to or below average then the student earned a point. Otherwise, a point was subtracted.

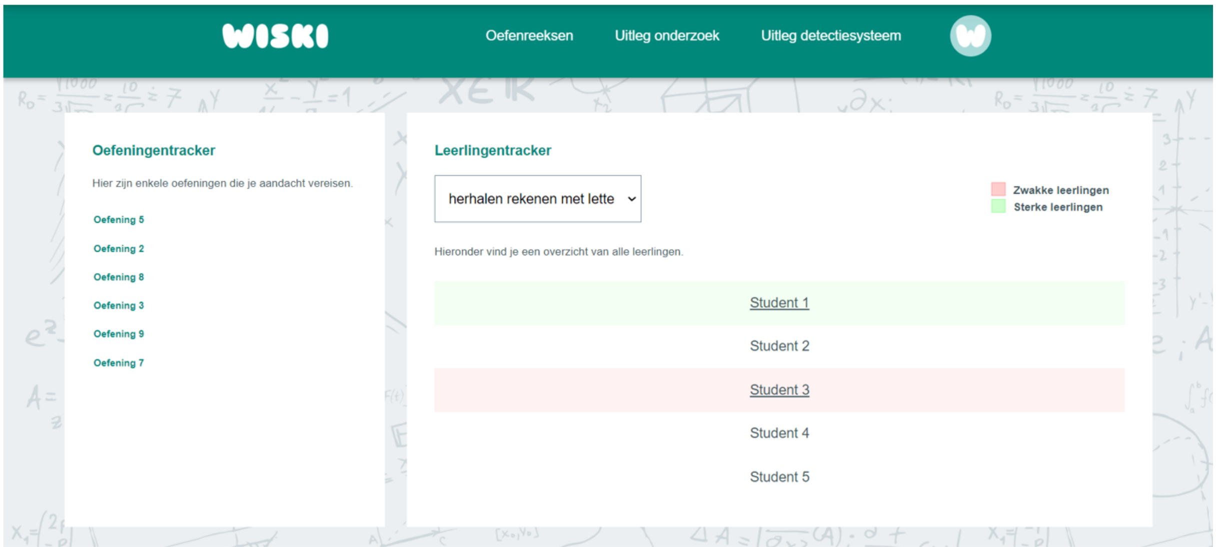
Attempt Score
The following formula was utilized to update the attempt score:

$$\text{new attempt score} = \begin{cases} \text{old attempt score} + \frac{1}{\#\text{ attempts}}, & \text{if the attempt is successful} \\ \text{old attempt score} - \left(1 - \frac{1}{\#\text{ attempts}+1}\right), & \text{if the attempt is unsuccessful} \end{cases} \tag{3.1}$$

The attempt score was determined by the number of attempts a student made to solve an exercise. This means that both their successes and failures in answering questions were taken into account. When a student answered a question correctly, their attempt score increased. However, if they needed multiple attempts to answer the question correctly, their score would increase less quickly. If a student answered an exercise incorrectly, their attempt score decreased. The amount of points deducted depends on the number of times the student has attempted to solve the exercise before. If the student had attempted the exercise many times before, their attempt score would decrease more than if it is their first or second attempt. The purpose of this was to give greater weight to guessing and its impact on the student's score.

3.1.3 Determing the Proficiency Level of Students with Elo

A multivariate Elo rating system was used to measure a student's proficiency level for a specific subject and the difficulty level of an exercise. It assigned an Elo rating to each student for each subject. In this case, it is used to provide teachers with the ability to assign more challenging exercises to students who were advancing quickly. Since the Elo ratings are relative, they can be chosen arbitrarily. In this case, the initial score for students was set to 1000. The initial score for questions was set to 1100. The K-parameter



Oefeningentracker

Hier zijn enkele oefeningen die je aandacht vereisen.

- Oefening 5
- Oefening 2
- Oefening 8
- Oefening 3
- Oefening 9
- Oefening 7

Leerlingentracker

herhalen rekenen met lette ⌄

🟥 Zwakke leerlingen
🟩 Sterke leerlingen

Hieronder vind je een overzicht van alle leerlingen.

- Student 1
- Student 2
- Student 3
- Student 4
- Student 5

Figure 3.11: Our baseline dashboard, which only offered an overview of students and outliers without any data-centric explanations.

is important as it decides the extent of a student's Elo score change. We used a K-value of 0.4, similar to previous research [21, 23, 24]. Note that the focus of this thesis was not to optimize these parameters, but to evaluate the explainable outlier detection system.

3.2 Switching Replications Study

To investigate the impact of our dashboard with visual explanations on teachers' trust and perceived accuracy, and their satisfaction with and effectiveness of the explanations, a switching replications study was conducted, which is a counterbalanced within-subjects study [42]. For this, an additional interface was developed as a baseline for the research (Figure 3.11). We used a switching replication design study due to the small number of participants. This design allowed us to compare the effect of explanations within the same group of teachers. This interface offered a view of student information and detected questions but lacked explanatory features. It allowed filtering based on detected students and provided the same intervention options as the explanation interface.

Figures 3.12 to 3.13 show how our study randomly assigned participants to either Group A or Group B. Both groups received two dashboards, but in different orders: Group A received the dashboard with explanations for the outlier detections for 25 minutes, followed by the dashboard without explanations for 15 minutes; Group B received the dashboards in the opposite order.

3.2.1 Recruiting Participants

A range of recruitment channels were utilized to find participants for the study. Initially, 12 schools across Flanders were contacted through e-mail. However, only one school expressed interest and agreed to relay information about the study to its teachers. Unfortunately, no teachers from that school responded. In addition, Facebook was also used to

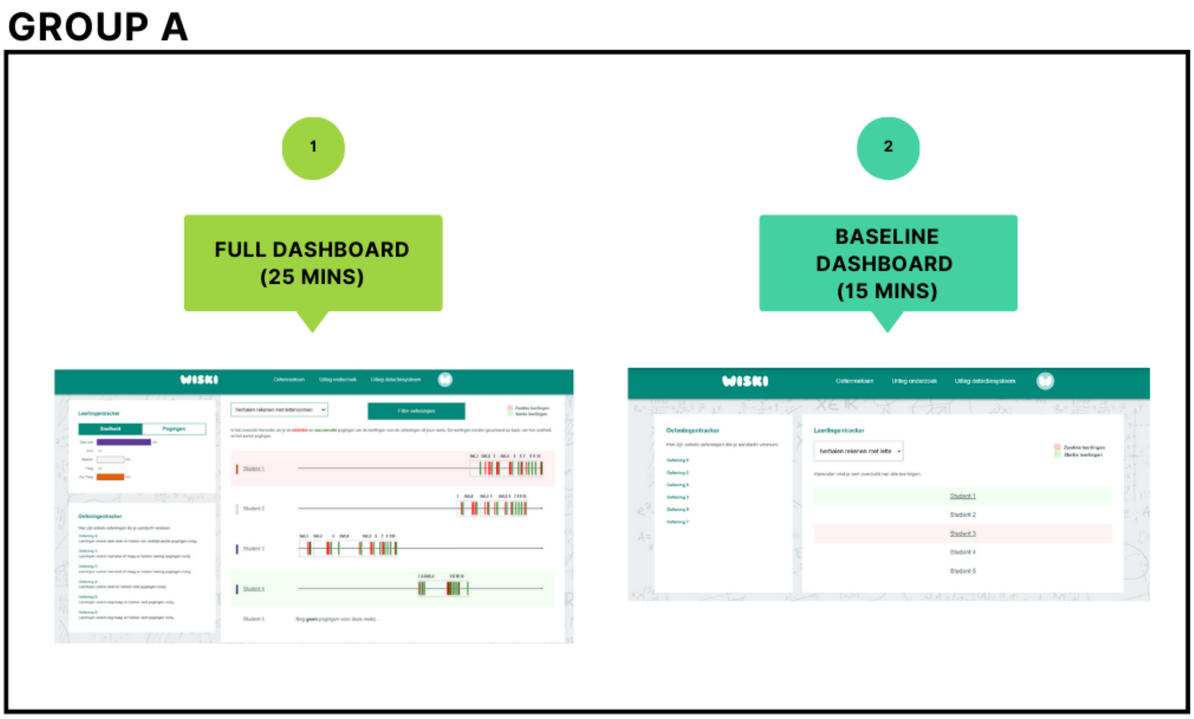


Figure 3.12: Participants in Group A first saw our dashboard with data-centric explanation and then our baseline dashboard.

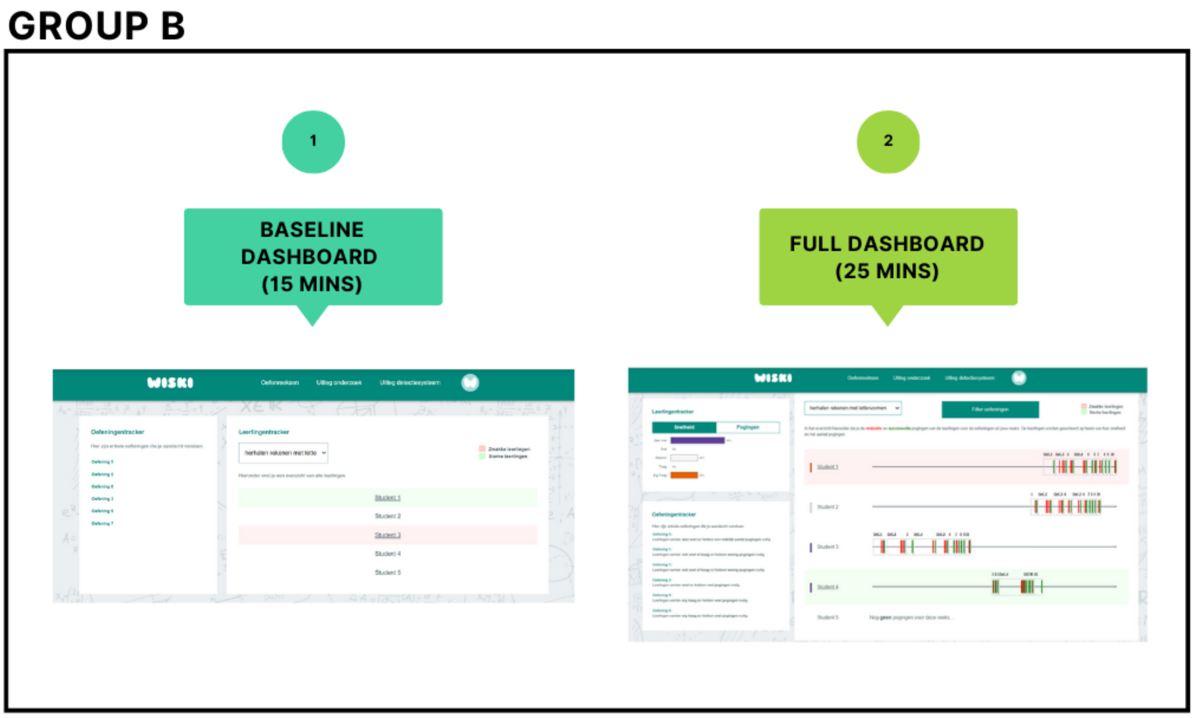


Figure 3.13: Participants in Group B first saw our baseline dashboard and then our dashboard with data-centric explanation.

contact teachers through groups targeting math teachers, education, and math studies. Although 29 teachers initially showed interest, only 8 ultimately participated. LinkedIn was also used to contact potential participants. Four teachers were contacted but only one responded and they were unable to participate. Additionally, two math teachers were contacted through lecturers of the educational bachelor program at UCLL. Unfortunately, only one responded and was unable to participate. Four other teachers were contacted through a personal connection with a teacher. Two of those teachers ultimately participated in the study. Lastly, three teachers were contacted from the author's workplace, but only one was able to participate. Despite these challenges, a small group of 11 participants were recruited to contribute to the study.

3.2.2 Flow of Study

We evaluated our dashboard in a real-life class setting with real students. However, since our study was focused on teachers' perceptions, we did not collect any data on students' behavior. Concretely, Figure 3.14 shows that our study involved 3 phases.

Pre-study

In the pre-study phase, interested teachers were informed about the study, registered on Wiski, and had to verify their teacher status. To inform participants about the study, a brochure outlined the objective of the research, the process involved, and what would be done with the data collected (see Appendix A.1). It was clearly stated that their participation was voluntary and they were free to end their involvement at any time. Moreover, teachers were informed that they should not use the information obtained from the dashboard to evaluate the students in any way. Finally, before participating in the study, teachers were required to sign an informed consent form (see Appendix A.2). Once registered and verified, the teachers were then asked to read the explanation of the detection system. This provided model-centric explanations on the outlier detection system and its parameters. Afterward, the teachers prepared a set(s) of exercises for their students for a particular subject either randomly (Figure 3.3) or by picking out exercises themselves (Figure 3.4).

Study

In the second phase, teachers conducted the study in their classes. They were silently assigned to either group A or B. Students registered and had access to the set(s) of exercises prepared by their teacher. Teachers then pressed the start button to initiate the study. From that point on, teachers were able to see the dashboard in the order according to their group. Additionally, teachers were able to put in interventions through the platform for students who are detected as struggling or advancing. A prompt was included when teachers put in an intervention, asking whether they agreed with the system's detection.

Many user interactions were logged. This included interventions and the agreement or disagreement of teachers with each intervention; the duration of students' detection period; hovering and clicking sets of exercises, applying filters to exercises or weak/strong students, interacting with detected exercises, student names, histogram bars, attempts, and exercise names on the timeline.

Post-study
In the post-study phase, teachers were asked to e-mail their first impressions of the dashboards right after they ended the study. Additionally, an interview was planned to discuss the dashboards in more detail.

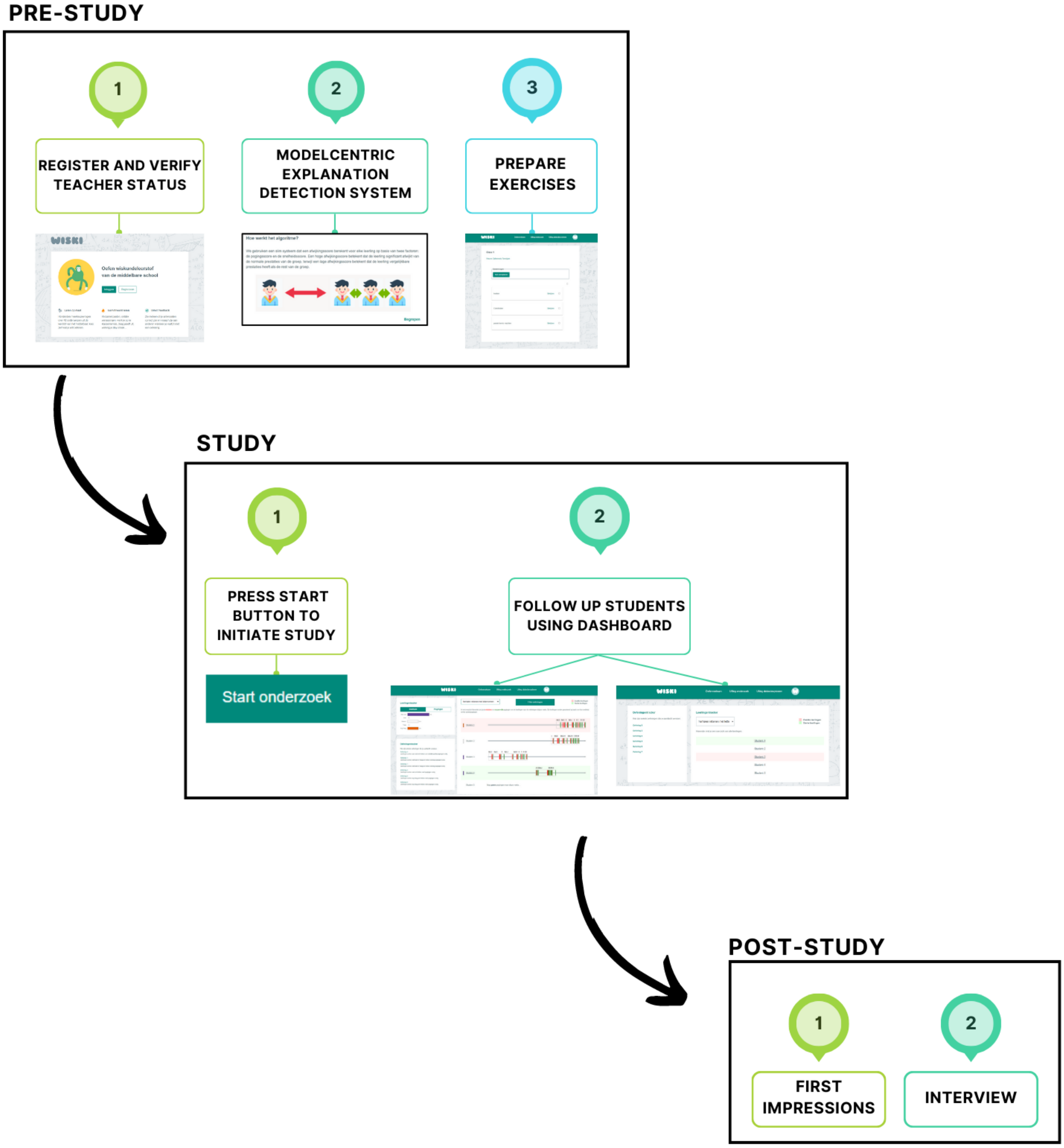

PRE-STUDY
1
REGISTER AND VERIFY TEACHER STATUS
2
MODELCENTRIC EXPLANATION DETECTION SYSTEM
3
PREPARE EXERCISES
STUDY
1
PRESS START BUTTON TO INITIATE STUDY
Start onderzoek
2
FOLLOW UP STUDENTS USING DASHBOARD
POST-STUDY
1
FIRST IMPRESSIONS
2
INTERVIEW

Figure 3.14: An overview of the process of the study.

3.3 Semi-Structured Interviews

Interviews were conducted with participants following the switching replications design study to gather their feedback and evaluation of the concepts related to the research questions. The questions of the interview can be found in appendix E.1. The responses were thematically analyzed according to guidelines of Braun and Clarke [84].

General questions

The interview began with general questions aimed at obtaining an overview of the participant's experiences with the system. This included their first impressions of the detection system and how they used it during the practice session. Participants were also asked whether they received different feedback from students compared to a typical lesson. Additionally, they were questioned about the insights they found most interesting or missing and the ones that encouraged them to intervene. The goal of these general questions was to provide participants with an opportunity to get comfortable and potentially mention topics related to trust, perceived accuracy, effectiveness, or satisfaction without prompting.

Trust

Questions on trust were based on constructs from the Madsen-Gregor trust scale [49] that were relevant to the research. To start the interview, a general question on the trustworthiness of the system was asked to prompt further conversation. In addition, constructs such as perceived reliability, perceived understandability, and faith were explored to gain a deeper understanding of trust beyond a one-dimensional construct or question.

Perceived accuracy

To gain a deeper understanding of participants' perceptions of the system, the interview also included questions on perceived accuracy. Participants were asked about how they viewed the accuracy of the system's detections and the extent to which they relied on the system's accuracy.

Satisfaction

To measure the satisfaction of the teachers with the provided explanations, a set of questions based on the explanation satisfaction scale developed by Hoffman et al. [53] was included in the interview. Concretely, we asked questions on understandability, feeling of satisfaction, sufficiency of detail, and usefulness were used to query the satisfaction with the explanations provided.

Effectiveness

To assess the effectiveness of explanations provided for the detections made by the system, participants were asked if the explanations helped them effectively utilize the detection system for intervention purposes. Considering that some teachers might prefer to intervene without relying on the platform's assistance, the qualitative analysis helped gain insight into the effectiveness of the explanations.

General Follow-up Questions

To gather further insights into participants' experiences, two general follow-up questions are included in the interview. The first question was aimed at verifying whether participants have read the model-centric explanation on the outlier detection system and if it helped them to understand the system better. The second question was a more general question asking participants for their overall impressions of the dashboard. This was included to potentially uncover any additional feedback that was not covered in the previous questions.

Chapter 4

Development of the Proof of Concept

This chapter discusses the user-centered development process of Wiski will in Figure 4.1. To initiate the development of the extention of Wiski, a pilot study was conducted with a specific aim of understanding the needs of teachers. Next, an iterative process was implemented to develop the final proof of concept. The first step was the creation of a paper prototype that was evaluated by two Phd student through a co-design session and followed by a Think-aloud study with seven teachers. The feedback from the first think-aloud study was used to develop the digital prototype. This was then evaluated by five education experts through a second think-aloud study. The feedback collected from this evaluation was used to further refine the prototype until a final proof of concept was developed.

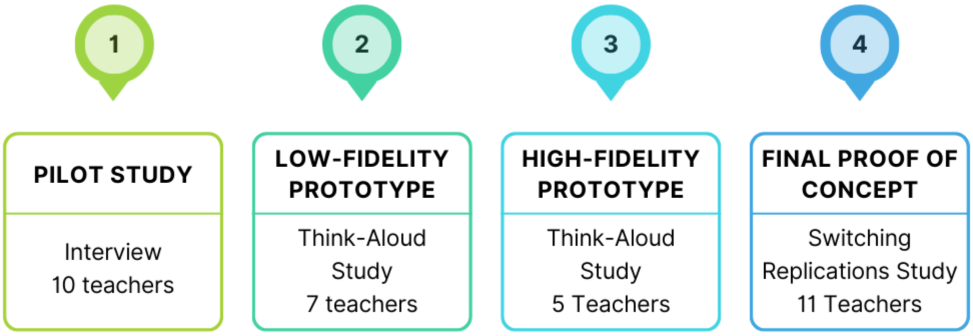


Figure 4.1: Development process of the wiski learning platform: from pilot study to final proof of concept.

4.1 Pilot Study

A pilot study was conducted for this research using a semi-structured interview approach. The goal of this study was to gain insights into the needs and expectations of teachers regarding the integration of an explainable artificial intelligence system into an e-learning platform. To ensure that the semi-structured interviews were conducted consistently and effectively, the guidelines provided by Adams on conducting semi-structured interviews [44] were followed closely throughout the pilot study.

Ten participants were recruited for the pilot study, consisting of teachers from various disciplines. The diversity of the participants allowed for a broad range of perspectives and experiences to be considered. Participants were asked about their use of digital platforms and learning environments (cfr. appendix B.1). More specifically, they were asked whether they utilize these platforms to gauge whether a student requires additional support, how they discern if a student is lagging or advancing, and what actions they take in such circumstances. Additionally, five distinct dashboards were presented to the participants to help them identify the most fitting visualizations for the dashboard displaying the students (see Appendix B.2).

4.1.1 Participant Perspectives on Digital Platforms for Student Support

Based on the responses provided by the participants, a variety of digital platforms and learning environments are being used in their classes. Google Classroom, Teams, Bookwidget, and Diddit were among the most commonly mentioned tools. Additionally, some teachers reported using specific digital platforms offered by textbooks, while others mentioned using a range of tools such as Kahoot and Quizlet.

When asked whether they use digital tools to assess whether a student requires more support, the majority of the participants responded negatively. However, one teacher reported using a platform that provided a dashboard to track student progress. Bookwidget was also mentioned by some teachers as a tool to monitor completed exercises, while another teacher mentioned using it to track activities. Some teachers expressed reluctance to use these tools due to inadequate feedback, with one teacher stating that the points are often incorrectly calculated. On the other hand, one teacher reported using a tool to test prior knowledge.

Teachers used various methods to assess students' progress in the classroom. Some relied on objective measures such as in-class exercises, self-tests, quizzes, tests, and tasks. Others considered students' behavior and study habits, the time required to complete tasks, and their own experience as a teacher. Some teachers also considered the level of assistance required.

Participants used a variety of interventions to help students who are either struggling or excelling in class. Remedial lessons and extra exercises were commonly implemented to provide additional support. Some teachers utilized online resources to assist students in developing specific skills. Others gave more challenging exercises or allowed strong students to serve as tutors. The strategy of within-class differentiation was also employed, with students given more challenging work based on their skill level.

4.1.2 Participant Assessment of Dashboard Visualizations for Student Monitoring

Teachers were shown five different dashboards and asked to provide feedback on each one (see Table 4.1.2). Specific questions were asked regarding the insights provided by each dashboard, which insights may be missing, liked or disliked aspects, and any additional suggestions for improvement.

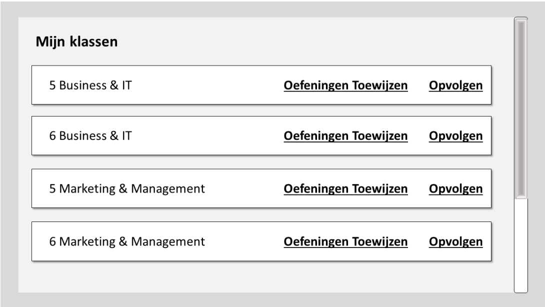
The results of the second part of the pilot study indicated that several crucial factors must be considered when developing a dashboard for teachers. All participants expressed that excessive information or visualizations could be overwhelming to analyze during a lesson. One participant described an elaborate dashboard as “clear as mud”, meaning that it does not provide the necessary insight for teachers to support their students. A table or overview visualizing completed exercises emerged as the most informative and preferred visualization for teachers, providing valuable insight into students’ progress.

In addition, participants stated that comparing a student to the class average or their peers is not preferable. They highlighted the importance of focusing on the individual learning process instead. Moreover, teachers identified the time that a student needs to complete an exercise as a critical indicator to be included.

Several participants emphasized the importance of providing personalized support to individual students and tailoring their learning process accordingly. They suggested that indications or suggested actions to take such as encouragement or investigation would be helpful, as well as implementing a warning or notification system for underperforming students. The use of clear color coding such as green, red, and yellow was helpful in quickly identifying students who need help.

4.2 Low-Fidelity Prototype

Figures 4.2 to 4.11 show the low-fidelity prototype, which is essentially a rough sketch of the initial design ideas in PowerPoint. The dashboard in Figure 4.7 displays information to the left and right of the student’s name. The left side shows a distribution graph with the student’s performance compared to the rest of the class. The right side displays a timeline of all the activities related to the exercises such as the exercise start, failed attempts, and successful completion. Detection of poor performance was indicated next to the student’s name with a red alert symbol, whereas students detected by the outlier algorithm with good performance were represented by a green race car icon. In the timeline, a blue, red, and green circle represented the start of an exercise, a failed attempt, and exercise completion, respectively.



Mijn klassen

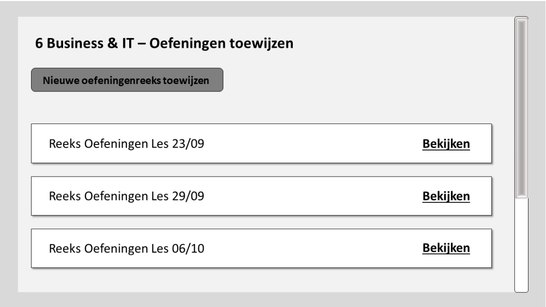
5 Business & IT — Oefeningen Toewijzen — Opvolgen

6 Business & IT — Oefeningen Toewijzen — Opvolgen

5 Marketing & Management — Oefeningen Toewijzen — Opvolgen

6 Marketing & Management — Oefeningen Toewijzen — Opvolgen

Figure 4.2: Low-fidelity prototype: Overview of the teacher’s classes.



6 Business & IT – Oefeningen toewijzen

Nieuwe oefeningenreeks toewijzen

Reeks Oefeningen Les 23/09 — Bekijken

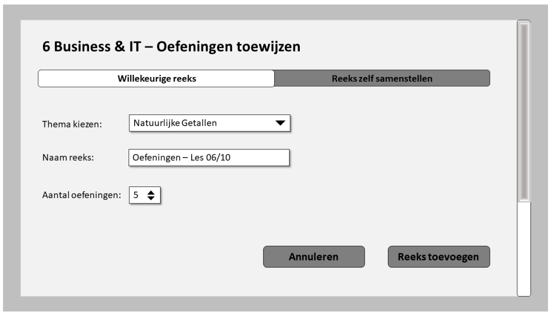
Reeks Oefeningen Les 29/09 — Bekijken

Reeks Oefeningen Les 06/10 — Bekijken

Figure 4.3: Low-fidelity prototype: Overview of all exercise sets assigned by the teacher to a class.

Table 4.1: Findings pilot study

Aspect	**Freq.**
Too much information/visualizations are overwhelming.	10/10
A table or overview showing which exercises have been completed gives insights.	10/10
No preference for comparing students to the class average.	7/10
Colors (green, red, yellow) provide clarity.	7/10
Time to solve an exercise is important in providing support.	6/10
Suggested actions to take are helpful (encouragement, investigation) helpful.	5/10
Do not compare two students with each other.	3/10
Icons that indicate a student's status (e.g., in danger, inactive) make tracking easier.	3/10
Focus on the individual learning process.	3/10
The discovery of patterns on how students achieve success or failure is not useful or clear during the teaching process.	3/10
Possibility to obtain more detailed information about individual students by clicking through.	2/10
Combination of class overview and individual overview for students.	2/10
Participation of students is visualized.	2/10
The class overview should display names.	2/10
Warning/notification system for students not performing as expected	2/10
A timeline of students' activities is useful for follow-up.	2/10



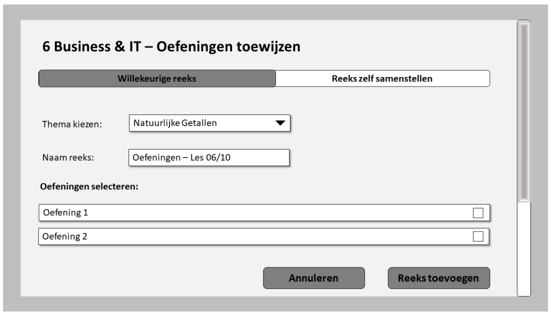
6 Business & IT – Oefeningen toewijzen

Thema kiezen: Natuurlijke Getallen

Naam reeks: Oefeningen – Les 06/10

Aantal oefeningen: 5

Figure 4.4: Low-fidelity prototype: Form for generating a random set of exercises of a subject.



6 Business & IT – Oefeningen toewijzen

Thema kiezen: Natuurlijke Getallen

Naam reeks: Oefeningen – Les 06/10

Oefeningen selecteren:

Oefening 1 □

Oefening 2 □

Figure 4.5: Low-fidelity prototype: Form for selecting exercises specific to a subject for a set.

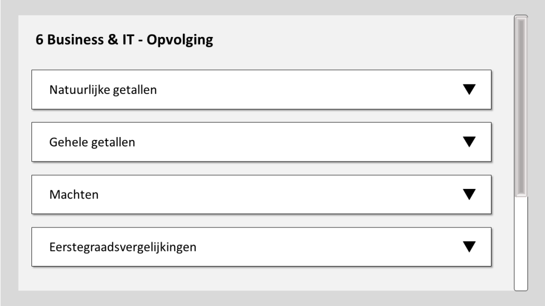
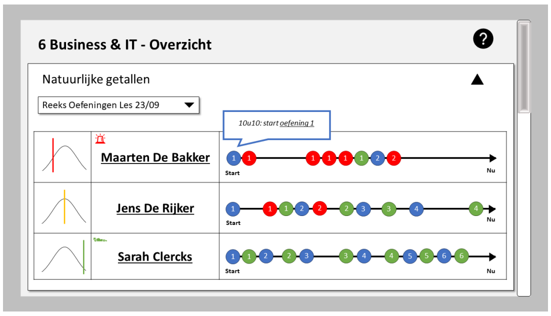
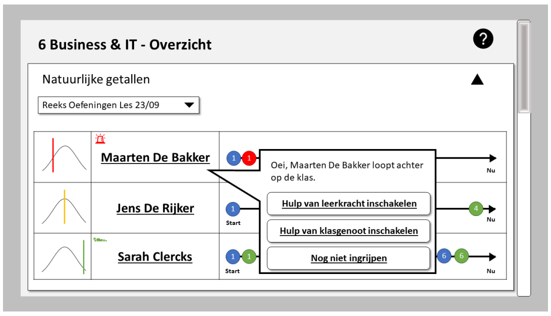


Figure 4.6: Low-fidelity prototype: All subjects for which exercise sets have been assigned by the teacher.



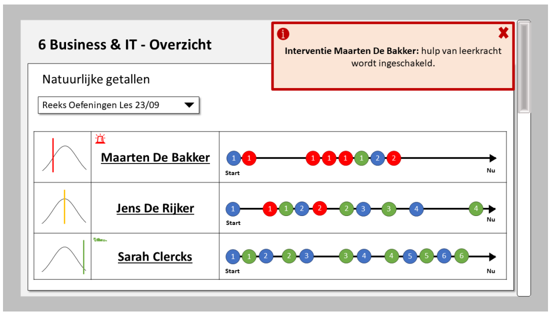
10u10: start oefening 1

Figure 4.7: Low-fidelity prototype: Hovering over the blue circle reveals the exercise start time while hovering over the red circle displays the time of the failed attempt. Similarly, hovering over the green circle reveals the time of exercise completion.


6 Business & IT - Overzicht
?
Natuurlijke getallen
Reeks Oefeningen Les 23/09
Maarten De Bakker
Jens De Rijker
Sarah Clercks
Start
Nu
Hulp van leerkracht inschakelen
Hulp van klasgenoot inschakelen
Nog niet ingrijpen

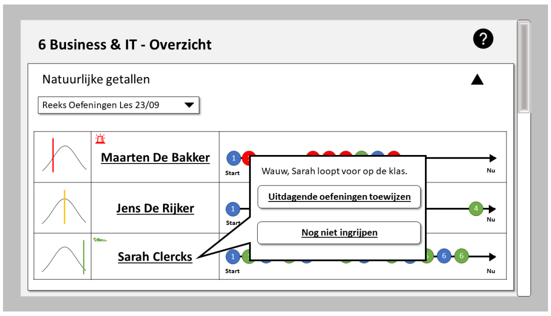
Oei, Maarten De Bakker loopt achter op de klas.

Figure 4.8: Low-fidelity prototype: Accessing intervention options by clicking on the name of a poorly performing student.



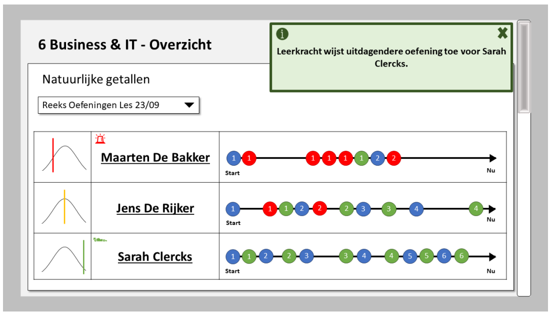
Interventie Maarten De Bakker: hulp van leerkracht wordt ingeschakeld.

Figure 4.9: Low-fidelity prototype: Teachers are notified when intervening for a "weak" student.



Wauw, Sarah loopt voor op de klas.

Figure 4.10: Low-fidelity prototype: Accessing intervention options by clicking on the name of a high-performing student.



Leerkracht wijst uitdagendere oefening toe voor Sarah Clercks.

Figure 4.11: Low-fidelity prototype: Teachers were notified when intervening for a "strong" student.

Teachers could choose a set of exercises to view visualizations for and can register interventions such as providing help, selecting a student to assist another, assigning a more challenging set of exercises, or not intervening. The platform only handled the intervention of assigning a more challenging set of exercises, while others were only provided to log their interventions for future analysis.

4.2.1 Co-Design Session with HCI Experts

Two PhD students from KU Leuven's HCI department were invited to provide their feedback on the low-fidelity prototype. The session was conducted virtually using Teams and Miro. This allowed for collaborative brainstorming and sketching as needed.

A suggestion was made to reduce clutter on the timeline by grouping exercises and eliminating the blue circles indicating the start of the exercise. They also recommended displaying a histogram for each parameter used for the outlier detection algorithm and indicating where students lay in the distribution for that parameter. Additionally, the idea of displaying which exercises students were struggling with was raised as a way of providing actionable insights for the teachers. Figure 4.12 and 4.13 show the sketches made to visualize the suggested changes for the low-fidelity prototype. Figure 4.14 and 4.15 display the updates to the low-fidelity prototype of the dashboard.

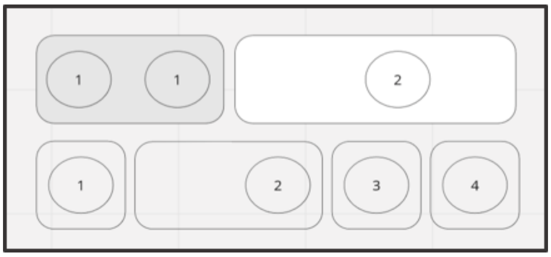


Figure 4.12: Sketch illustrating the grouping of attempts together.

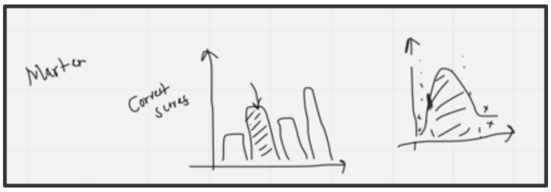


Figure 4.13: Sketch demonstrating the visualization and highlighting of the distribution of parameter values.

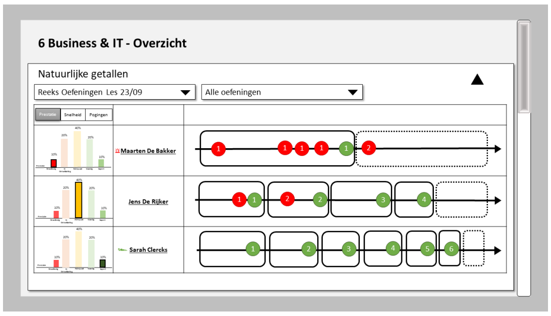


Figure 4.14: Low-fidelity prototype: Dashboard visualizations updated based on the feedback gathered during.

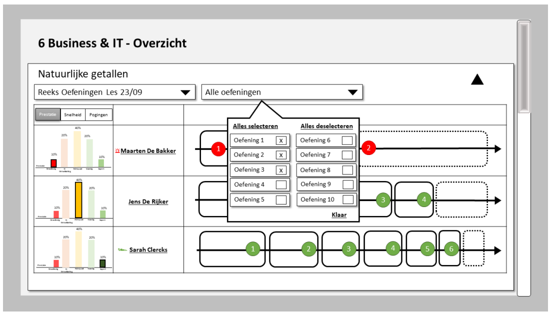


Figure 4.15: Low-fidelity prototype: Exercise filtering feature allows for displaying only the attempts made for the selected exercises.

4.2.2 Think-Aloud with Teachers

A think-aloud study was conducted with 7 teachers after the co-design session. The purpose was to evaluate the usability and navigation of the extended Wiski platform with the visualizations providing local explanations for detections. Appendix C.1 contains the questions posed during the study, which focused on the teachers' ability to create exercise sequences, navigate to the follow-up dashboard, interpret the histograms, interpret the timeline, execute interventions, and filter based on the needed information. At the end of the think-aloud session, participants were encouraged to discuss further suggestions, actionable insights, or missing aspects. Participants were also asked whether they thought the indication that a student was detected should disappear after an intervention or after some time.

Table 4.2.2 shows the results of the first think-aloud study. The study revealed that the intervention alert could stay visible as long as the student is falling behind or making significant progress ahead of the class. Teachers focused primarily on the timeline, while histograms were not initially noticed or found necessary. One solution to address these issues is to display a single histogram for a certain parameter and position a bar next to each student to indicate their position in the distribution. This approach may allow users to focus on one histogram at a time and not oversee this visualization.

Teachers suggested various functionalities and interventions, including filtering based on detected students, assigning general feedback, and allowing students to ask questions. Additionally, an intervention was recommended to pair strong students with weaker ones and to provide extra learning materials to weaker students. The study also suggested consistency in the colors of histogram distributions and the inclusion of exercise completion duration in the frame. One teacher noted that the timeline's appearance for sequences of exercises completed over several days was unclear.

The overall feedback from teachers regarding the dashboard was positive. One participant expressed, "It is very well-organized, neat, and easy to interpret. It allows for both individual student assessment and easy comparison between students." Another participant appreciated the program's feature that automatically indicates which students are ahead and which ones are lagging, stating: "This would certainly simplify our work as teachers." Regarding specific features, a teacher mentioned, "I find the timeline to be excellent. The surrounding boxes are also visually appealing. Everything becomes clear very quickly with red and green indicating attempts. I also liked the names with the accompanying icons. The ability to implement interventions such as providing additional exercises, was also very interesting."

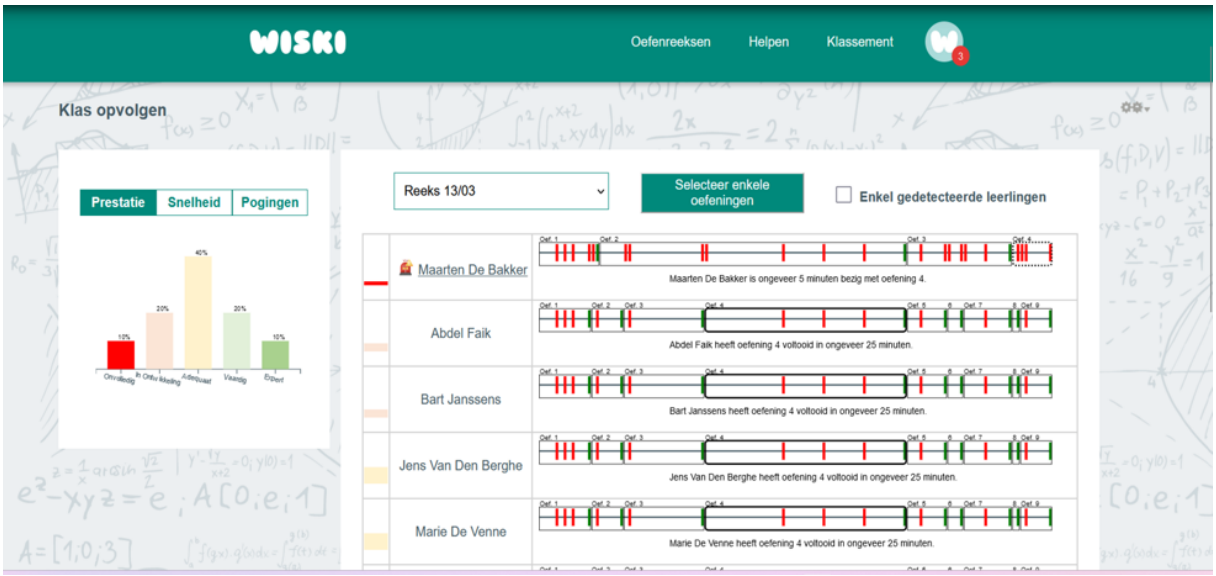
4.3 High-Fidelity Prototype

The feedback gathered from the low-fidelity prototype was incorporated into a high-fidelity version. Figures 4.16 to 4.19 show the evolution of the interface. Instead of displaying a separate histogram next to each student's name, the decision was made to show a single histogram. A colored bar was placed next to each student's name, indicating their group within the displayed distribution on the histogram. This approach aimed to capture their attention more effectively. Additionally, the histogram featured a hovering functionality,

Table 4.2: Feedback and suggestions from teachers on the low-fidelity prototype during the think-aloud study.

Issue	**Freq.**
The intervention alert may remain as long as the student is lagging or progressing quickly ahead of the class.	6/7
The focus of the teacher is directly on the timeline, histograms are not noticed (initially).	4/7
The teacher finds the timeline to be sufficient, bar charts are not needed.	2/7
A functionality was suggested by the teacher: filtering based on detected students.	1/7
A functionality was suggested by the teacher: assign general feedback to students.	1/7
A functionality was suggested by the teacher: students can ask questions.	1/7
A functionality was suggested by the teacher: view solution attempts.	1/7
An intervention was suggested by the teacher: assign strong students to weaker ones.	1/7
An intervention was suggested for weak students by the teacher: consult additional learning materials.	1/7
Consistency in colors of distributions of histograms.	1/7
Duration for exercise completion to be included in the frame.	1/7
Unclear how the timeline would look for sequences of exercises that are completed over multiple days	1/7

allowing users to see only the students within a specific group when hovering over a bar. A filtering option was also implemented to display only the detected students.



Maarten De Bakker is ongeveer 5 minuten bezig met oefening 4.

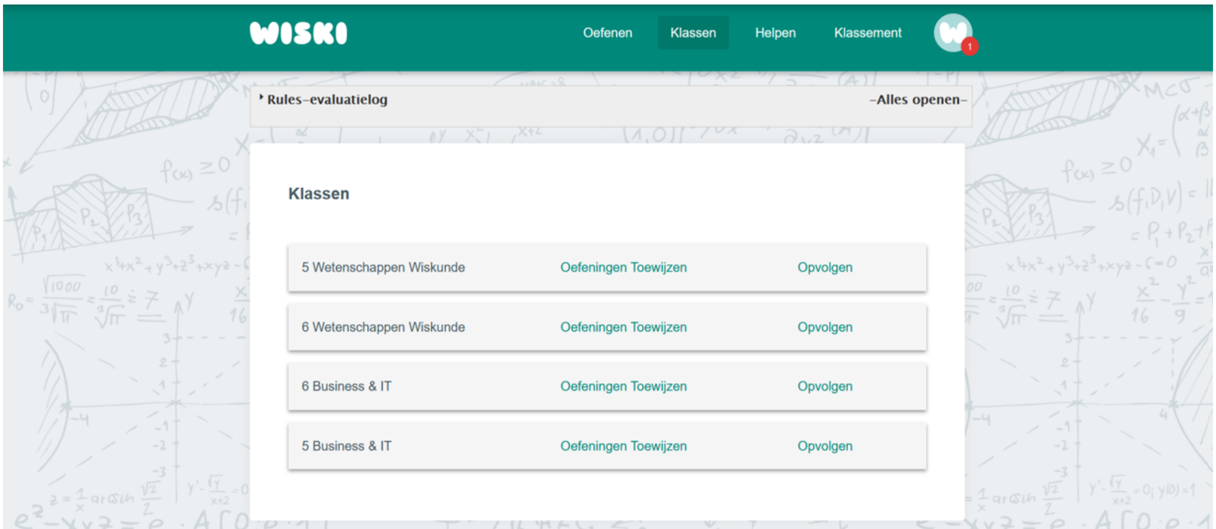
Abdel Faik heeft oefening 4 voltooid in ongeveer 25 minuten.

Bart Janssens heeft oefening 4 voltooid in ongeveer 25 minuten.

Jens Van Den Berghe heeft oefening 4 voltooid in ongeveer 25 minuten.

Marie De Venne heeft oefening 4 voltooid in ongeveer 25 minuten.

Figure 4.16: High-fidelity prototype: Dashboard of the high-fidelity prototype, which included visual explanations for why students are identified as outliers.

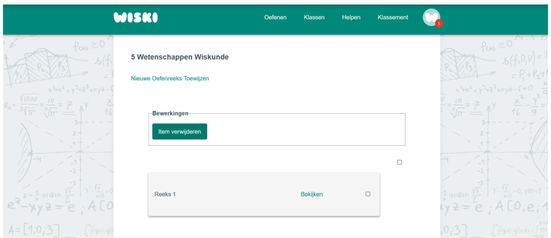


5 Wetenschappen Wiskunde	Oefeningen Toewijzen	Opvolgen
6 Wetenschappen Wiskunde	Oefeningen Toewijzen	Opvolgen
6 Business & IT	Oefeningen Toewijzen	Opvolgen
5 Business & IT	Oefeningen Toewijzen	Opvolgen

Figure 4.17: High-fidelity prototype: An overview of all classes of a teacher

For the timeline of all students, instead of using circles, vertical green or red bars were chosen to optimize the timeline's compactness. These bars represented the attempts and when hovering over them, users could view the solution for a particular attempt. A notification system was integrated to inform both students and teachers about interventions that have been taken. To maintain consistency, the distribution was presented with consistent colors.

Furthermore, when clicking on the full box of an exercise, the system displayed the duration of time taken by the student to complete the exercise. Other functionalities such as



5 Wetenschappen Wiskunde

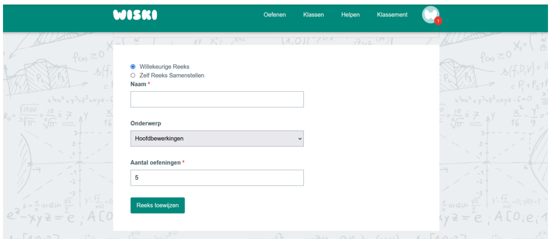
Nieuwe Oefenreeks Toewijzen

Bewerkingen

□

Reeks 1 □

Figure 4.18: High-fidelity prototype: Overview of the exercise sets that a teacher has prepared for a class.



● Willekeurige Reeks
○ Zelf Reeks Samenstellen

Naam *

Onderwerp: Hoofdbewerkingen

Aantal oefeningen *: 5

Figure 4.19: High-fidelity prototype: Form to generate a randomized set of exercises for a specific subject.

asking questions through the interface, were intentionally omitted to reduce the teacher's workload.

4.3.1 Think-Aloud with Education Experts

The high-fidelity prototype was then evaluated through a second think-aloud study involving four teachers and one lecturer from the educational bachelor at UCLL. The results of this study are summarized in table 4.3.1 and the same set of questions was posed as in the first think-aloud study (see appendix C.1).

One of the issues identified was related to the hover functionality for histograms and the vertical bars for the attempts. All participants found it difficult to use. Participants recommended incorporating a click functionality for these elements to enhance accessibility. Additionally, the icons for detections were unnoticed by some participants. They opted to look for poorly or strongly performing students using the histograms instead. This indicated a need to increase their prominence or make them stand out more clearly.

Some participants had difficulty understanding that the bars next to the student's name were related to the distribution bar charts on the left because they shared the same color scheme as the attempts. This highlighted the need for a better visual representation of the data to ensure that users could easily understand the relationship between different elements of the interface.

Furthermore, they suggested increasing the size of the bars for the attempts as well as enlarging the timeline and text to enhance visibility. One participant recommended changing the label "Name" to "Name of exercise set" in the form for assigning a more challenging set of exercises to avoid confusion with the name of the student. Moreover, the participants suggested providing an explanation to help users interpret the timeline and the attempts visualized. One participant also recommended including an explanation of the detection algorithm, as well as the basis on which the detections were made. Lastly, one participant proposed incorporating a feature to indicate the exercises that students find challenging. This suggestion aligned with one of the ideas proposed during the co-design session.

Overall, a lot of positive feedback was given as well. One participant said *"I think it's really good. I would want it automatically for all the exercises I prepare for my students."* Another said *"I find it very user-friendly. You need to get used to it, but it's clear. I can't* *think of any other way to visualize this information."* As well as another participant said *"This tool is very useful for getting a nice overview of the class."*

Table 4.3: Feedback and suggestions from teachers on the high-fidelity prototype

Issue	**Freq.**
Challenging to use the hover functionality for histograms and vertical bars for the attempts.	4/5
Icons for detections go unnoticed.	3/5
Initially does not recognize the relationship between the bars next to the student's name and the distribution bar charts.	1/5
Increase the size of the bars representing attempts as well as enlarging the timeline and text.	1/5
Change the label from "Name" to "Name of exercise sequence" for assigning more challenging exercises.	1/5
Missing explanation to interpreting timeline and the detection algorithm.	1/5
Show which exercises the students are struggling with.	1/5

4.4 Final Proof of Concept

After receiving feedback from the second think-aloud study, several changes were implemented in the final proof of concept (see Figures 4.20 to 4.22).

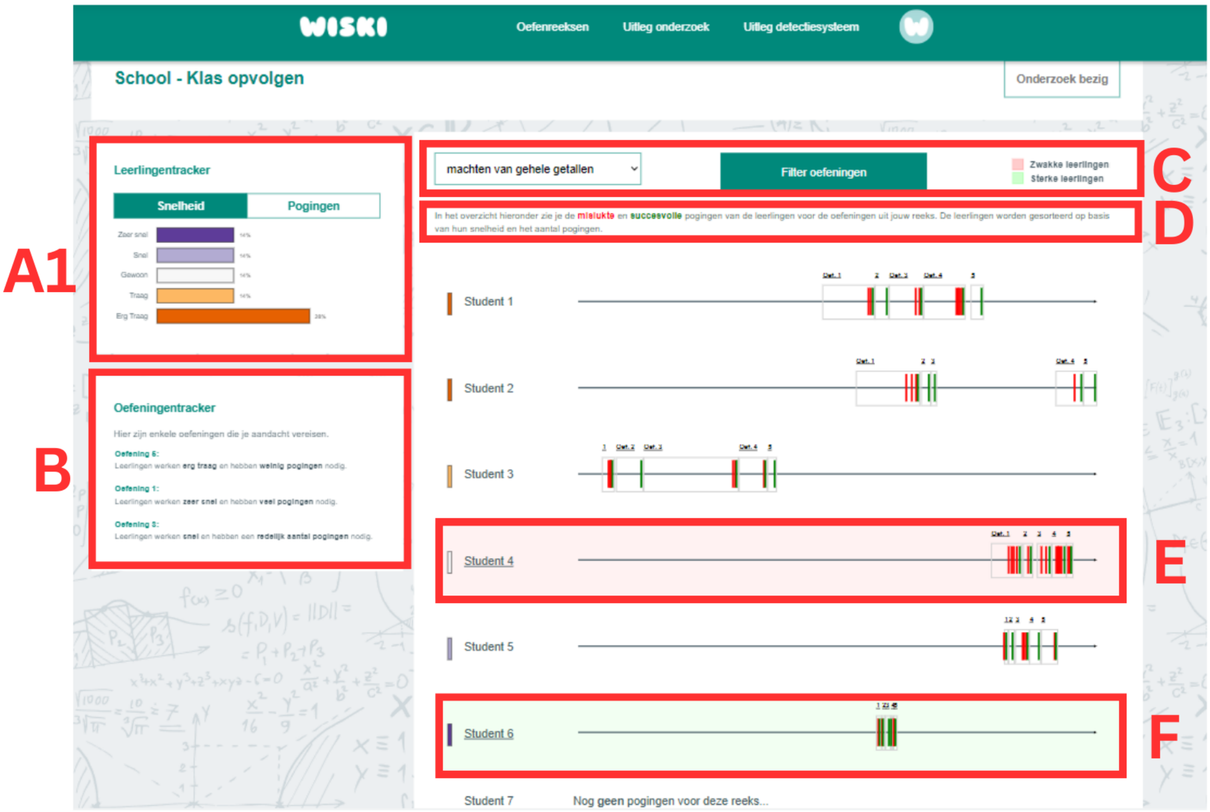
4.4.1 Dashboard with Visual and Textual Explanations

Figures 4.20 to 4.21 show the final dashboards with visual and textual explanations.

Histograms (A1 and A2) The histograms were updated with a new color scheme. That made it easy to distinguish between the timeline and the bar representing each student's position in the distribution. Additionally, a clicking functionality was added to display the same information when a user hovers over the bars of the current histogram or attempts on the timeline.

Exercise tracker (B) An *exercise tracker* was added to the dashboard to display the exercises that students are either excelling at or struggling with. Additionally, textual explanations were added to clarify the reasons behind the detected exercises. The calculation of these scores is explained in detail in section 4.5.

Filtering options (C) The filtering feature for detected students was updated to include a legend that differentiates between the filtering options for students who are either performing poorly or excelling.



School - Klas opvolgen

Leerlingentracker

Oefeningentracker

Hier zijn enkele oefeningen die je aandacht vereisen.

Oefening 5:
Leerlingen werken **erg traag** en hebben **weinig pogingen** nodig.

Oefening 1:
Leerlingen werken **zeer snel** en hebben **veel pogingen** nodig.

Oefening 3:
Leerlingen werken **snel** en hebben een **redelijk aantal pogingen** nodig.

In het overzicht hieronder zie je de **mislukte** en **succesvolle** pogingen van de leerlingen voor de oefeningen uit jouw reeks. De leerlingen worden gesorteerd op basis van hun snelheid en het aantal pogingen.

Student 7 Nog geen pogingen voor deze reeks...

Figure 4.20: Final proof of concept: An overview of all students for the teachers with visual and textual explanations.

Short explanation on the timeline (D) Additionally, a short explanation was added above the timeline to help users interpret the visualization: "In the overview below, you can see the failed and successful attempts of the students for the exercises in your series. The students are sorted based on their speed and the number of attempts." To align with the colors used for the attempts on the timeline, the terms "failed" and "successful" were displayed in red and green, respectively.

Timeline (E and F) The full row is highlighted with green or red to indicate whether a student is excelling or performing poorly instead of using icons. In addition, the bars next to the students' names were kept at a smaller and uniform size. Moreover, the size of the bars on the timeline and the spacing between students were increased.

Clear labels The label of the form for assigning more challenging exercises was edited to avoid confusion.

4.4.2 Dashboard Without Explanations

Figure 4.22 shows the final baseline dashboard that does not provide any data-centric explanations. The dashboard also features an exercise tracker, similar to the previous version but lacks any textual explanation regarding the reasons for detection (A). Additionally, the filtering options were limited to detecting students as there is no timeline

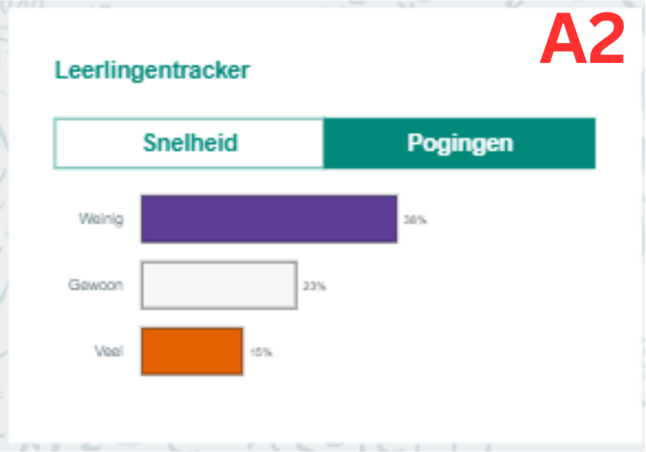
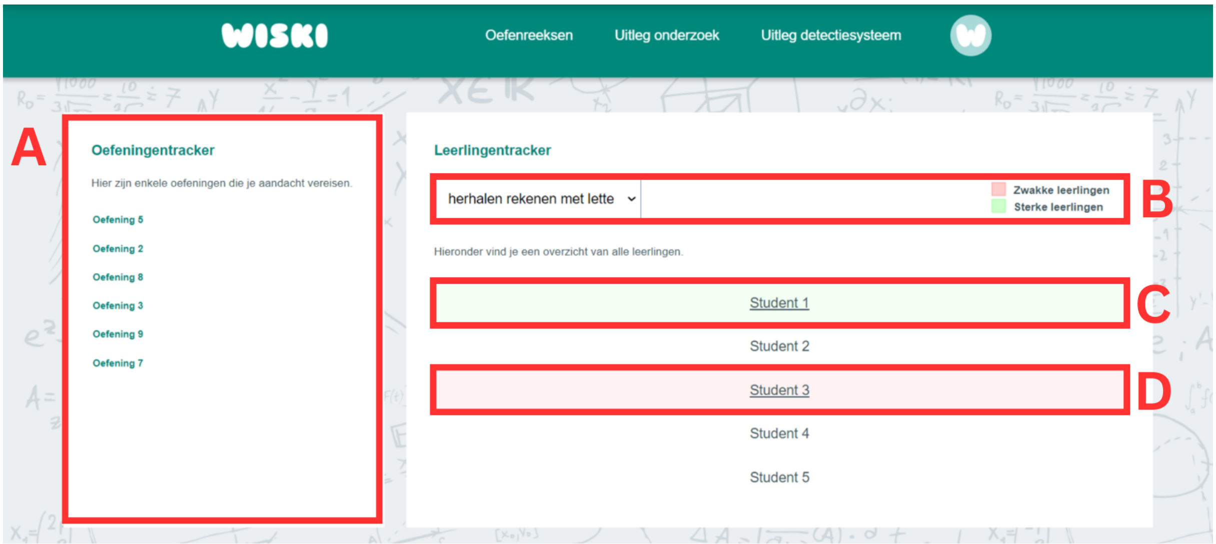


Figure 4.21: Final proof of concept: Histogram for the distribution of the attempts scores.

available for exercise filtering (B). Similar to the dashboard with explanations, detected students were highlighted with green or red colors to indicate their detection status and performance level (C and D).



Hier zijn enkele oefeningen die je aandacht vereisen.

Hieronder vind je een overzicht van alle leerlingen.

Figure 4.22: Final proof of concept: An overview of all students for the teachers without explanations.

4.4.3 Model-Centric Explanation

To enhance transparency on how the detection algorithm works and its parameters, a model-centric explanation was included as shown in Figures 3.8 to 3.10.

4.5 Technical Implementation

This section provides a technical overview of the key components implemented in Wiski, including the outlier detection algorithm, student, and exercise score calculation, the Elo scoring system, the visualizations, the assignment of challenging exercises, and creation of a set of exercises. Wiski was developed using Drupal 7 with HTML, CSS, and JavaScript for the front-end and PHP for the back-end. For more information on Drupal 7, refer to Ooge's thesis [43].

4.5.1 Sets of Exercises

A custom module *Reeks* provided functionalities regarding managing sequences of exercises created by teachers. This includes all the forms for creating either a random or self-selected set of exercises on a subject. SQL is used to obtain the needed data on the subjects and questions.

4.5.2 Interventions

The *Reeks* module can also generate challenging recommendations for a student. To do this, it first identifies the subject of the current sequence that the student is working on and gets their current Elo score for that subject from the database using SQL. The module then finds all questions related to that subject and selects those that have an Elo score between the student's current Elo score and 100 above it, ensuring that the questions are challenging. The recommended questions are then sorted based on their Elo score and the top n recommendations are returned, where n is the number of exercises that the teacher wants to assign.

This module also has a functionality that enables assigning a buddy to a student who has been detected as struggling within the same class. To accomplish this, PHP and SQL are utilized to fetch the list of all students in the class from the database and to designate a buddy selected by the teacher for the student. Every intervention whether it be assigning a buddy, recommending challenging exercises, or providing help from a teacher, is recorded in the database and stored for future analysis. This module also stores the teacher's agreement or disagreement with an intervention in the database. The module also handles notifications for both students and teachers regarding interventions. Students receive notifications when teachers assign more challenging exercises, while teachers receive notifications when the intervention is successfully implemented.

4.5.3 Outlier Detection System

The *Reeks* module provides several functionalities related to a sequence of exercises. One of the main functions is to detect outliers for a specific set of exercises made by students in a class. Each student is given a speed and attempt score for each set of exercises, which are used as parameters for the outlier detection algorithm. The local outlier factor algorithm is implemented in PHP to detect outliers. The algorithm calculates the Euclidean distance between two points in a dataset and uses this distance to calculate the local reachability density and the local outlier factor for each point. These factors are then used to identify points that are outliers. All data points with a factor above 1 are classified as outliers. Additionally, those data points that have both an attempt and speed score above the mean are classified as "strong" students, while the outliers that do not meet this criterion are classified as "weak" students. The period for which the student was detected as an outlier is also stored.

4.5.4 Visualisations

The visualization of the detections and the students' data is presented through a dashboard developed using d3.js, HTML, and CSS. This dashboard is powered by the custom *Opvolgingsdashboard* module, which is responsible for generating visualizations. The module provides a block that can be added to a Drupal page, consisting of HTML, CSS, and JavaScript code. Hovering and clicking data is also stored in the database for various events including: changing the current set of exercises, filtering exercises, filtering weak/strong students, clicking/hovering on detected exercises, clicking/hovering on detected student names, clicking/hovering on histogram bars, clicking/hovering over attempts, and clicking/hovering on exercise names on the timeline.

4.5.5 Student and Exercise Scoring System

In the *Wiski* module offered by Ooge, there is additional PHP code that communicates with the database containing student scores through SQL. This code updates the scores of students when they submit a new answer for an exercise. Additionally, the scores of the exercises are also updated. The attempt and speed scores gained or lost by the student are added or subtracted from the exercise's attempt and speed score, respectively. In this manner, the exercise contains a global attempt and speed score that indicates how much time and how many attempts students need to solve it.

4.5.6 Elo Scoring System

The *Wiski* module incorporates features to calculate the Elo score for students and questions after students have answered the questions. It utilizes a multivariate Elo system to assess student performance, where Elo scores are maintained for each subject and student in the database. When a student attempts a question, the module retrieves the latest subject score and question score from the database using SQL and adjusts them accordingly. The expected scores for the student and the question are calculated, and the Elo score is updated based on the actual score achieved.

Chapter 5

Results

This chapter presents the findings of the switching replications design study and the subsequent interviews. The study involved a group of 11 teachers who volunteered to participate. However, due to time constraints, one participant was unable to participate in the interview. The allocation of teachers to group A and group B is depicted in Figure 5.1. Recall that group A first saw the full dashboard with data-centric explanations and then the baseline dashboard without; group B experienced the reverse order. The quantitative and qualitative findings for perceived accuracy, trust, satisfaction, effectiveness, and the use of Wiski are presented. In addition, it should be noted that the quotes have been translated to the best of our ability to ensure an accurate representation of participants' responses. Finally, as participants had 15 minutes of exposure to the baseline dashboard, they expressed their struggle in evaluating and drawing comparisons with the data-centric dashboard. As a result, many participants primarily directed their attention towards the explanations dashboard, considering it the most vivid and memorable aspect of the experience.

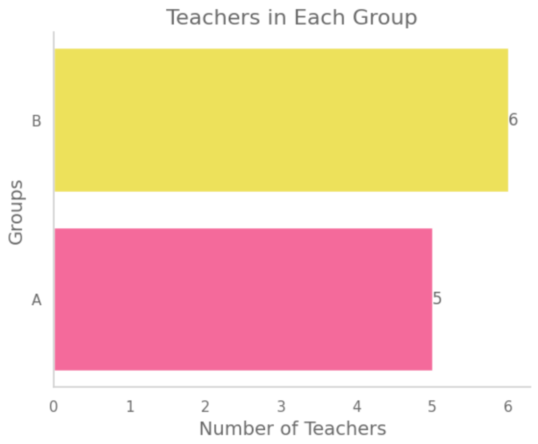


Figure 5.1: Teacher allocation to group A and B.

5.1 Perceived Accuracy

In our analysis of perceived accuracy, we discuss two key themes: (1) the (mis)alignment of perceived accuracy with teachers' findings, and (2) the recognition of the need for additional time and evaluation to assess accuracy more accurately.

5.1.1 Quantitative Analysis

As part of the intervention, teachers were required to indicate their agreement with a detection. Figures 5.2 to 5.3 showcase the distribution of agreements among teachers based on two factors: the presence of an explanation and their allocated groups. In the first figure, it is evident that teachers tend to agree more with a detection when an explanation is provided. The second figure suggests that teachers in Group B tend to agree slightly more than those in Group A, although the difference is not substantial. This difference in agreement could also be attributed to the fact that Group B had one more teacher than Group A.

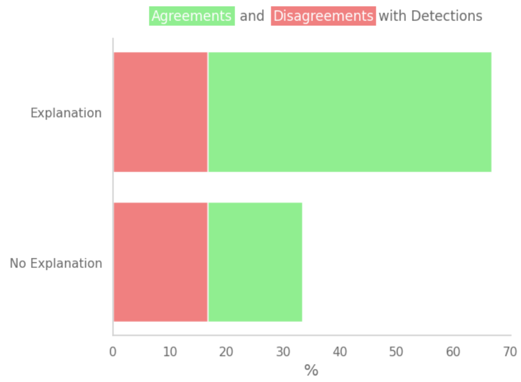


Figure 5.2: Distribution of agreement with detections, categorized by the presence or absence of explanations.

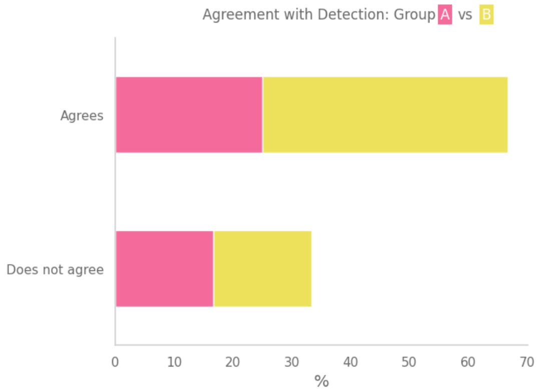


Figure 5.3: Distribution of agreement with a detection by group.

5.1.2 Qualitative Analysis

(Dis)alignment with own findings: Perceived accuracy relies on observations and intuition in evaluating students' performance.

Some teachers expressed confidence in the system's performance and acknowledged its alignment with their expectations. **P5** expressed confidence in the system's accuracy, stating: *"Yes, I do think the detections were accurate. I expected them, but only because the exercises were a bit more challenging. Maybe, yes."* **P5**'s belief in the system's performance aligns with the expected outcomes based on the difficulty level of the exercises. **P3** supported the system's accuracy, exclaiming: *"Yes, except for that one student, I found everything to be in perfect order. Everything matched and made sense, just as I expected. I also knew that one student would guess, and I already had that information. So yes, apart from that one student, I found it all good."* **P3**'s statement demonstrates that the system met their expectations by accurately reflecting the behavior of most students, except for a specific case they were already aware of.

P10 agreed that the system was accurate as they found that there were no surprises. They explained: *"But overall, I followed how it was going, whether it was going well, fast, better or worse. But there were no surprises, really. The students who I expected to perform well actually did so."* **P10**'s experience aligns with their expectations as the students she anticipated to excel matched the system's indications. **P7** noted: *"Yes, the weaker and stronger students corresponded to my expectations."* **P7**'s understanding of her students' abilities allowed them to recognize that the system's assessment aligned with their observations in the classroom.

However, **P9** voiced some concerns about the system's accuracy, perceiving a discrepancy between what they observed in the students and what the dashboard indicated. They pointed out: *"There was quite a difference, I think, between what I observed in the students and what the dashboard said. That wasn't always 100% correct."* **P9**'s experience indicates that the system may not always accurately reflect the students' performances as perceived by the teacher.

Need for more time: Teachers recognize the need for additional time and evaluation to accurately assess the accuracy of the system.

Some teachers expressed a neutral stance and emphasizing the necessity for additional evaluation. **P1** stated: *"[..] But to assess it better, I need to try it a few more times. It's still a bit early to say at the moment. I think it needs to grow. And even then, you know, okay, everything can still change. It should span over a longer period, so more time is required."* **P1** believes that additional trials and an extended evaluation period are necessary to form a more conclusive judgment on its accuracy.

P8 found it challenging to assess the system's accuracy as well, noting: *"Yes, it is difficult to respond to that. I think I haven't been engaged enough with it to give a proper answer. I only spent a day on it."* **P8**'s limited exposure to the system prevents them from offering a thorough assessment, emphasizing the need for more time and involvement to form a well-informed opinion on the system's accuracy. **P2** acknowledged the importance of further exploration, stating: *"I would need to play around with it more to extract valuable insights like that. After a lesson like this, where I also noticed it was the last period [of the day], it's not ideal for that."* The teachers' viewpoints emphasized the necessity for increased engagement and longer-term observation to provide comprehensive feedback on the system's accuracy.

5.2 Trust

In our analysis of trust, we explore several themes: (1) the fostering of trust through data-centric explanations, (2) the time required for teachers to build trust by using the dashboard and observing the detections, (3) blind trust stemming from the belief in the system's objectivity and robust design, (4) trust based on (mis)alignment with teachers' findings, and (5) doubts arising from fluctuations in detections, leading to uncertainty in the system's reliability.

5.2.1 Qualitative Analysis

Data-centric explanations foster trust: The teachers gained trust in the system as they analyzed the visualisations.

P1 used Wiski for multiple classes. The participant thoroughly analyzed the visualisations in the first group and upon observing consistent patterns, they gained trust in the system. The participant believed that the system would continue to be effective in subsequent lessons: *"At the beginning, I had the feeling that I could see it with my own eyes, using the timeline. I wanted to verify if it was accurate. Like, is it being displayed correctly, showing me the time taken and the number of correct attempts? Does it match up? But so far, I do believe that I trusted it."*

P1 expressed how their trust in the system developed as they carefully examined the visualisations. They noticed a correlation between the number of attempts made by students and their detection. *"[...] on the timeline, you could effectively see that the students made only one or two attempts and those who took less time were also chosen as better learners. So yes, I could follow that my strong learners, who tried a lot, were then indicated as weaker. [...] I had looked into that thoroughly with the first group, and with the second group, I could already say, 'Yes, that actually makes sense.' So I had trust that it would work in the next lessons as well."*

Additionally, **P1** observed that the system identified students they considered weaker as stronger performers. They found it surprising but understood the reasoning behind it. They explained: *"I also noticed that students whom I know are weaker were being highlighted as stronger. And I thought, 'Really? Can that be?' But there were instances where it was true. The students who diligently completed their exercises on paper without attempting them multiple times before moving on. I wasn't surprised because I knew that you receive more points for providing the correct answer quickly or getting it right on the first try. In my eyes, a strong student is someone who finds the correct solution, even if it's done quickly. Those strong students don't have a specific problem-solving strategy or method; they just think, 'I need to know right away,' and they start solving. On the other hand, other students take their time. It's funny to see that the weaker students also rank higher. So, in that sense, I think it's nice for them to see that they're also earning points or being ranked higher because they have a good approach to solving problems. So, I found that quite enjoyable."*

Similarly, **P2** also utilized the timeline to investigate why a student was detected. This exploration led to a greater level of trust in the system. **P2** shared an example of a student who typically performed better but ended up marked as a weaker student due to guessing: *But, I could see on the timeline that they were guessing because they genuinely believed they always had a one-in-four chance and would eventually select the correct answer. [...] So yeah, I had trust that the system was identifying the students correctly.*

P3 sometimes disagreed with certain detections but gained an understanding of why a student was being detected by analyzing their behavior through the timeline. **P3** describes the reasoning process behind understanding the detection as follows: *"Yes, yes, I have a few students who are strong but were still rated as less strong because they are very slow. One of my students is extremely slow but very strong in mathematics, so he just needs a lot of response time to react. And I could follow that on the timeline. And yes, that can give a distorted image of his abilities because he was often seen as a weaker student. But yeah, that's purely because he needs so much time, really an extreme amount of time. So he was in the lowest category for speed. And that gave a wrong impression. I found that unfortunate, but yeah, that's how it is with such special students, right? [...] but overall, I thought it was good that there was a balance between the number of times they answered and the speed at which they answered. I thought that was a good way to sort because that's also a difficulty he naturally faces. So yeah, it's logical that it's taken into account."* Additionally, **P3** demonstrated trust in the system, acknowledging that the explanation helped them comprehend the underlying concept: *"[...] it's because I know the underlying idea, the reasoning behind the classifications, that I can say that."*

Building trust takes time: Teachers build trust by using the dashboard and observing the detections.

P1 also highlighted the need for time when building trust and working with the system: *"Yes, because I didn't see it often, I had to rely more on the timeline. Of course, the more I work with the platform, the more my trust in it grows. You learn that the system responds in the right way. But in the beginning, you have to test the system with the exercises. But it did prove to be accurate."*

Blind trust due to assumed objectivity: Teachers' belief in the system's objectivity and robust design.

Some participants who expressed blind trust in the system, believing in its objectivity and setup. **P4** stated: *"I trust it purely based on what I see. If I encounter a student I don't know, I can already deduce their level based on that. [...] I believe it works because there's an objective system that generates judgments based on data and accuracy."* **P4**'s trust stemmed from the perception that the system was designed objectively and relied on reliable data to make assessments. **P5** also expressed a similar perspective, stating, *"Yes, it seems to me that if it is set up that way, it should work correctly. Right?"*

P6 shared a similar sentiment, assuming that the detection system was well-designed. They said, *"Yes, I suppose that if the detection system is well-made, you can trust it, right? But of course, you don't know exactly how it works."* Despite not having a precise understanding of the system's inner workings, **P6** still placed trust in its overall reliability, relying on the assumption that the system was well-crafted. These participants exhibited blind trust in the system, largely because they believed in its objectivity and assumed that it was well-designed. They acknowledged that they may not have a full understanding of how the system operated but still placed their faith in its capabilities.

Trust based on (Dis)alignment with own findings: Teachers trust the system if the outlier detections align with their own observations and knowledge of the students, but distrust it otherwise.

Some participants expressed trust in the system because it aligned with their findings and perceptions. **P2** shared: *"Yes, And I also found that it provided a very quick overview. It was fast enough for me to see, 'Okay, this student is doing very well, while that one is only partially correct,' so it matched how I thought the students would be."*

P7's trust stemmed from their familiarity with the students. They stated: *"Because I've known the students for a while now, I know which ones are strong in mathematics and which ones are slightly weaker. The system also indicated that the high-performing students were doing well and those who were a bit weaker aligned with how I see them in class."*

P3, too, expressed trust in the system based on their knowledge of the students. They explained: *"Yes because I really know my students, I can say that I trust the system. But you know, if I were to start the school year with a completely new class or for any other reason find myself in an unfamiliar setting, I can't be sure if it would be as reliable. It's all about knowing my students, you see, and being aware of their weaknesses. So, in that sense, I can put my trust in it and I also get where it's coming from. I really understand it, what it tells me is definitely on point."* **P3** also highlights that having no prior knowledge of the student might influence how trustworthy they would find the system.

These participants placed trust in the system because its outlier detections aligned with their observations and understanding of their students. They appreciated the system in reflecting their perceptions and assessments.

On the other hand, some participants expressed skepticism and lacked trust in the system due to its inconsistencies with their perceptions and observations. **P8** highlighted this discrepancy, saying: *"Yes, some students were indeed fast and correct, but based on the color coding, I thought they were being labeled as weak learners. I found that strange, yes."* **P8** further emphasized that the system's assessments did not always align with their findings. They remarked: *"No, it didn't always correspond with my observations. In terms of distinguishing between weak learners and strong learners, I interpret it differently in my lessons."*

P9 also noticed a notable difference between what they observed in the students and what the system's dashboard indicated. They expressed distrust and elaborated: *"No, there was quite a difference, I think, between what I observed in the students and what the dashboard said. It wasn't always 100% accurate. [...] The dashboard primarily highlighted which students were fast or slow. I noticed that some students were engaged in guessing, which might have led to them being labeled as fast. Even in terms of strong and weak students, it didn't fully correspond to what I personally observed. So, that's what I would say..."* Despite finding the concept intriguing and appreciating the system's ability to provide an overview of students' speed and performance, **P9** admitted that it didn't entirely align with their intuition.

P10 echoed the sentiment of mistrust, stating: *"According to my knowledge of the students, [I do] not really [trust it]. [...] It could be on their part as well. [...] Some students scored poorly, while in class, I had the impression that they were actually stronger. [...]"* **P10** did acknowledge that there might be some reasoning as to why some of their students were detected but still did not trust it.

These participants expressed a lack of trust in the system as it often contradicted their assessments. They highlighted specific cases where the system's judgments did not align with their own observations and intuition. Subjective interpretation and students' individual understanding of the material were identified as factors contributing to their skepticism and limited trust in the system.

Doubts arise as detections fluctuate: Teachers express uncertainty in the system's reliability.

One participant expressed less or mixed trust in the system due to the fluctuating detections. **P6** mentioned: *"Yes, sometimes, but sometimes not. [...] I don't think so because that also changed, right? It would say 'weak,' and then suddenly that student would be labeled as 'strong,' and then they would become 'weak' again. Do you understand? I think it's really related to the constant running algorithm. But yeah, it does cause confusion at times because, you know, it's like... I mean, things change."* **P6**'s lack of trust stemmed from the inconsistency they observed in the system's detections. The labels assigned to students would shift between "weak" and "strong" intermittently, which they attributed to the running algorithm. This inconsistency and constant change led to confusion and undermined their trust in the system.

5.3 Satisfaction

In our analysis of the satisfaction with explanations, we discuss several themes: (1) overall satisfaction with the model- and data-centric explanations; (2) the value of data-centric and model-centric explanations in facilitating a clearer understanding; and (3) the need for increased exposure to the dashboard to further enhance understanding.

5.3.1 Qualitative Analysis

Overall satisfaction: Teachers generally found the model- and data-centric explanations to be satisfactory, although concerns were raised regarding the method of labeling students as"strong" or "weak"."

Some teachers found the model-centric explanations more than sufficient to get a basic understanding of the underlying algorithm. **P2** appreciated the model-centric explanation, saying: *"That explanation I read beforehand, which you have at the top. Yeah, I found that quite nice. It was handy. I don't think it needs to be much more than that, otherwise it would be too long."* **P3** believed it was concise yet comprehensive, catering to even individuals with limited computer knowledge. They believed it was concise yet comprehensive, catering to even individuals with limited computer knowledge.

Some participants expressed satisfaction with the data-centric explanations provided. As highlighted by **P10**: *"If you can already see this in the first trimester, then you have an idea about your students before you fully know them."* This early visibility into students' abilities was considered significant by the participants. One participant, **P4**, appreciated the timeline feature on the dashboard, stating, *"I really appreciate the timeline... It provides a clear snapshot [...]"* **P2** emphasized the question tracker feature of the dashboard: *"I found it particularly useful when a question required extra attention."* Highlighting the strengths of the dashboard, **P5** stated: *"The dashboard's strength lies in its ability to present this information concisely and accessible."*

However, it should be noted that there were also teachers who expressed a need for more detailed explanations due to their uncertainties about how the system made its detections. **P6** wondered about the system's process for determining whether someone is strong or weak, questioning how teachers themselves determine a student's weakness. They hoped for a space where more questions could be answered before reaching a judgment. They emphasized that the labels of "strong" or "weak" were presented without any accompanying explanations. **P1** also expressed the desire for more details about the precise allocation of points per attempt but emphasized that it was not necessarily essential information: *"You can see the score, but you don't know how many points are awarded for it. Whether it's 5 points, 2 points, or 3 points, it's not really important to know, I think. It's not displayed visibly, like 'This student has 350 points.' Maybe it would be nice to know how many points they receive for the first correct answer, but it's not necessary information."*

Facilitating Understanding: Teachers found the data-centric and model-centric explanations valuable for gaining a clearer understanding of the algorithm and model parameters.

Some teachers used the information they read from the model-centric explanation to go into detail about the working of the parameters used to detect outlier. **P1** described the scoring system based on their understanding: *"The detection system looks at how you solve the exercise correctly. You receive a higher score if you get it right on the first or second attempt and as you make more attempts, the score decreases. So, you get a lower score. [...] But anyway, as you make more attempts, you get a higher or lower result."* **P3** mentioned that the detection system determines the students' ratings and generates scores for learners. They noted that there seemed to be a formula behind it awarding points based on the number of attempts students made to arrive at the correct answer along with their speed or slowness in responding. **P3** noticed that the ratings changed when more students completed the exercise series, resulting in higher scores for them.

Some teachers utilized the data-centric visual components to gain insight into the algorithm's functioning and parameters. **P4** and **P8** shared their interpretations in this regard. **P4** speculated: *"The detection will probably be based on the accuracy of the exercises, I think. And probably also on speed, I assume. Speed and correctness."* **P8** offered a similar perspective, stating: *"Yes, based on correct and incorrect answers and probably also based on speed. So, I suppose fast and accurate answers represent strong students."*

P7 contributed their observations based on the visual components. They noted: *"I* *think it is mainly based on how fast and how accurately they answer. You could see that the students who turned green were the ones who were very fast and finished the exercises quickly, while others were a bit behind. So, I think speed and correctness are two important aspects."*

Regarding understanding the algorithm, some teachers had limited knowledge. **P9** admitted to not focusing much on the explanation and rushed through it due to time constraints. **P2** mentioned not having any idea of how the students are detected, while **P10** had a hard time explaining the parameters of the model.

> **Developing expertise:** Teachers emphasized the need for more exposure to the dashboard in order to improve their understanding of the data-centric explanations.

Furthermore, teachers highlighted feeling overwhelmed during their lessons. They had to interpret the visual components and data-centric explanations. They emphasized the need for additional information. **P1** specifically pointed out the lack of visibility into the student's perspective on the platform. They stated: *"I could select my exercises in advance, but I couldn't see how it would appear for the students. If there were some example screenshots, it would have been helpful."*

P8 echoed this sentiment. They expressed their preference for having prior exposure to the dashboard. This would enable a better understanding of the colors and information presented. They remarked: *"You have different colors, right? You have the red, green for correct and incorrect. You have weak students, strong students. You have the speed. There are so many things. Especially when I saw it for the first time, I found it really difficult to grasp everything at once."* **P8** suggested providing a preview of the dashboard with fictitious names. This would allow teachers to analyze and familiarize themselves with the various colors and information in advance.

5.4 Effectiveness

In our analysis of effectiveness of explanations, we explore utilization of data-centric explanationsw of Wiski.

5.4.1 Quantitative Analysis

Figure 5.4 provides an overview of the interventions carried out by participants, comparing the presence or absence of data-centric explanations. Interestingly, participants tended to register more interventions when no explanation was present. Figure 5.4 also depicts the changes in intervention frequency for group A and group B before and after their respective changes. Group B showed a higher tendency to intervene when no explanation was available. Furthermore, after the introduction of explanations, group B intervened less frequently compared to group A. In contrast, group A exhibited consistent intervention rates regardless of the presence or absence of explanations.

Figure 5.5 displays the different types of interventions and the corresponding counts for both scenarios. The exploration of the dashboard's visual components is represented in

Figure 5.6 , illustrating the number of times participants hovered over specific elements. Notably, Group A demonstrated a higher level of exploration, engaging with a broader range of components compared to Group B.

Figure 5.7 focuses on the click count for the dashboard with explanations, providing an indication of participants' interaction with the interface. Conversely, Figure 5.8 presents the click count when no explanation were available, offering a comparison between the two conditions.

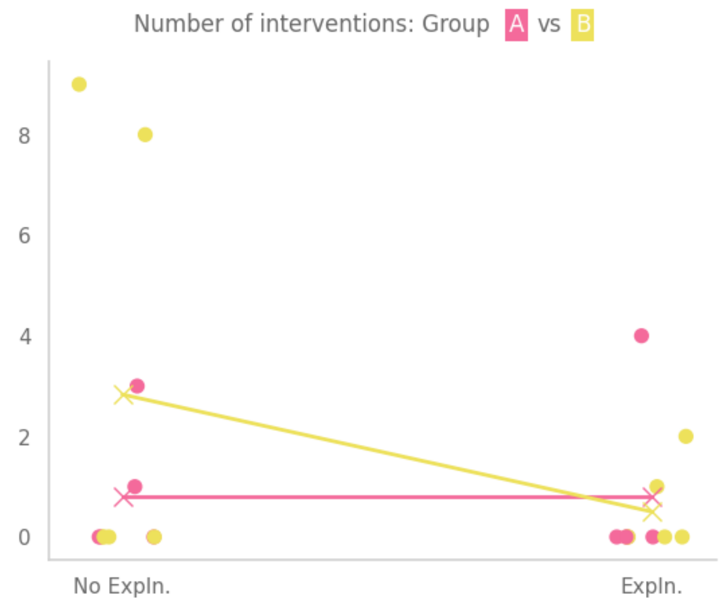


Figure 5.4: Changes in intervention frequency for group A and group B with and without data-centric explanations

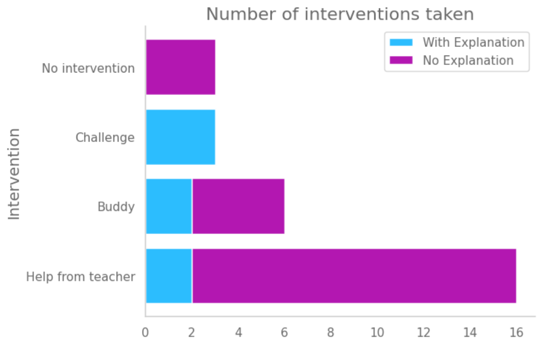


Figure 5.5: Distribution of different interventions with and without explanations

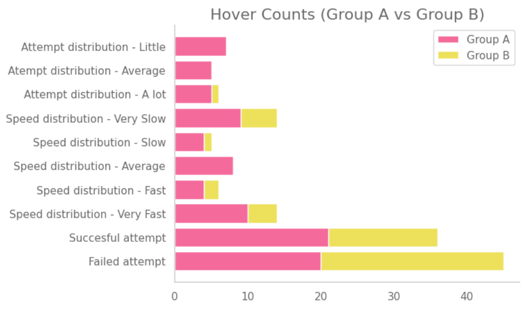


Figure 5.6: Figure 5.7: Exploration visual components of explanations

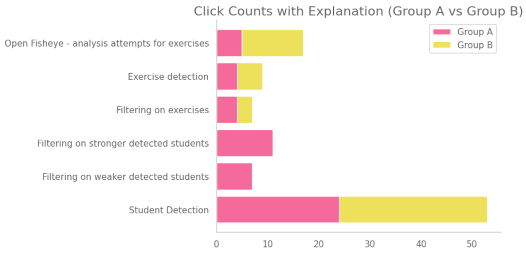


Figure 5.7: Click count for dashboard with explanations.

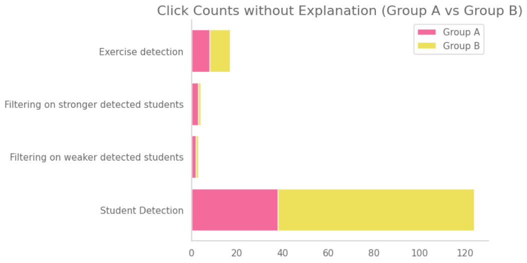


Figure 5.8: Click count for dashboard without explanations.

5.4.2 Qualitative Analysis

Visual components for effective teaching: Teachers utilized the timeline, the histograms, colour coding and question tracker to monitor student progress and intervention.

The visual components helped in supporting teachers' decision-making processes and optimizing classroom management. **P10** highlights the significance of early visibility into students' abilities, stating: *'If you can already see this in the first trimester, then you have an idea about your students before you fully know them. You know how quickly they can complete exercises. This allows you to assign more tasks to some students within the same timeframe as others.'* **P1** shares their observations, noting how the timeline helped them decide to intervene, saying: *"I did want to help the person who had been working on it for so long when I looked at the timeline. I thought, 'Let me go and help that person.' But then I had to decide whether to assign a buddy, help myself, or just do nothing."*

P4 expresses their initial impression of the visual components, stating: *"The dashboard without the timeline is quite straightforward and doesn't show much progression. It was rather simplistic. I didn't have a comprehensive overview since I had selected three classes and students, but I didn't really see any follow-up or evolution. I had only tried it once, so I didn't witness any ongoing progress or development."* **P4** prefers the dashboard with the timeline feature and found it particularly valuable, expressing: *"I really appreciate the timeline... It provides a clear snapshot of when students start and complete their exercises, allowing me to assess their progress and identify areas that need attention.*

P5 emphasizes the dashboard's tracking capabilities, stating: *"For students, I can easily track their completion time, performance, and problem-solving approach... The dashboard's strength lies in its ability to present this information concisely and accessibly."* **P5** acknowledges the potential of using the visual components to assess which students needed remediation or using it to differentiation: *"I think it really useful for situations where I need to assess which students need remediation. Just looking at the dashboard, I would pick out the student who make a lot of mistakes. Or even when differentiating, the bars really help in grouping students together and seeing how much time or attempts they need."* Highlighting the value of the histograms in classifying students based on their performance.

P2 emphasized how the timeline effectively assisted them in identifying students who were guessing, needed more time, or performed exceptionally well. They mentioned: *"You can see when they have done the exercises, and at the top, you can see how long they spent on the exercises. That was also nice to see. Like, 'okay, that went very smoothly,' or 'they apparently had more difficulty with this one'. So I could check up on students when necessary. You could also precisely observe when they all went to get paper because they didn't do it the whole time. One would go get paper, which influenced others to think, 'I should do that too,' and so on. So, you could definitely notice these patterns and extract valuable insights from them. That's why it was handy for me to have access to this information. It helped me understand that an exercise took longer because they also went to get paper, and I could observe their movements. Additionally, it allowed me to spot when someone was guessing because for one person, it went very quickly, and I thought, 'yeah, they're just guessing.' and address the student on it. So, I found that aspect useful. The timeline provided useful information, and when you opened the exercise for each student, it was easy to quickly see what was correct and what deviated. I found it handy that I could grasp that information at a glance since they worked with colors."* They also expressed a preference for the dashboard with data-centric explanations, saying, *"I found the second one much more convenient. I would use the second one more than the first. The first one was okay, but with the second one, I had a feeling like, 'okay, yeah, I spent a lot more time on it, clicking and exploring.' So, I found the second one more practical, especially as a teacher."*

P2 highlights the dashboard's immediate feedback feature, explaining: *"I found it particularly useful when a question required extra attention."* The visual components allow **P2** to identify questions that demand additional support and intervene promptly to assist her students. **P2** also shares their thoughts on using the dashboard for remediation and for task differentiation. They said: *"I would love to use it to differentiate tasks. For example, if a student performed poorly on a particular test section, I could assign them additional exercises. They would receive feedback on whether their answers were correct or incorrect."* The use of color-coded indicators within the dashboard is highly valued by the teachers.

P7 describes their impact, stating: *"The red and green indicators are so clear... They quickly catch my attention and prompt me to take action."* These visual cues enable teachers to promptly address students accordingly. Moreover, **P7** highlights how they utilized the timeline to evaluate if students required interventions. They mentioned, *"I really enjoyed being able to track in real-time which exercise the students were working on. Otherwise, you have to literally look over their shoulders, and they don't really like that. Now, I could simply see on my screen which students were engaged in which exercise. Especially if a student had been working on an exercise for a long time, I could instantly notice that. So, it was very convenient because then I could think, 'Hey, that student might need to do something different.' It was truly nice to be able to track it so quickly. [...] Yes, especially the time aspect. And sometimes, when you see that they were not doing the exercises correctly, you also notice whether they have tried it multiple times and failed. Then you think, 'Maybe I should take a look myself.' [...] I found it handy to see my entire class listed and in a learning path, I could see where they were and whether they had completed the exercises correctly or not. I simply found that convenient. I didn't need* *to know more than that. And if I wanted to see what went wrong, I could click on it. It didn't need to be immediately visible."*

P10 also expresses their preference for the dashboard with data-centric explanations as they found it beneficial for assessing the amount of time students required. They stated: *"[...] with the first dashboard, apart from the colors indicating strength or weakness, you didn't see much. But with the second dashboard, you could see more of those answers. You would see lines that were close together and lines that were further apart. That was more interesting at that moment. [...] But actually, I found the second dashboard more organized because you can see better how long they work on an exercise and how smoothly it's going. You can gauge that better. It also provides more assistance as a teacher."*

The real-time timeline feature, color-coded indicators and question tracker provide valuable tools for monitoring student progress, identifying areas of concern and tailoring instruction.

5.5 Wiski in a Classroom Setting

In analyzing the use of Wiski, several themes emerged regarding its implementation and impact in the classroom, such as: (1) sharing insights with students, (2) finding a balance between interpreting the visual component and classroom management, (3)valuing personalized teaching, (4)and the need for control over the system.

5.5.1 Quantitative Analysis

Sharing insights with students: Exploring the motivating and potential negative impact.

Utilizing classroom dashboards includes the ability to share insights with students, giving them a glimpse into their own performance. Some teachers chose to share dashboard information directly with their students, while others express their reservations about doing so. **P1** shares their positive experience, stating: *"I could say to them, 'Hey, the system recognizes you as a strong student' and you could see them light up with pride. So, I found that to be great."* **P5** describes the students' reactions, saying: *"They found it fun to see if they were considered good or average students. It even sparked some laughter."*

P2 expresses their reservations about sharing the dashboard with students, explaining: *"As I mentioned, it's not easy to share the dashboard with the students. However, if they see other students who didn't score as well, the boys in the class can become quite competitive and start mocking others. So, I believe we can use it for ourselves, but not share it with the students. I'm afraid they might approach others and say things like "oh, that person made such a silly mistake" or simply because they discovered which exercises were challenging for them. I would like to avoid that."* **P10** reflects on the impact of the dashboard on her dynamic and energetic class, noting: *"It was indeed different from the norm. The class was quite lively, and they were definitely active, especially when you could say, 'Come on, you're in the red zone, let's get back to green.' They would make an effort, although some more than others."*

Finding balance: Combining interpretation of components and classroom management

Integrating a classroom dashboard into the teaching environment presents both challenges and opportunities for teachers. **P8** shared their initial struggles, remarking: *"I had to figure out how it worked and what it exactly meant. It was a lot to process on the computer while also attending to the students."* **P9** echoed the sentiment, emphasizing the delicate nature of motivating students while monitoring the dashboard. *"It's good, but it's difficult to monitor the dashboard and walk around at the same time. You would need to be in two places at once."* Striking the right balance between offering assistance and leveraging the dashboard's insights proved to be a complex endeavor.

P2 explained: *"Identifying which exercises required attention was very useful. However, I had to physically turn around to see who was actually working on them."* Balancing the need to observe individual progress on the dashboard and physically interact with students posed logistical challenges. **P2** further elaborated: *"I found myself shouting in two directions from the center of the room. It didn't make much sense."*

Valuing personalized teaching: Balancing technology and personal connection.

P1 acknowledges the significance of nuanced understanding when using dashboards. They highlighted that students' classification as strong or weak based on performance alone may overlook individual circumstances and growth. **P1** believed that the personal connection remains essential in catering to each student's unique needs and providing appropriate challenges. They states, *"You know, it's a combination of a good teacher and the system that together ensure a good lesson or exercise."*

P8 expressed hesitance in fully relying on computers, emphasizing their preference for the personal factor in teaching. They believed that the teacher's presence, ability to explain concepts and observe students' facial expressions are crucial elements that technology cannot fully replace. **P8** states: *"I'm not naturally inclined to trust the computer in that sense. But yes, I think the human factor is important."*

P9 concurs with the importance of personal interaction, particularly when teaching students who may have had negative experiences with mathematics in the past. **P9** highlights the efficiency of standing beside students, providing explanations and observing their reactions. They admit: *"It's less evident in the direction I executed, but it's more efficient if you're next to the student, showing and explaining. Being able to see their facial expressions. Especially for those that haven't always had good experiences with math."*

P6 acknowledges the sentiment of valuing personal contact with students. They found it more effective to physically move around the classroom, observing individual progress, and tailoring instruction accordingly. **P6** states: *"I'm actually a strong advocate for personal contact with students, so in that sense, I find it much more enjoyable to walk around and see who can do the exercises and who cannot."*

P7 recognizes the potential benefits of technology in distance learning scenarios. However, in a traditional classroom setting, they find that personal interaction is more logical and effective. **P10** expressed that with the dashboard without the data-centric explanations, their reliance on face-to-face engagement was greater, as it did not offer the capability to provide additional explanations. However, with the visual components, they felt less inclined to approach individual students as they already had a clear understanding of the situation.

Need for control: Tuning parameters and customizing visual components

P3 and **P10** highlighted the importance of self-determined parameter weights and customizable visual components. **P3** recognizes the value of assigning appropriate weights to parameters based on individual student needs. By adjusting the emphasis on certain aspects, such as slowness or speed, teachers can tailor their approach to cater to each student's unique learning style. **P3** suggests: *"Maybe it would be an idea to make adjustments for such a student, so that less emphasis is placed on slowness in some learners because you're aware of it."*

P10 highlights the need for dashboard personalizing, specifically regarding the visual components. Acknowledging that different teachers have varying preferences and teaching styles, they suggests the ability to seamlessly switch between different dashboard layouts: *"Actually, I should be able to switch between the two. I find that quite interesting as well. Or even just choosing yourself what you want to see on your screen as the teacher."*

Chapter 6

Discussion

This chapter addresses the research questions posed in Chapter 3 by utilizing the insights from both quantitative and qualitative data presented in Chapter 5. The first topic examined is the theme of perceived accuracy and trust in explanations. Next, the trade-off between satisfaction and effectiveness of explanations is explored. Additionally, this chapter highlights the limitations encountered during the study.

6.1 Trust in an Explainable Outlier Detection System

Research Question 1

Which factors affect teachers' trust in an explainable outlier detection system? For example, how do perceived accuracy and understanding of the system affect trust?

Trust in the system is influenced by various factors, which were explored in this study. Four key themes emerged as significant in shaping teachers' trust:

- **Explanations:** Providing explanations played a pivotal role in fostering trust within the educational detection system. Through the analysis of visualizations provided by the system, teachers gained a deeper understanding of how the system arrived at its detections. This aligns with the findings of Bussone et al. [76] showing that fuller explanations of the data used in making a decision had a positive effect on trust.

- **Blind trust in the detection system:** Some teachers exhibited a belief in the objectivity and accuracy of the detection system. They placed blind trust in the system, assuming that it would provide accurate assessments without questioning its reliability. This form of trust was based on the belief that the system was designed to be objective and unbiased. This finding is consistent with previous research, which has demonstrated that users may not always be "sensitive" to the reliability of automated systems and, in some cases, place even greater trust in them than in their judgments [77, 78].

- **(Dis)alignment with own findings:** Teachers relied on and trusted the system only when its results aligned with their observations and knowledge of the students. This finding is consistent with the research conducted by Nourani et al. [51], which emphasizes the significance of domain experience in influencing users' decisions to trust a system. In this particular case, teachers tend to place their trust in the system when its outlier detections align with their knowledge and observations of their students.

- **Doubts arise as detections fluctuate:** Teachers expressed uncertainty and doubts about the system's reliability when they observed fluctuations in the detections. Inconsistencies or unexpected variations in the system's output led to questioning its accuracy and thus decreasing their trust. This finding corresponds to the research conducted by Dzindolet et al. [79], which suggests that users often have high expectations of automated systems, anticipating near-perfection. Consequently, users tend to be less forgiving of any mistakes made by these systems.

The study's findings indicate that trust and perceived accuracy in the educational detection system are influenced by two key factors:

- **Explanations:** Explanations play a crucial role in influencing perceived accuracy. When outlier detections deviate from teachers' expectations, they are initially perceived as inaccurate. However, the presence of explanations facilitates understanding of the outlier detections, leading to an enhanced perception of the system's accuracy. This aligns with the findings of Nourani et al. [61, 62], who emphasized the significant influence of meaningful explanations on the perception of system accuracy. Figure 5.2 demonstrates the possible impact of explanations on teachers' agreement with detections. It reveals that when explanations are available, teachers tend to show a higher level of agreement with the system's detections. However, it is worth noting that this effect could also be attributed to the alignment between the outlier detection and teachers' expectations.

- **(Dis)alignment with own findings:** This means that teachers' perceptions of the system's accuracy were influenced by how well the system's detections aligned with what they expected to see based on their observations and intuition. If the outlier detections aligned with their findings, they perceived the system to be accurate. This finding indicates that teachers exhibit confirmation bias when it comes to AI. They hold the belief that their pedagogic experience, human intuition, and prior knowledge about students surpass the AI system [80]. As a result, they tend to dismiss it when its predictions conflict with their existing perceptions of individual students and the entire class.

It becomes evident that explanations and expectations have a notable influence on perceived accuracy. When users are provided with meaningful explanations or if the outlier detection aligns with their expectations, their perception of the accuracy of a system is likely to be positively influenced. Similar to the alignment between the system's detections and the teacher's expectations.

Importantly, perceived accuracy itself acts as a key driver in shaping trust. When users perceive a system's performance as accurate, they tend to develop a higher level of trust in that system. Several studies provide evidence supporting the connection between perceived accuracy and trust in systems. Yu et al. [59] observed that users establish trust in systems based on the observed accuracy. Users possess the ability to estimate system accuracies and adjust their trust levels accordingly. When users perceive the system's performance as accurate, it has a positive impact on their trust, resulting in decisions that align with the system's recommendations. Similarly, Yin et al. [60, 62] discovered that perceived accuracy influences user trust, leading them to increase their trust in a model when its observed accuracy exceeds their own.

Therefore, this highlights that explanations and expectations exert their influence on trust indirectly through their impact on perceived accuracy. Figure 6.1 illustrates the influence of explanations and expectations on a user's perception of system accuracy. This perceived accurarcy, in turn, plays a significant role in shaping trust in the system. When teachers have an initial perception of low accuracy in the system, their trust in it is diminished. However, providing explanations that demonstrate the reasoning behind outlier detections can positively influence their perception of accuracy, subsequently increasing trust. Similarly, if the system's outlier detections align with teachers' expectations, they perceive the system as accurate and trust it. Conversely, misalignment between expectations and outlier detections can lead to a perception of low accuracy and diminished trust in the system.

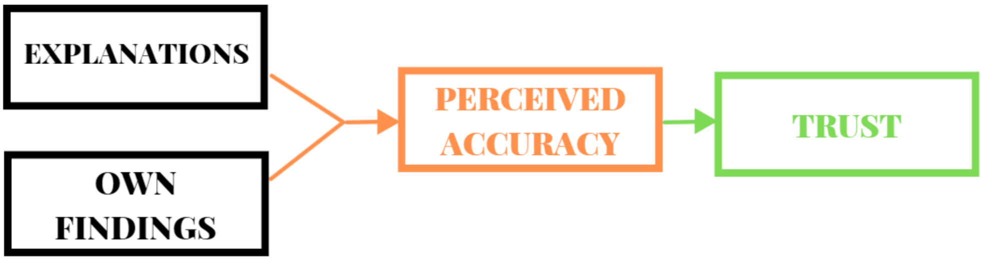


Figure 6.1: Influencing chain: (1) Expectations and explanations have an impact on perceived accuracy. (2) Perceived accuracy, in turn, influences trust.

6.2 Model-Centric and Data-centric Explanations: Effectiveness and Satisfaction

Research Question 2

How do teachers assess model-centric and data-centric explanations in terms of effectiveness and satisfaction?

6.2.1 Satisfaction

We further delve into the satisfaction of teachers with the provided explanations:

- **Overall satisfaction:** The teachers' were overall satisfied with both model- and data-centric explanations. The concise and accessible nature of model-centric explanations was appreciated by teachers, while the benefits of early visibility into students' abilities through data-centric explanations were acknowledged.

- **Explanations facilitated model understanding:** Some teachers used the model-centric explanation to understand the working of the outlier detection parameters, such as the scoring system based on the number of attempts and speed. However, some teachers interpreted the data-centric visual components as indicators of the algorithm's functioning and parameters, such as speed, and correctness of answers. Highlighting how visual explanations enhance the comprehension of the prediction output generated by the black-box model [81].

- **Need for time:** Teachers emphasized the need for time to develop familiarity and experience with the dashboard to feel satisfied and confident in its use. Initially, they felt overwhelmed by the visual components and data-centric explanations presented to them. The sheer amount of information, such as different colors representing correctness or incorrectness, weak or strong students, and speed, can be challenging to grasp all at once. They express a desire to see how the exercises appear to students and suggest that example screenshots would be beneficial in this regard. By having a preview or simulated environment with fictitious names, teachers can analyze and familiarize themselves with the various elements of the dashboard before implementing it in their actual classrooms.

6.2.2 Effectiveness

We further delve into the effectiveness of the provided explanations:

- **Effective visual components:** The effectiveness of explanations in the educational detection system was investigated through a comprehensive analysis of both qualitative and quantitative data. The qualitative analysis revealed that teachers actively utilized visual components such as the timeline, color coding, and detected questions to monitor student progress and inform their intervention strategies. During the semi-structured interviews, teachers highlighted how they used the timeline specifically to strategically identify those students they wanted to intervene. This indicated the potential effectiveness of these visualizations in supporting teaching practices.

- **Difficult to draw definite conclusions:** To further assess the effectiveness of the data-centric explanations quantitatively, we examined the impact on teacher interventions. Figure 5.4 illustrates the intervention frequency changes for groups A and B before and after the introduction of explanations. It shows how group B made slightly more interventions when no data-centric explanations were available. However, it appears that only a few participants made interventions. In Figure 5.6, we observe that group A, initially exposed to explanations, exhibited an exploratory approach on the data-centric dashboard. However, it is important to note that the overall number of interventions by teachers was low, making it difficult to draw definitive conclusions about the effectiveness of the data-centric explanations. Several factors may have influenced these results. For instance, group A and B might have had a more exploratory approach at the beginning of the study. Additionally, the availability of data-centric explanations might have prompted participants to reconsider the need for interventions and adopt a more strategic approach based on the provided data. It is also worth considering that teachers may have physically intervened in the classroom, as their movements were not solely dependent on the platform.

Research Question 3

How do teachers (intend to) utilize a dashboard with outlier detection in a classroom setting?

- **Sharing insights with students:** Interestingly, certain teachers found value in sharing the dashboard and its insights with their students. They saw it as a motivational tool to enhance performance and effort. However, caution is advised to prevent unnecessary competition among students.

- **Hard to find balance:** Some teachers faced challenges in assessing the system while simultaneously providing necessary assistance to their students.

- **Loss of personal connection:** Concerns were raised regarding the potential loss of personal connection when using a tool like Wiski. However, it was emphasized that a combination of a skilled teacher and a reliable system ensures an effective lesson.

- **Need for control:** Teachers expressed the need for control in the system. They suggested implementing customizable thresholds for specific parameters based on individual student learning processes. Additionally, they desired the ability to select and view specific visualizations on their screens.

6.3 Limitations and Future Work

6.3.1 Limitations

The study has several limitations, which are as follows:

1. Need for more time: Teachers recognized the need for additional time and evaluation to accurately assess the accuracy of the system. It takes time for the scores to adjust and the detections to become more accurate, which may influence the overall perception of the system.

2. Limited number of participants: The study had a small number of participants, which may limit the generalizability of the findings. A larger sample size would provide a more representative perspective.

3. Parameter settings and simplified scoring: The effectiveness of the system is dependent on the parameter settings of the detection algorithms. The study did not explore the impact of different parameter configurations, which could influence the accuracy and reliability of the system. Additionally, the scoring and detection systems used in the study were relatively simple. This simplicity may not fully capture

the complexity of student performance and may limit the depth of analysis and insights that can be obtained from the system. However, it is important to consider the potential value of a multivariate approach of the Elo-rating system in future research. This approach can provide a more comprehensive understanding of students' mastery levels.

4. During the study, teachers may have offered assistance to students outside the system's recorded interventions, as they actively moved around the classroom. This discrepancy between the recorded interventions and the actual interventions performed by teachers introduces a level of uncertainty and may affect the accuracy of intervention measurements [82]. Future studies should consider implementing measures to ensure more accurate tracking of interventions.

These limitations should be taken into account when interpreting the findings of the study and considering the applicability of the results in broader contexts. Future research could address these limitations to provide a more comprehensive understanding of the system's effectiveness and potential improvements.

6.3.2 Future work

Based on the findings and limitations of the current study, the following suggestions for future work can be considered:

1. Conduct a switching replications study over a longer period: Extending the study duration would provide a more comprehensive understanding of the long-term effects of explanations on teacher intervention rates and the overall effectiveness of the educational detection system.

2. Implement control over parameters and visual components: Allowing teachers to customize and control the parameters and visual components of the explanation dashboard could enhance their satisfaction and engagement. This customization would enable teachers to tailor the system to their specific needs and preferences.

3. Explore alternative scoring systems: Consider incorporating factors beyond speed and the number of attempts into the scoring system. By considering additional variables, such as student engagement, comprehension, or mastery of specific topics, a more comprehensive understanding of student performance and progress can be obtained.

4. Explore the application in distance learning scenarios: Investigate the applicability and effectiveness of the explanation dashboard in distance learning settings. With the increasing prevalence of online education, understanding how explanations impact teacher intervention and student performance in remote learning environments would be valuable.

5. Include long-term progress in pilot studies and think-aloud sessions: When conducting pilot studies and think-aloud sessions, consider assessing not only short-term student performance but also their progress over time across different topics. This broader perspective would provide insights into how the explanations can support

long-term learning and intervention strategies.

6. Analyze the impact on students' behavior when sharing the dashboard with them: Even though some participants shared their experiences in doing so. More potential benefits and drawbacks of sharing the explanation dashboard with students can be explored. Analyzing how students respond to the information provided and their engagement with the system could offer valuable insights into student-centered learning and self-regulation.

Chapter 7

Conclusion

This thesis aimed to investigate the topic of explainable outlier detection in education. To accomplish this, the Wiski platform was extended with a teacher-facing dashboard that utilized an outlier detection algorithm to identify struggling or well-performing students based on their speed and number of attempts. The explanations provided for this model were developed through an iterative process, which included a pilot study, a co-design session, and two think-aloud studies.

To evaluate the system, a switching replications study was conducted, involving 11 participating teachers. This study aimed to investigate the factors of trust in explanations such as perceived accuracy, the effectiveness and satisfaction of the explanations, and the use of Wiski in a classroom setting. (1) The findings indicate that explanations and alignment with teachers' perceptions have an impact on perceived accuracy, which, in turn, influences trust. Thus, these factors indirectly affect trust through their influence on perceived accuracy. (2) Teachers expressed overall satisfaction with both model- and data-centric explanations, finding them effective in understanding the system. They utilized the data-centric explanations to comprehend the detection parameters and to gain insights into the algorithm's functioning, while also highlighting the need for time to develop familiarity with this dashboard. The qualitative data highlighted that visual components were effectively used by teachers to monitor student progress and guide interventions. However, due to low overall intervention rates and various influencing factors, it was difficult to determine the precise effectiveness of data-centric explanations. (3) Certain teachers found value in sharing the dashboard and its insights with students as a motivational tool, but caution is advised to prevent unnecessary competition. Teachers faced challenges in assessing the system while assisting, and concerns were raised about the potential loss of personal connection. They expressed the need for added control, such as customizable thresholds and the ability to select specific visualizations.

Appendix A

Pre-Study Phase

A.1 Brochure

Uitleg over het onderzoek
Dag bezoeker! Bedankt voor je interesse om deel te nemen aan het wetenschappelijke onderzoek van de masterthesis van Anissa Faik! Hier zal je lezen hoe het onderzoek precies zal verlopen. Als je alles begrijpt dat je op deze pagina kan lezen, kan je kiezen of je akkoord gaat. Indien wel, kan je je registreren voor Wiski en deelnemen aan het onderzoek.

Doel van het onderzoek
Elke leerling is anders. Sommige hebben meer ondersteuning nodig en andere missen uitdaging. Ik heb voor mijn thesis voortgebouwd op een online wiskundeplatform "Wiski". Dit platform bevat duizenden oefeningen van Die Keure, uitgeverij van de wiskundehandboeken zoals Van Basis Tot Limiet. Daarnaast bestaat het platform ook uit een detectiesysteem dat aangeeft welke leerlingen moeilijkheden ondervinden of meer uitdaging nodig hebben. Met mijn onderzoek wil ik te weten komen of dit systeem de leerkracht effectief ondersteund.

Hoe verloopt het onderzoek?
Om toegang te krijgen tot de oefeningen op de website, dienen zowel leerkrachten als leerlingen zich te registreren. Na registratie kan een leerkracht een reeks oefeningen klaarzetten voor zijn/haar klas, welke de leerlingen vervolgens kunnen maken. Alle digitale activiteiten gerelateerd aan het oplossen van de oefeningen worden in de achtergrond bijgehouden en later door mij geanalyseerd. Indien een leerling gedetecteerd wordt, heeft de leerkracht verschillende opties, waaronder: geen interventie, een buur toewijzen om te helpen, interventie door de leerkracht, of zelfs een uitdagendere reeks oefeningen toewijzen. Het is niet de bedoeling dat de leerkracht de bevindingen van dit onderzoek gebruikt om leerlingen te beoordelen of te evalueren.

Wat gebeurt er met mijn gegevens?
De gegevens worden veilig bewaard en zijn alleen toegankelijk voor Anissa. Bij de registratie zal de leerkracht en leerling zijn/haar e-mailadres moeten ingeven. Deze wordt alleen gebruikt door mij om hem/haar te contacteren en zal dus ook verwijderd worden

hierna. Er worden verder geen persoonlijke gegevens gevraagd die gebruikt kunnen worden om de gebruiker direct te identificeren. De anonieme resultaten van het onderzoek worden alleen gebruikt voor wetenschappelijke doeleinden. Alles wordt gepubliceerd in een openbare thesistekst. Na de studie worden de resultaten bijgehouden op een beveiligde server van de KU Leuven.

Ben ik verplicht om deel te nemen aan het onderzoek?
Nee, je neemt volledig vrijwillig deel aan het onderzoek. Elk moment kan je de beslissing nemen om te stoppen met deelnemen. Hiervoor heb je ook geen reden nodig en er zal ook geen nadeel aan verbonden zijn.

Heb je nog vragen?
Indien er onduidelijkheden zijn of je nog vragen/feedback/opmerkingen hebt, kan je steeds contact opnemen met Anissa op anissa.faik@student.kuleuven.be.

A.2 Informed Consent Form

Geïnformeerde toestemming

Titel van het onderzoek: Bringing a New Perspective to the Classroom: Detecting and Explaining Student Outliers

Naam + contactgegevens [*e-mail, telefoonnummer, faculteit/departement/onderzoekseenheid, werkadres*] promotor en onderzoeker(s):

Faik Anissa (`anissa.faik@student.kuleuven.be`)

Ooge Jeroen (`jeroen.ooge@kuleuven.be`)

Verbert Katrien (`katrien.verbert@kuleuven.be`)

Duur van het experiment: Duur van een lesuur (50 minuten)

- Ik begrijp wat van mij verwacht wordt tijdens dit onderzoek.
- Ik weet dat ik zal deelnemen aan volgende proeven of testen:
  - Gebruik maken het e-learning platform met een ingebouwd detectiesysteem tijdens mijn les.
- Ikzelf of anderen kunnen baat bij dit onderzoek hebben op volgende wijze:
  - Gebruik maken van duizenden oefeningen gegeven door uitgeverij Die Keure.
- Ik weet dat er een beloning of compensatie gekoppeld is aan mijn deelname aan het onderzoek:
  - Gebruik maken van duizenden oefeningen gegeven door uitgeverij Die Keure.
- Ik begrijp dat mijn deelname aan deze studie vrijwillig is. Ik heb het recht om mijn deelname op elk moment stop te zetten. Daarvoor hoef ik geen reden te geven en

ik weet dat daaruit geen nadeel voor mij kan ontstaan.

- Voor de verdere verwerking van de verzamelde gegevens geldt het algemeen belang als rechtsgrond volgens de AVG/GDPR. Stopzetting van deelname aan de studie houdt dus in dat de eerder verzamelde gegevens nog verder rechtsgeldig kunnen worden betrokken in de studie en niet moeten worden verwijderd door KU Leuven.

- De resultaten van dit onderzoek kunnen gebruikt worden voor wetenschappelijke doeleinden en mogen gepubliceerd worden. Mijn naam wordt daarbij niet gepubliceerd. Doorheen het onderzoek zullen mijn gegevens steeds vertrouwelijk behandeld worden. Anonimiteit van de gegevens worden in elk stadium van het onderzoek gewaarborgd.

- Ik wil graag op de hoogte gehouden worden van de resultaten van dit onderzoek. De onderzoeker mag mij hiervoor contacteren op het volgende e-mailadres:

- Voor vragen evenals voor de uitoefening van mijn rechten (inzage gegevens, correctie ervan,...) weet ik dat ik na mijn deelname terecht kan bij:

  - *Faik Anissa* (*anissa.faik@student.kuleuven.be*)

  Meer informatie met betrekking tot privacy in onderzoek kan ik terugvinden op www.kuleuven.be/privacy. Verdere vragen over privacyaspecten kan ik richten tot de functionaris voor gegevensbescherming: `dpo@kuleuven.be`

- Deze studie werd beoordeeld en goedgekeurd door de Sociaal-Maatschappelijke Ethische Commissie (SMEC) van KU Leuven (G-2022-6129). Voor eventuele klachten of andere bezorgdheden omtrent ethische aspecten van deze studie kan ik contact opnemen met SMEC: `smec@kuleuven.be`

Ik heb bovenstaande informatie gelezen en begrepen en heb antwoord gekregen op al mijn vragen betreffende deze studie. Ik stem toe om deel te nemen.

Datum:

Naam en handtekening proefpersoon:

Naam en handtekening onderzoeker:

Appendix B

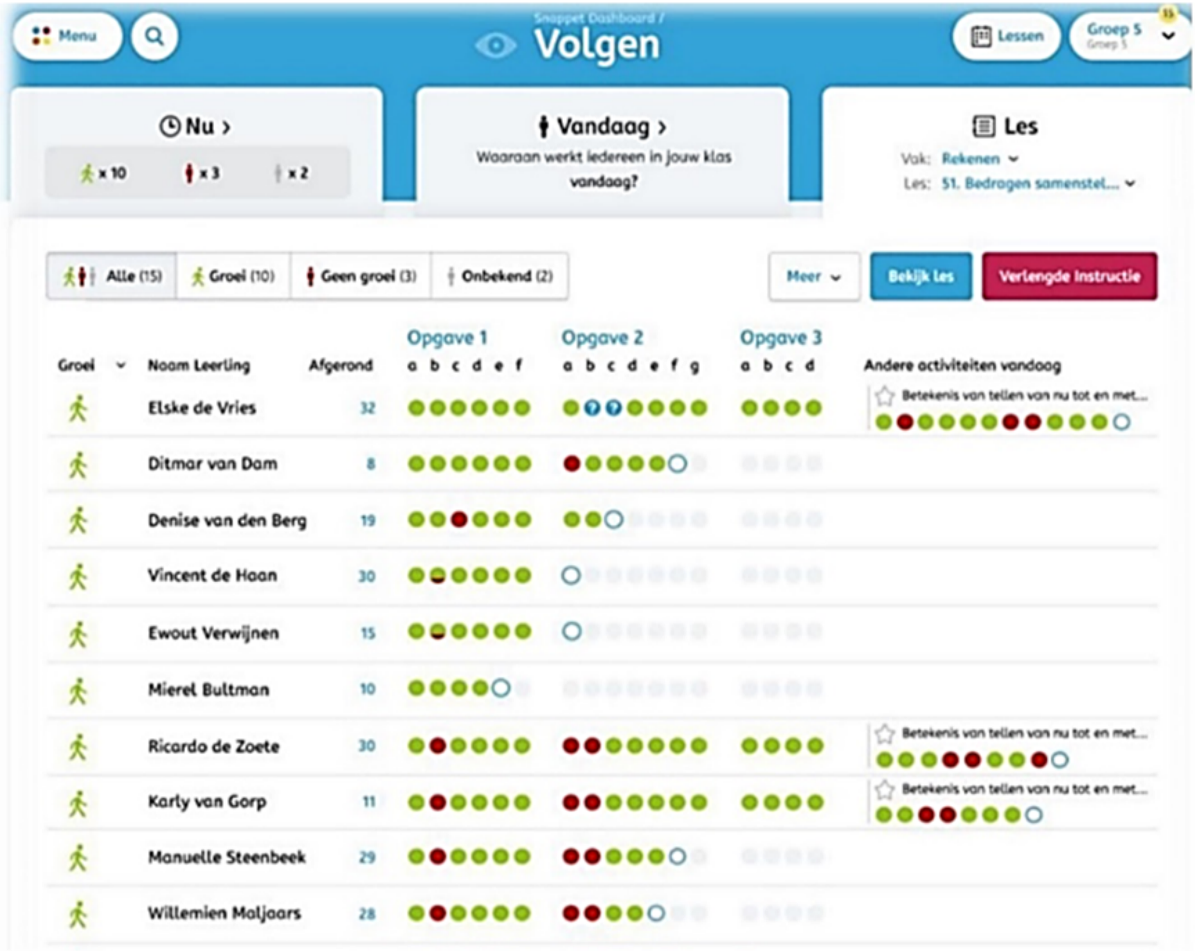
Pilot Study

B.1 General Questions

Table B.1: General questions asked during pilot study

English	**Dutch**
Do you use digital platforms or learning environments in your class?	Gebruik je digitale platformen of leeromgevingen in jouw les?
Do you use them to assess whether a student needs more support?	Gebruik jij deze om in te schatten of een leerling meer ondersteuning nodig heeft?
How do you determine if a student is falling behind or progressing ahead of the class?	Hoe beoordeel jij dat een leerling achteruit- of vooruitloopt op de klas?
What interventions do you use in such cases?	Welke interventies zet je hiervoor in?

B.2 Dashboards



Waaraan werkt iedereen in jouw klas vandaag?

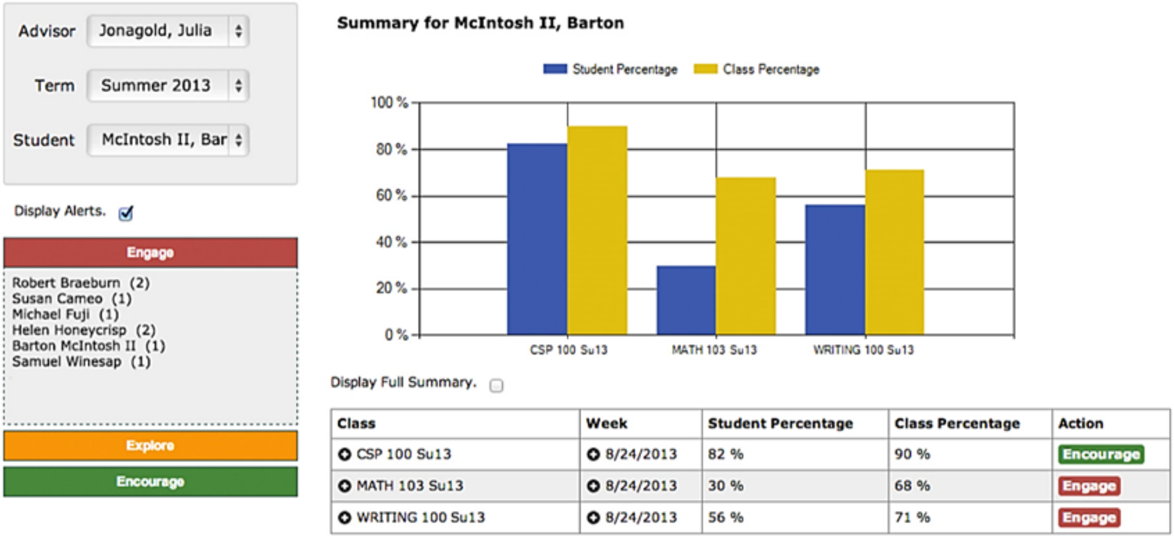
Vak: Rekenen

Les: 51. Bedragen samenstel...

Alle (15) | Groei (10) | Geen groei (3) | Onbekend (2)

Groei	Naam Leerling	Afgerond	Andere activiteiten vandaag
	Elske de Vries	32	Betekenis van tellen van nu tot en met...
	Ditmar van Dam	8	
	Denise van den Berg	19	
	Vincent de Haan	30	
	Ewout Verwijnen	15	
	Mierel Bultman	10	
	Ricardo de Zoete	30	Betekenis van tellen van nu tot en met...
	Karly van Gorp	11	Betekenis van tellen van nu tot en met...
	Manuelle Steenbeek	29	
	Willemien Maljaars	28	

Figure B.1: Dashboard taken from[66].



Advisor: Jonagold, Julia

Term: Summer 2013

Student: McIntosh II, Bar

Display Alerts. ☑

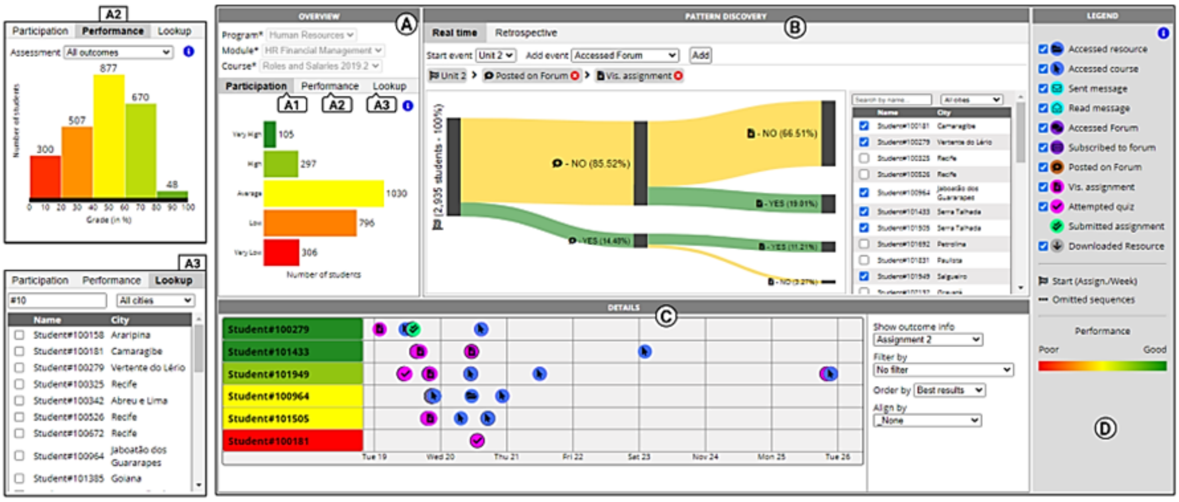
Robert Braeburn (2)
Susan Cameo (1)
Michael Fuji (1)
Helen Honeycrisp (2)
Barton McIntosh II (1)
Samuel Winesap (1)

Summary for McIntosh II, Barton

Display Full Summary. □

Class	Week	Student Percentage	Class Percentage	Action
CSP 100 Su13	8/24/2013	82 %	90 %	Encourage
MATH 103 Su13	8/24/2013	30 %	68 %	Engage
WRITING 100 Su13	8/24/2013	56 %	71 %	Engage

Figure B.2: Dashboard taken from[67].



Program* Human Resources
Module* HR Financial Management
Course* Roles and Salaries 2019.2

Start event Unit 2   Add event Accessed Forum

Unit 2 > Posted on Forum > Vis. assignment

#10   All cities

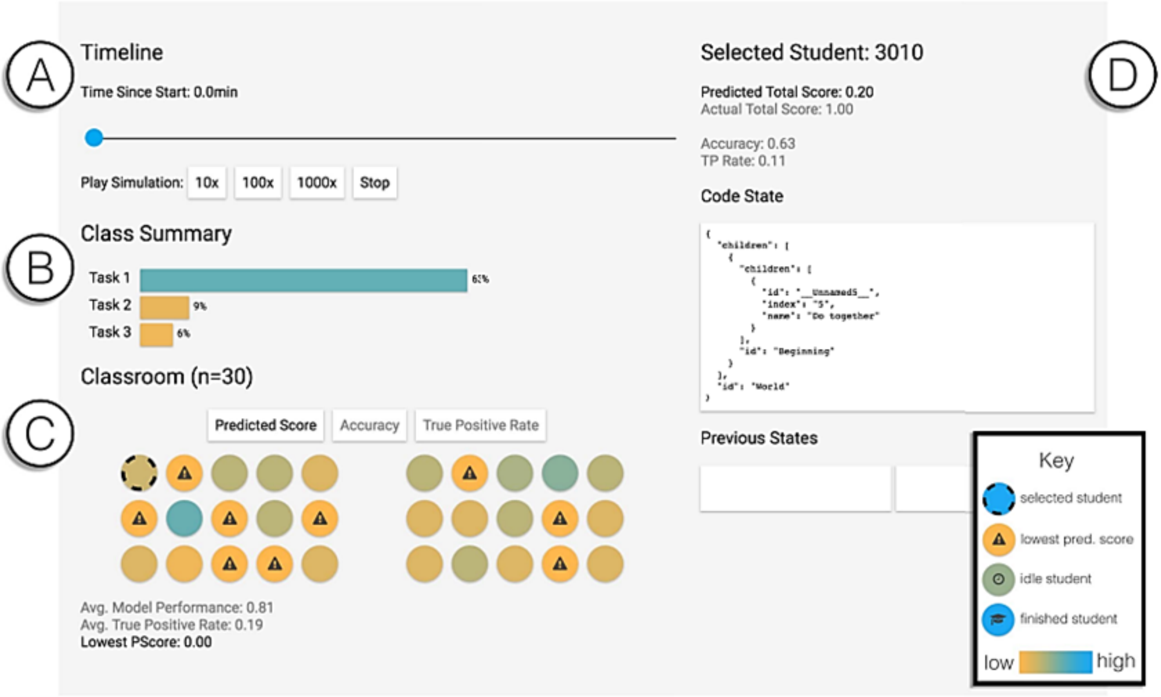
	Name	City
□	Student#100158	Araripina
□	Student#100181	Camaragibe
□	Student#100279	Vertente do Lério
□	Student#100325	Recife
□	Student#100342	Abreu e Lima
□	Student#100526	Recife
□	Student#100672	Recife
□	Student#100964	Jaboatão dos Guararapes
□	Student#101385	Goiana

	Name	City
☑	Student#100181	Camaragibe
☑	Student#100279	Vertente do Lério
□	Student#100325	Recife
□	Student#100526	Recife
☑	Student#100964	Jaboatão dos Guararapes
☑	Student#101433	Serra Talhada
☑	Student#101505	Serra Talhada
□	Student#101692	Petrolina
□	Student#101831	Paulista
☑	Student#101949	Salgueiro

Student#100279
Student#101433
Student#101949
Student#100964
Student#101505
Student#100181

Show outcome info Assignment 2
Filter by No filter
Order by Best results
Align by _None

Figure B.3: Dashboard taken from[68].



Timeline

Time Since Start: 0.0min

Class Summary

Classroom (n=30)

Avg. Model Performance: 0.81
Avg. True Positive Rate: 0.19
Lowest PScore: 0.00

Selected Student: 3010

Predicted Total Score: 0.20
Actual Total Score: 1.00

Accuracy: 0.63
TP Rate: 0.11

Code State

```
{
  "children": [
    {
      "children": [
        {
          "id": "__Unnamed5__",
          "index": "5",
          "name": "Do together"
        }
      ],
      "id": "Beginning"
    }
  ],
  "id": "World"
}
```

Previous States

Figure B.4: Dashboard taken from[69].

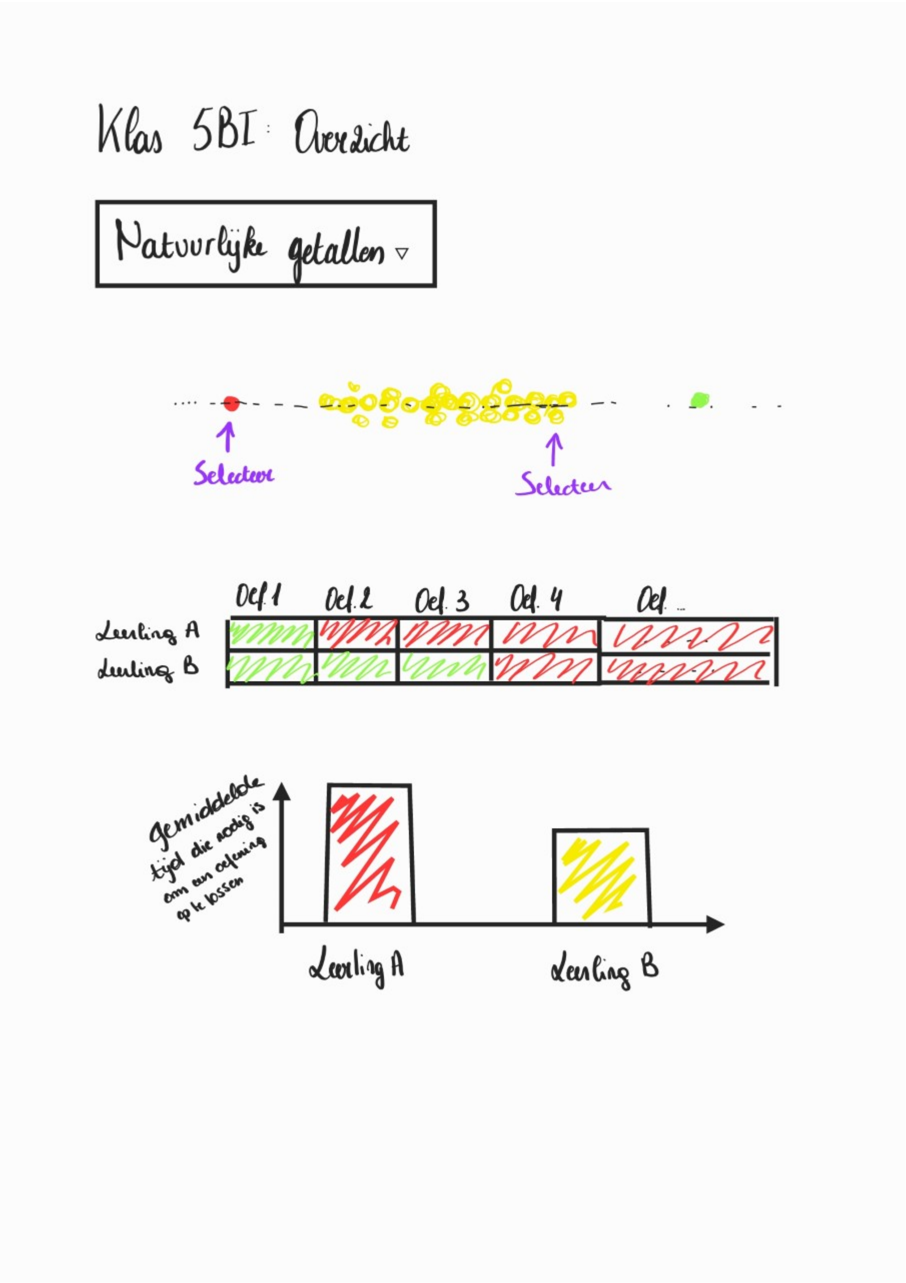


Figure B.5: Initial sketch made by author.

Appendix C

Think-aloud studies

C.1 Questions/Tasks Asked During the First and Second Think-aloud Study

Table C.1: Tasks for creating sets of exercises.

Compose a set of exercises.	**Oefeningenreeks samenstellen**
a. Compile a series of exercises for 5 Business & IT on the topic of “Natural Numbers,” with the name “Set 1.” Include exercises 1 and 2.	a. Stel zelf een reeks oefeningen samen voor 5 Business & IT. De oefeningen zijn van het thema “Natuurlijke getallen” en hebben als naam “Reeks 1”. Voeg de volgende oefeningen toe: oefening 1 en 2.
b. Create a series of random composite exercises on the topic of “Natural Numbers” with the name “Set 2.”	b. Maak een reeks aan met willekeurige samengestelde oefeningen voor het thema “Natuurlijke getallen” met als naam “Reeks 2”.

Table C.2: Tasks and questions related to student follow-up.

Follow-up students.	**Opvolging**
a. Follow up with the students of 5 Business & IT for the series of exercises on the topic of natural numbers, specifically for the exercise series of the lesson on 23/09.	a. Volg de leerlingen van 5 Business & IT op voor de reeks oefeningen voor het thema Natuurlijke getallen en specifiek voor de reeks oefening van de les op 23/09.
b. What information can you gather from this overview, and how would you use it to determine which students need extra attention?	b. Welke informatie kun je uit dit overzicht halen en hoe zou je deze gebruiken om te bepalen welke leerlingen extra aandacht nodig hebben?

Table C.3: Tasks and questions related to the histograms.

Histograms	**Histogrammen**
a. Is there any way to see how Maarten De Bakker performs compared to other students?	a. Kun je ergens zien hoe Maarten De Bakker presteert in vergelijking met andere leerlingen?
b. Is there any way to see how much time Maarten De Bakker takes to solve exercises compared to other students?	b. Kun je ergens zien hoeveel tijd Maarten De Bakker nodig heeft om oefeningen op te lossen in vergelijking met andere leerlingen?
c. Is there any way to see how many attempts Maarten De Bakker takes to solve exercises compared to other students?	c. Kun je ergens zien hoeveel pogingen Maarten De Bakker nodig heeft om oefeningen op te lossen in vergelijking met andere leerlingen?
d. How would you use these graphs to determine which students need extra attention?	d. Hoe zou je deze grafieken gebruiken om te bepalen welke leerlingen extra aandacht nodig hebben?

Table C.4: Tasks and questions related to timelines.

View the timeline of the first student.	**Bekijk de tijdslijn van de eerste leerling.**
a. Can you find out how many attempts Maarten De Bakker needed to solve exercise 1 correctly? And how much time was needed for that?	a. Kun je terugvinden hoeveel pogingen Maarten de Bakker nodig had om oefening 1 correct op te lossen? En hoe veel tijd was daarvoor nodig?
b. What patterns do you see, and how would you use this information to help guide the student?	b. Welke patronen zie je en hoe zou je deze informatie gebruiken om te helpen bij de begeleiding van de leerling?

Table C.5: Tasks and questions related to the interventions.

Interventions	**Interventies**
a. Is there any way to see which students require attention?	a. Kun je ergens zien welke leerlingen aandacht vereisen?
b. Report that a student who is falling behind in the class will be helped by a classmate.	b. Meldt dat je een leerling dat achteruitloopt op de klas laat helpen door een klasgenoot.
c. Assign a more challenging exercise series to a student who is ahead of the class.	c. Wijs een uitdagendere reeks oefening toe aan een leerling dat vooruitloopt op de klas.
d. What do you notice after you have implemented the intervention?	d. Wat valt er jou op na dat je de interventie hebt ingezet?
e. The alert remains. Should it disappear after implementing an intervention, or would you prefer to keep it until the student catches up?	e. Het alert blijft staan, is dat iets dat zou moeten verdwijnen na het inzetten van een interventie of zie je het liever staan zolang de leerling achterloopt?

Table C.6: Tasks and questions related to the filtering options.

Filtering options	**Filteropties**
a. Filter to show information only for exercises 1, 2, and 3. What changes on the screen?	a. Filter zodat enkel informatie voor oefening 1, 2 en 3 getoond wordt. Wat verandert er aan het scherm?
a. Filter to only show the detected students. What changes on the screen? (Only for second Think-aloud study.)	a. Filter zodat enkel gedetecteerde leerlingen worden getoond. Wat verandert er aan het scherm? (Enkel voor tweede think-aloud studie.)

Table C.7: General questions

General questions	**Algemene vragen**
What do you think of the dashboard? Are there any improvements you would suggest?	Wat vind je van het dashboard? Zijn er verbeteringen die je zou voorstellen?
What information do you find useful for monitoring students? Are there any other data or insights you would like to see?	Welke informatie vindt je nuttig bij het opvolgen van leerlingen? Zijn er andere gegevens of inzichten je zou willen zien?

Appendix D

Final Proof of Concept

D.1 Model-Centric Explanations: Translations

Figure 3.8 translate to the following: "Explanation of the detection system (1/3) On this page, I will explain how the detection system works, which is capable of determining which students may be struggling with the exercises and which students are performing very well. How does the algorithm work? We use an intelligent system that calculates a deviation score for each student based on two factors: the attempt score and the speed score. A high deviation score means that the student significantly deviates from the normal performance of the group, while a low deviation score means that the student has similar performance to the rest of the group."

Figure 3.9 translate to the following: "Attempt score: The attempt score is determined by the number of attempts a student has made to solve an exercise. This means that we consider both their successes and failures when answering questions. When a student answers a question correctly, their attempt score increases. However, if they require multiple attempts to answer the question correctly, their score will increase at a slower rate. If a student answers an exercise incorrectly, their attempt score decreases. The amount deducted depends on the number of times the student has previously attempted to solve the exercise. If the student has attempted it many times before, their attempt score will decrease more than if it is their first or second attempt. This approach aims to prevent students from simply guessing."

Figure 3.10 translate to the following: "Speed score: The speed score is based on the time a student takes to submit an attempt for an exercise. We compare this time with the average of all students and calculate a score. This allows us to determine whether a student is relatively fast or slow compared to the rest of the group."

Appendix E

Evaluation

E.1 Interview Questions

Table E.1: General questions during introduction

English	Dutch
In what order did you see the dashboard?	In welke volgorde heb jij het dashboard gezien?
What were your first impressions of the detection system?	Wat waren jouw eerste indrukken van het detectiesysteem?
How did you use it during the practice session?	Hoe heb je het gebruikt tijdens de oefensessie?
Did you receive different feedback from students compared to a typical lesson?	Kreeg je van leerlingen andere respons dan tijdens een gebruikelijke les?
Which insights did you find most interesting and why?	Welke inzichten vond je het interessantstie en waarom?
Which insights encourage you to use interventions?	Welke inzichten moedigen jou aan om interventies in te zetten?
Which insights do you feel were missing?	Welke inzichten miste volgens jou nog?

Table E.2: General question on trust

English	Dutch
Do you feel that you can trust the detection system?	Heb je het gevoel dat je het detectiesysteem kunt vertrouwen?

Table E.3: Questions on perceived reliability

English	Dutch
Can you describe how reliable you find the detection system?	Kun je beschrijven hoe betrouwbaar je het detectiesysteem vindt?
According to you, what are important indicators of the reliability of the system?	Wat zijn volgens jou belangrijke indicatoren van de betrouwbaarheid van het systeem?
To what extent do you trust that the system functions correctly?	In hoeverre vertrouw je erop dat het systeem correct functioneert?

Table E.4: Questions on perceived understandability

English	Dutch
Can you tell me how the detection system works?	Kun je me vertellen hoe het detectiesysteem werkt?
Where did you find information about how the detection system works?	Waar heb je informatie gevonden over hoe het detectiesysteem werkt?
Would you like more detailed explanations about how the detection system works? If so, what kind of information would you like to have?	Zou je graag meer gedetailleerde uitleg hebben over hoe het detectiesysteem werkt? Zo ja, wat voor soort informatie zou je graag willen hebben?
Can you describe how the detection system helps you make decisions and whether this is understandable for you?	Kun je beschrijven hoe het detectiesysteem je helpt bij het nemen van beslissingen en of dit voor jou begrijpelijk is?
What do you find easy or difficult to understand about the behavior of the detection system?	Wat vind je gemakkelijk of moeilijk te begrijpen aan het gedrag van het detectiesysteem?

Table E.5: Questions on perceived accuracy

English	Dutch
How accurate do you find the detections of the system?	Hoe nauwkeurig vind je de detecties van het systeem?
To what extent do you rely on the accuracy of the system and why?	In hoeverre vertrouw je op de nauwkeurigheid van het systeem en waarom?

Table E.6: Question on faith

English	Dutch
To what extent do you have faith in the detections of the system, even if you are not sure if they are correct?	In hoeverre heb je vertrouwen in de detecties van het systeem, zelfs als je niet zeker weet of ze correct zijn?

Table E.7: Questions on satisfaction

English	**Dutch**
Do you understand how the detection system works based on the explanations?	Begrijp je hoe het detectiesysteem werkt op basis van de uitleg?
Are you satisfied with the explanations about the detection system?	Ben je tevreden met de uitleg over het detectiesysteem?
Does the explanation help you evaluate when you can trust the detection system?	Helpt de uitleg je om te beoordelen wanneer je het detectiesysteem kunt vertrouwen?

Table E.8: Questions on effectivity

English	**Dutch**
Does the explanation help you know how to use the detection system to achieve your goals?	Helpt de uitleg je om te weten hoe je het detectiesysteem kunt gebruiken om je doelen te bereiken?

Table E.9: General follow-up questions

English	**Dutch**
Have you read the general explanation about the detection system? Does this help you understand the detection system?	Heb je de algemene uitleg over het detectiesysteem gelezen? Helpt dit bij het begrijpen van het detectiesysteem?
What did you generally think of the dashboard?	Wat vond je in het algemeen van het dashboard?

Bibliography

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6: 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281 Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk

[3] Towards A Rigorous Science of Interpretable Machine Learning [1702.08608] Towards A Rigorous Science of Interpretable Machine Learning (arxiv.org), Finale Doshi-Velez, Been Kim

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine 40, 2 (June 2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[6] Explainable Artificial Intelligence (XAI), David Gunning, National-Security-Archive-David-Gunning-DARPA.pdf (gwu.edu)

[7] Michael Hind. 2019. Explaining explainable AI. XRDS: Crossroads, The ACM Magazine for Students 25, 3: 16–19. https://doi.org/10.1145/3313096

[8] Explainable Artificial Intelligence in education https://doi.org/10.1016/j.caeai.2022.100074

[9] Salvador Garcia-Martinez and Abdelwahab Hamou-Lhadj. 2013. Educational Recommender Systems: A Pedagogical-Focused Perspective. In Multimedia Services in Intelligent Environments, George A. Tsihrintzis, Maria Virvou, and Lakhmi C. Jain (Eds.). Vol. 25. Springer International Publishing, Heidelberg, 113–124. https://doi.org/10.1007/978-3-319-00375-7_8

[10] Bull, S., Kay, J. (2010). Open Learner Models. In: Nkambou, R., Bourdeau, J., Mizoguchi, R. (eds) Advances in Intelligent Tutoring Systems. Studies in Computational

Intelligence, vol 308. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14363-2_15

[11] Robert Bodily, Judy Kay, Vincent Aleven, Ioana Jivet, Dan Davis, Franceska Xhakaj, and Katrien Verbert. 2018. Open learner models and learning analytics dashboards: a systematic review. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK '18). Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/3170358.3170409

[12] Jordan Barria-Pineda. 2020. Exploring the Need for Transparency in Educational Recommender Systems. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20). Association for Computing Machinery, New York, NY, USA, 376–379. https://doi.org/10.1145/3340631.3398676

[13] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in Education Needs Interpretable Machine Learning: Lessons from Open Learner Modelling. arXiv:1807.00154 [cs] (June 2018). arXiv:1807.00154 [cs] https://arxiv.org/pdf/1807.00154.pdf

[14] Khosravi, H., Kitto, K., & Williams, J. J. (2019). RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities. Journal of Learning Analytics, 6(3), 91–105. https://doi.org/10.18608/jla.2019.63.12

[15] Cambridge advanced learner's dictionary E. Walter Cambridge University Press (2008)

[16] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact. Intell. Syst. 11, 3–4, Article 24 (December 2021), 45 pages. https://doi.org/10.1145/3387166

[17] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. Comput. Surveys 51, 5 (Jan. 2019), 1–42. https://arxiv.org/abs/1802.01933

[18] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 579, 1–13. https://doi.org/10.1145/3290605.3300809

[19] Fatih Gedikli, Dietmar Jannach, Mouzhi Ge, How should I explain? A comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies, 2014, Pages 367-382, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2013.12.007.

[20] N. Tintarev and J. Masthoff, "A Survey of Explanations in Recommender Systems," 2007 IEEE 23rd International Conference on Data Engineering Workshop, 2007, pp. 801-810, doi: 10.1109/ICDEW.2007.4401070.

[21] Kelly Wauters, Piet Desmet, Wim Van Den Noortgate, Item difficulty estimation: An auspicious collaboration between data and judgment, Computers & Education, 2012, Pages 1183-1193, https://doi.org/10.1016/j.compedu.2011.11.020.

[22] Helma Torkamaan and Jürgen Ziegler. 2022. Recommendations as Challenges: Estimating Required Effort and User Ability for Health Behavior Change Recommendations. In 27th International Conference on Intelligent User Interfaces (IUI '22). Association for Computing Machinery, New York, NY, USA, 106–119. https://doi.org/10.1145/3490099.3511118

[23] Radek Pelánek, Applications of the Elo rating system in adaptive educational systems, 2016, Pages 169-179, https://doi.org/10.1016/j.compedu.2016.03.017.

[24] Antal, Margit. (2013). ON THE USE OF Elo RATING FOR ADAPTIVE ASSESSMENT. Studia Informatica. LVIII.

[25] Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2019). A multivariate Elo-based learner model for adaptive educational systems. https://arxiv.org/abs/1910.12581

[26] Rémi Domingues, Maurizio Filippone, Pietro Michiardi, Jihane Zouaoui, A comparative evaluation of outlier detection algorithms: Experiments and analyses,2018,Pages 406-421, https://doi.org/10.1016/j.patcog.2017.09.037

[27] Christophe Leys, Olivier Klein, Yves Dominicy, Christophe Ley, Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance, Journal of Experimental Social Psychology, 2018, Pages 150-156, https://doi.org/10.1016/j.jesp.2017.09.011.

[28] H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Outlier detection in axis-parallel subspaces of high dimensional data, in: Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27–30, 2009 Proceedings, Springer, 2009, pp. 831–838, doi:10.1007/978-3-642-01307-2_86.

[29] H.-P. Kriegel, M. S hubert, A. Zimek, Angle-based outlier detection in high dimensional data, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '08, ACM, 2008, pp. 444–452, doi:10.1145/1401890.1401946.

[30] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation forest, in: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, in: ICDM '08, IEEE Computer Society, 2008, pp. 413–422, doi:10.1109/ICDM.2008.17

[31] U. Porwal and S. Mukund, "Credit card fraud detection in e-commerce: An outlier detection approach," 2018, arXiv:1811.02196. [Online]. Available: https://arxiv.org/abs/1811.02196

[32] G. Gebremeskel, C. Yi, Z. He, and D. Haile, "Combined data mining techniques based patient data outlier detection for healthcare safety," Int.J. Intell. Comput. Cybern., vol. 9, no. 1, pp. 42–68, 2016.

[33] Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2020. Development and Adop-

tion of an Adaptive Learning System: Reflections and Lessons Learned. In Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20). Association for Computing Machinery, New York, NY, USA, 58–64. https://doi.org/10.1145/3328778.3366900

[34] Kaliisa, R., Dolonen, J.A. CADA: a teacher-facing learning analytics dashboard to foster teachers' awareness of students' participation and discourse patterns in online discussions. Tech Know Learn (2022). https://doi.org/10.1007/s10758-022-09598-7

[35] Denden, M. et al. (2019). iMoodle: An Intelligent Gamified Moodle to Predict "at-risk" Students Using Learning Analytics Approaches. In: Tlili, A., Chang, M. (eds) Data Analytics Approaches in Educational Games and Gamification Systems. Smart Computing and Intelligence. Springer, Singapore. https://doi.org/10.1007/978-981-32-9335-9_6

[36] Bull, S., Kay, J. (2010). Open Learner Models. In: Nkambou, R., Bourdeau, J., Mizoguchi, R. (eds) Advances in Intelligent Tutoring Systems. Studies in Computational Intelligence, vol 308. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14363-2_15

[37] Meng, F., Yuan, G., Lv, S. et al. An overview on trajectory outlier detection. Artif Intell Rev 52, 2437–2456 (2019). https://doi.org/10.1007/s10462-018-9619-1

[38] Knox, E. M., Ng, R. T. (1998, August). Algorithms for mining distance based outliers in large datasets. In Proceedings of the international conference on very large data bases (pp. 392-403). Citeseer.

[39] Bo Tang, Haibo He, A local density-based approach for outlier detection, Neurocomputing, Volume 241, 2017, Pages 171-180, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2017.02.039.

[40] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00). Association for Computing Machinery, New York, NY, USA, 93–104. https://doi.org/10.1145/342009.335388

[41] MANSUR, M. O.; SAP, Md; NOOR, Mohd. Outlier detection technique in data mining: a research perspective. In: Postgraduate Annual Research Seminar. CMS Press, 2005. p. 23-31.

[42] Rienties, Bart; Cross, Simon and Zdrahal, Zdenek (2016). Implementing a Learning Analytics Intervention and Evaluation Framework: what works? In: Kei Daniel, Ben and Butson, Russell eds. Big Data and Learning Analytics in Higher Education: Current Theory and Practice. Heidelberg: Springer, pp. 147–166.

[43] J. Ooge. Het personaliseren van motivationele strategieën en gamificationtechnieken m.b.v. recommendersystemen. 2019

[44] Adams, William. (2015). Conducting Semi-Structured Interviews. 10.1002/9781119171386.ch19.

[45] Hans-Peter Kriegel, Peer Kroger, Erich Schubert and Arthur Zimek. Interpreting and Unifying Outlier Scores. Proceedings of the 2011 SIAM International Conference on Data Mining (SDM). 13-24. https://epubs.siam.org/doi/abs/10.1137/1.9781611972818.2

[46] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In 27th International Conference on Intelligent User Interfaces (IUI '22). Association for Computing Machinery, New York, NY, USA, 93–105. https://doi.org/10.1145/3490099.3511140

[47] Jeroen Ooge, Leen Dereu, and Katrien Verbert. 2023. Steering Recommendations and Visualising Its Impact: Effects on Adolescents' Trust in E-Learning Platforms. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23). Association for Computing Machinery, New York, NY, USA, 156–170. https://doi.org/10.1145/3581641.3584046

[48] J. Ooge and K. Verbert, "Trust in Prediction Models: a Mixed-Methods Pilot Study on the Impact of Domain Expertise," 2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX), New Orleans, LA, USA, 2021, pp. 8-13, doi: 10.1109/TREX53765.2021.00007.

[49] Madsen, Maria, and Shirley Gregor. "Measuring human-computer trust." 11th australasian conference on information systems. Vol. 53. Australasian Association for Information System, 2000.

[50] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User Trust in Intelligent Systems: A Journey Over Time. In Proceedings of the 21st International Conference on Intelligent User Interfaces. ACM, Sonoma California USA, 164–168. https://doi.org/10.1145/2856767.2856811

[51] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (Oct. 2020), 112–121.

[52] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). Association for Computing Machinery, New York, NY, USA, 397-407. https://doi.org/10.1145/3301275.3302313

[53] Hoffman, Robert R., et al. "Metrics for explainable AI: Challenges and prospects." arXiv preprint arXiv:1812.04608 (2018).

[54] Arissa J. Sato, Zefan Sramek, and Koji Yatani. 2023. Groupnamics: Designing an Interface for Overviewing and Managing Parallel Group Discussions in an Online Classroom. In Proceedings of the 2023 CHI Conference on Human Factors in Computing

Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 701, 1–18. https://doi.org/10.1145/3544548.3581322

[55] LI, Zhong; ZHU, Yuxuan; VAN LEEUWEN, Matthijs. A Survey on Explainable Anomaly Detection. arXiv preprint arXiv:2210.06959, 2022.

[56] Barbora Micenková, Raymond T Ng, Xuan-Hong Dang, and Ira Assent. 2013. Explaining outliers by subspace separability. In 2013 IEEE 13th international conference on data mining. IEEE, 518–527

[57] Nguyen Xuan Vinh, Jeffrey Chan, Simone Romano, James Bailey, Christopher Leckie, Kotagiri Ramamohanarao, and Jian Pei. 2016. Discovering outlying aspects in large datasets. Data mining and knowledge discovery 30, 6 (2016), 1520–1555.

[58] Xuan Hong Dang, Barbora Micenková, Ira Assent, and Raymond T Ng. 2013. Local outlier detection with interpretation. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 304–320.

[59] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I trust my machine teammate? an investigation from perception to decision. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[60] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK. ACM, New York, NY, USA 13 Pages. https://doi.org/10.1145/3290605.3300509

[61] Nourani, M., Kabir, S., Mohseni, S., Ragan, E. D. (2019). The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7(1), 97-105. https://doi.org/10.1609/hcomp.v7i1.5284

[62] Lopes, Pedro, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. 2022. "XAI Systems Evaluation: A Review of Human and Computer-Centred Methods" Applied Sciences 12, no. 19: 9423. https://doi.org/10.3390/app12199423

[63] Ribera, M., Lapedriza, A. (2019, March). Can we do better explanations? A proposal of user-centered explainable AI. In IUI workshops (Vol. 2327, p. 38).

[64] Chounta, IA., Bardone, E., Raudsep, A. et al. Exploring Teachers' Perceptions of Artificial Intelligence as a Tool to Support their Practice in Estonian K-12 Education. Int J Artif Intell Educ 32, 725–755 (2022). https://doi.org/10.1007/s40593-021-00243-5

[65] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074.

[66] Knoop-van Campen, C., Molenaar, I. (2020). How Teachers integrate Dashboards into their Feedback Practices. Frontline Learning Research, 8(4), 37–51. https://doi.org/10.14786/flr.v8i4.641

[67] Steven Lonn, Stephen J. Aguilar, Stephanie D. Teasley, Investigating student motivation in the context of a learning analytics intervention during a summer bridge program, Computers in Human Behavior, 2015, Pages 90-97, https://doi.org/10.1016/j.chb.2014.07.013.

[68] Steven Lonn, Stephen J. Aguilar, Stephanie D. Teasley, Investigating student motivation in the context of a learning analytics intervention during a summer bridge program, Computers in Human Behavior, 2015, Pages 90-97, https://doi.org/10.1016/j.chb.2014.07.013.

[69] Nicholas Diana, Michael Eagle, John Stamper, Shuchi Grover, Marie Bienkowski, and Satabdi Basu. 2017. An instructor dashboard for real-time analytics in interactive programming assignments. In Proceedings of the Seventh International Learning Analytics amp; Knowledge Conference (LAK '17). Association for Computing Machinery, New York, NY, USA, 272–279. https://doi.org/10.1145/3027385.3027441

[70] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 75, 1–13. https://doi.org/10.1145/3411764.3445736

[71] R. Matulionyte and A. Hanif, "A call for more explainable AI in law enforcement," 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), Gold Coast, Australia, 2021, pp. 75-80, doi: 10.1109/EDOCW52865.2021.00035.

[72] Ooge, J., Stiglic, G., Verbert, K. (2022). Explaining artificial intelligence with visual analytics in healthcare. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(1), e1427.

[73] T. Spinner, U. Schlegel, H. Schäfer and M. El-Assady, "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning," in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 1064-1074, Jan. 2020, doi: 10.1109/TVCG.2019.2934629.

[74] Spinner, T., Schlegel, U., Schäfer, H., El-Assady, M. (2019). explAIner: A visual analytics framework for interactive and explainable machine learning. IEEE transactions on visualization and computer graphics, 26(1), 1064-1074.

[75] N. Andrienko, G. Andrienko, L. Adilova and S. Wrobel, "Visual Analytics for Human-Centered Machine Learning," in IEEE Computer Graphics and Applications, vol. 42, no. 1, pp. 123-133, 1 Jan.-Feb. 2022, doi: 10.1109/MCG.2021.3130314.

[76] A. Bussone, S. Stumpf and D. O'Sullivan, "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems," 2015 International Con-

ference on Healthcare Informatics, Dallas, TX, USA, 2015, pp. 160-169, doi: 10.1109/ICHI.2015.26.

[77] J.D. Lee, and K.A. See, ”Trust in Automation: Designing for Appropriate Reliance,” Human Factors: The Journal of Human Factors and Ergonomics Society, vol. 46(1), pp.50-80, 2004.

[78] P. Madhavan, D.A. Wiegmann, ”Similarities and differences between human-human and human-automation trust: an integrative review,” Theoretical Issues in Ergonomics Science, vol. 8(4), pp.277-301, 2007.

[79] Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. Human Factors, 44(1), 79–94. https://doi.org/10.1518/0018720024494856

[80] Nazaretsky, T., Cukurova, M., Ariely, M., Alexandron, G. (2021, September). Confirmation bias and trust: human factors that influence teachers' attitudes towards AI-based educational technology. In CEUR Workshop Proceedings (Vol. 3042).

[81] Satya M. Muddamsetty, Mohammad N.S. Jahromi, Andreea E. Ciontos, Laura M. Fenoy, Thomas B. Moeslund, Visual explanation of black-box model: Similarity Difference and Uniqueness (SIDU) method, Pattern Recognition, Volume 127, 2022, 108604, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2022.108604

[82] Timms, M.J. Letting Artificial Intelligence in Education Out of the Box: Educational Cobots and Smart Classrooms. Int J Artif Intell Educ 26, 701–712 (2016). https://doi.org/10.1007/s40593-016-0095-y

[83] Guerra-Hollstein, J., Barria-Pineda, J., Schunn, C. D., Bull, S., Brusilovsky, P. (2017, July). Fine-grained open learner models: Complexity versus support. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (pp. 41-49).

[84] Braun, Virginia Clarke, Victoria. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology. 3. 77-101. 10.1191/1478088706qp063oa.

AFDEL

Straat nr bus
3000 LEUVEN, BEL
tel. + 32 16 00 0
fax + 32 16 00 0
www.kuleuve