KU LEUVEN

FACULTEIT INGENIEURSWETENSCHAPPEN

Practicing the Right Math: Enhancing Trust in an E-Learning Platform Using an Explainable Recommender System

Shotallo Kato

Thesis voorgedragen tot het behalen van de graad van Master of Science in de ingenieurswetenschappen: computerwetenschappen, hoofdoptie Artificiële intelligentie

Promotor:
Prof. dr. K. Verbert

Assessoren:
Prof. dr. ir. G. Janssens
Dr. ir. F. Gutiérrez

Begeleider:
J. Ooge

Academiejaar 2020 – 2021

Preface

Before I started my thesis, I heard from various sources what my experience would be like.

"*One of the toughest experiences you will face in your life.*"
"*The most amount of hours you will ever need to work every week.*"
"*The least amount of sleep you will be able to get during your days at university.*"

And yes, it was all of this, and much more. However, *solitary*, it was *not*. I'm very fortunate for the amount of support I received to get me through this challenge. First, I would like to thank Jeroen Ooge for the immeasurable amount of support he provided me throughout the year. Whether it would be giving me ideas, proofreading my work, or any other form of help, he gladly guided me every step of the way. Without his help, I'm sure this master's thesis would look very different... and *not* in a positive way.
I would further like to thank those from the Augment research group, especially Professor Verbert, my supervisor, for their kind yet valuable feedback after each of my presentations. Also, a special thanks to Robin de Croon for assisting me with the deployment of my platform.
Thank you to all of my friends and family for also helping me along the way.
Of course, I cannot forget to thank all the participants of my think-aloud studies and user studies. In particular, I would like to thank the teachers that offered their valuable time in such a difficult period. Without all of you, this research would not have been possible.
Finally, a special thanks to Ellen for all of the emotional support and motivation she provided me throughout this challenging year. Any person would be fortunate to have you by their side.

Shotallo Kato

Contents

Abstract

User trust is a compelling goal in the field of Human-Computer Interaction due to the vast array of benefits it can provide, such as acceptance of technology. Increasing user trust has especially been sought after with explainable recommender systems. However, the literature comprises various research with university students, and research with participants outside of this age group is often hard to find. Furthermore, the context in which such systems are studied are usually the same. This begs the question of whether results from prior studies are generalizable to a different age group or application.

In this thesis, we wish to investigate the effects of using an explainable recommender system on initial user trust in an e-learning platform for high school students. We first augment an existing e-learning platform called Wiski with a recommender system. An accompanying explanation interface is designed with the specific goal to increase initial user trust. We further compare our interface to both a no-explanation baseline and a placebo explanation interface to test its effectiveness. Finally, we investigate whether there is a correlation between initial user trust and acceptance of recommendations to see whether acceptance of recommendations can potentially be used as an implicit measurement.

A user study with 37 high school students from Flanders was conducted. Our findings give us a somewhat two-sided conclusion, where using an explanation interface can only significantly increase initial user trust according to some measures. The results further show that the population of high school students is diverse and that the importance of explanations most likely differs depending on the end-user. We also find that using placebo explanations in user studies can potentially be interesting, as it provides qualitative information not obtainable by only using a no-explanation baseline. Finally, we observe little to no correlation between acceptance of recommendations and initial user trust, as most users accept the recommendations regardless of how much they trust the system.

Samenvatting

Vertrouwen (*trust*) is een belangrijke onderzoeksrichting binnen het domein van Mens-Machine Interactie door de verschillende voordelen die het kan opleveren. Voor het creëren van vertrouwen in de omgang met zowel artificiële intelligentie systemen en aanbevelingssystemen, is het belangrijk dat de acties te verklaren zijn (*explainable*). Zo kan een aanbevelingssysteem niet alleen de aanbeveling geven, maar ook uitleggen waarom het deze aanbeveling geeft. Verschillende bronnen geven aan dat de uitleg afhankelijk is van de eindgebruiker. Desondanks beperkt het huidige onderzoeksveld zich hoofdzakelijk tot een beperkte leeftijdsgroep. De huidige literatuur slaagt er dus niet in om een eenduidig antwoord te geven voor de leeftijdsgroep van jongeren onder de 18 jaar.
Verder focust de meerderheid zich slechts op een beperkt deel van de applicaties, zoals bijvoorbeeld het aanbevelen van films. Het levert dus de interessante vraag op of dezelfde resultaten gelden in andere toepassingsdomeinen, zoals e-learning.

In deze masterthesis onderzoeken wij het effect van het gebruik van een aanbevelingssysteem dat uitleg geeft (een zgn. *explainable recommender system*) om het initiële vertrouwen van leerlingen van de tweede en derde graad van het middelbaar onderwijs te verhogen in de context van een e-learningplatform. Deze thesis bouwt verder op een bestaand platform, Wiski, en voegt een aanbevelingssysteem toe. Verder ontwerpen we een interface die uitleg geeft over waarom de aanbevolen oefening goed bij het niveau van de gebruiker past (een zgn. *explanation interface*). Naast het vergelijken van ons systeem met een interface zonder uitleg (zoals de meeste literatuur), vergelijken wij onze interface ook met een *placebo explanation interface*. *Placebo explanations* kunnen worden gezien als (textuele) verklaringen die geen nuttige informatie geven aan de eindgebruiker. Recente literatuur stelt voor dat *placebo explanations* misschien een betere controlemethode kunnen zijn voor gebruikersstudies dan helemaal geen uitleg tonen. Ten slotte gaan we na of het aanvaarden van aanbevelingen, gemeten door *click-through rate*, een correlatie vertoont met het initiële vertrouwen in deze context. Zulke impliciete methodes om vertrouwen te meten kunnen nuttig zijn aangezien momenteel voor gelijkaardig onderzoek er voornamelijk gebruik wordt gemaakt van vragenlijsten.

Een gebruikersstudie met 37 leerlingen van de tweede en derde graad van het middelbaar onderwijs is uitgevoerd om de resultaten te verzamelen. De resultaten zijn niet eenduidig: de *explanation interface* slaagt erin om het initiële vertrouwen te

verhogen wanneer het gemeten wordt aan de hand van de verschillende *constructs* van vertrouwen, maar niet wanneer er expliciet naar gevraagd wordt. Een mogelijke verklaring is dat de nood voor een uitleg bij een aanbeveling minder belangrijk is in een e-learningplatform dan bijvoorbeeld een e-commerceplatform waar keuzes zwaarder kunnen doorwegen. Verder kunnen bijvoorbeeld hoe accuraat de gebruikers de website vinden, kwaliteit van de oefeningen of uiterlijk van de website misschien een even grote of zelfs een grotere rol spelen bij het beoordelen van het initiële vertrouwen voor het platform. Anderzijds kan het misschien ook een bewijs zijn dat vertrouwen meten aan de hand van de *constructs* een genuanceerder beeld kan geven in plaats van slechts met één vraag het vertrouwen te meten.
Verder merken we ook geen kwantitatief significant verschil tussen de *placebo explanation interface* en de interface zonder enige uitleg. We zien wel dat er leerlingen zijn die kwalitatief aangeven dat de *placebo explanation* voldoende is, terwijl sommige vinden dat het geen nuttige informatie geeft. *Placebo explanations* kunnen dus wel extra informatie geven in gebruikersstudies (e.g., "Is transparantie echt belangrijk voor gebruiker x?") t.o.v. geen verklaringen gebruiken.
Ten slotte, de thesis vindt geen correlatie tussen het initiële vertrouwen en het aanvaarden van aanbevelingen gemeten door *click-through rate*. We zien echter wel dat de groep die geen uitleg gekregen heeft minder vaak voor de aanbevolen oefeningen koos. Het kan dus zijn dat wanneer de gebruikers wel een verklaring krijgen, ze voor het gemak een aanbevolen oefening kiezen, ongeacht het vertrouwen in de website. Anderzijds wanneer de gebruikers niet weten waarom een oefening is aanbevolen, willen ze misschien zelf controle hebben over welke oefening ze willen maken in plaats van blindelings een aanbeveling te volgen.

List of Abbreviations and Symbols

Abbreviations

IRE	Interface for Real Explanation
IPE	Interface for Placebo Explanation
INE	Interface for No Explanation

Chapter 1

Introduction

It has become quite improbable not to have come into contact with recommender systems in today's age. Whether it would be when making purchases on Amazon or scrolling through endless options on Netflix, recommender systems have made their way into our daily lives. However, we are often left in the dark when it comes to why something has been recommended, leaving us only to make educated guesses. Therefore, accompanying recommendations with explanations and studying their influences have been popular topics of interest in the field of Human-Computer Interaction (HCI). Many researchers agree that using explanations can accomplish various goals, such as improving user trust [70, 54, 4]. However, most of the literature in this field uses recommender systems for media (e.g., movies) [10, 29, 52, 71] or e-commerce [64, 31, 9] as the basis for their research. The question thus remains whether previous results can be generalized to other domains. One application domain of particular interest is e-learning. State-of-the-art platforms such as ALEKS[1] and Knewton[2] use complex AI techniques to model the students' knowledge and curate content for them. There is also a plethora of literature concerning automatic adaptation techniques in e-learning. However, there is still quite a lot of room for work to be done concerning the transparency and explainability of such systems.

This thesis aims to research the effects of accompanying recommendations with explanations in the context of an e-learning platform for high school students. More specifically, we attempt to answer whether using explanations leads to higher initial user trust in an e-learning context with this user base.
First, an existing e-learning platform, *Wiski* [60], is augmented to include an automatic adaptation strategy. The platform attempts to recommend questions of the correct difficulty level to the end-user using an amalgamation of the Elo rating system and collaborative filtering [23]. Furthermore, building upon prior research, the effects of explanations and transparency on user trust are investigated to examine whether they hold for high school students in an e-learning context. For example, young users may not place as much importance on transparency as older, more mature users. A

[1] https://www.aleks.com/
[2] https://www.knewton.com/

randomized controlled experiment using both a no-explanation control group and a placebo explanation control group is conducted to examine the effectiveness of the developed explanation interface. Placebo explanations [25] are explanations that do not convey any useful information to the end-user. We further investigate the effects of such explanations to gain insight into whether they are helpful for user studies. Finally, we look into whether there exists a correlation between the acceptance of recommendations, measured through click-through rate, and initial user trust. Such a correlation could lead to using the acceptance of a recommendation as an implicit measurement for initial user trust in user studies instead of traditional methods that require questionnaires.

What this thesis does not aim to answer is, for example, "What algorithm leads to the best accuracy in an e-learning application." Furthermore, although an explanation interface is designed and evaluated (against a no-explanation interface and a placebo explanation interface), the thesis does not compare and contrast multiple "real" explanation interfaces to determine which amongst them is the most effective.

The thesis is unique as it addresses the following gaps in the literature:

- The target audience consists purely of high school students, whereas most of the literature comprises university students or young adults. We also use an explainable recommender system in an e-learning context as opposed to most of the traditional applications such as movie recommender systems.
- We compare our explanation interface to a placebo explanation baseline in addition to a no-explanation baseline. Placebo explanations [25] are yet to be widely applied in the literature. Their effects and usefulness are thus not fully determined.
- We investigate whether a correlation exists between initial user trust and acceptance of recommendations (measured through click-through rate) to see whether it could be used as an implicit measure. Such implicit measures for trust are scarce in the literature for explainable recommender systems.

The thesis is structured as follows. We first take a look at the literature, covering background information and related work in Chapter 2. Next, we discuss the methodology used for the research in Chapter 3. Chapter 4 addresses the development process of the platform, Wiski. Afterward, we present the results of the research in Chapter 5 and consequently discuss them, as well as outline the limitations of the work in Chapter 6. Finally, we conclude the thesis in Chapter 7, where we summarize our findings and open a discussion for possible future work.

Chapter 2

Literature Study

2.1 Adaptation Techniques

There exists a myriad of automatic adaptation techniques for e-learning. A quick look through the literature highlights approaches such as *Item Response Theory*, the *Elo Rating System*, and traditional recommender system algorithms such as *content-based* and *collaborative filtering*. More recent approaches include complex machine learning tools such as neural networks. This section discusses the relevant adaptation techniques for this thesis.

2.1.1 Recommender Systems

Recommender systems have been successfully implemented in various commercial applications, such as Amazon and Netflix. Traditionally, recommender system algorithms can be split into three categories: *content-based filtering*, *collaborative filtering*, and *hybrid methods*. Apart from these algorithms, one can find other algorithms such as knowledge-based algorithms and context-aware algorithms. Collaborative filtering is the focus of this section as it is used in the thesis. The following information concerning collaborative filtering is based on and adapted from Aggarwal's book [5].

Collaborative filtering leverages the users' or items' past data to make recommendations. The algorithm builds upon one fundamental assumption: ratings of the past form a representation of the ratings given in the future. Collaborative filtering can primarily be split into two categories: *memory-based collaborative filtering* and *model-based collaborative filtering*. Only the former is discussed as that is the type that is used in this thesis. Memory-based collaborative filtering (also known as *neighborhood-based collaborative filtering* [5]) leverages the similarity of past ratings of the users or items to make predictions. Central to memory-based methods are *similarity measures*. One can choose to either base the predictions of an unknown item on the similarity between items (*item-based*) or users (*user-based*). Take for example the scenario of predicting the rating user U will give movie M. The prediction can either be based upon the ratings that similar users as U have given M, or the ratings that U has given to movies similar to M in the past. Typically, the *cosine*

similarity or *Pearson correlation* is used. Given two users A and B, the cosine similarity of A and B is defined as

$$\text{sim}(A, B) = \frac{\sum_i A_i * B_i}{\sqrt{\sum_i A_i^2} * \sqrt{\sum_i B_i^2}} \tag{2.1}$$

where A_i and B_i are the ratings given by the respective users. The Pearson correlation for two users A and B is defined as

$$\text{sim}(A, B) = \frac{\sum_i (A_i - \mu_A) * (B_i - \mu_B)}{\sqrt{\sum_i (A_i - \mu_A)^2} * \sqrt{\sum_i (B_i - \mu_B)^2}} \tag{2.2}$$

where μ_A and μ_B are the mean ratings the users have given and $i \in A \cap B$ (items that both users have rated).
The subtle difference between the two similarity functions comes in the form of a subtraction of the mean rating for the respective users. The range of similarity functions is between 0 and 1, where 0 indicates no similarity, and 1 indicates a complete match. From here on out, user-based similarities are used. The equations can easily be modified to obtain item-based collaborative filtering. It is important to note that the Pearson correlation is undefined when the standard deviation is zero. With the similarity measure chosen, a rating r for a new item for user u can be predicted with the following *prediction function*:

$$r_{(u,new\_item)} = \mu_u + \frac{\sum_{v \in N_u(new\_item)} sim(u, v) * (r_{(v,new\_item)} - \mu_v)}{\sum_{v \in N_u(new\_item)} |sim(u, v)|} \tag{2.3}$$

where $N_u(new\_item)$ is the set of users that have rated the new[1] item. When no other user has rated the new item, the predicted rating reduces to the user's average rating.
Instead of using the mean μ of the respective users, one can also use *baseline estimates* [41] to account for user biases when calculating an item's rating. The baseline estimate considers whether or not an item tends to receive higher/lower ratings than average or if users give higher/lower ratings compared to others.
The formulas above (e.g., similarity) outline a significant disadvantage for the algorithm: complexity. Furthermore, the algorithm's reliance on other users' data makes it impossible to make predictions if not enough data is present. This complication is referred to as the *cold start problem*.

As expected, the accuracy of the recommendations plays a prominent role in the evaluation of the recommender system. Numerous methods and improvements to these methods have been made to increase the accuracy of said algorithms. However, a recommender system focused on accuracy alone may not be satisfactory for end-users. Users may, for example, fail to use a recommender system to its fullest potential due to a lack of trust in the system [21]. Some users may not always want to be recommended action movies similar to past recommendations. One often-used

[1] The item is in this case only new for user u.

evaluation method in the literature is ResQue, developed by Pu and Chen [66]. Pu et al. [67] outline guidelines for designing recommender systems in their work, focusing on aspects outside of algorithm accuracy.

Recommender system algorithms have also been applied for educational purposes. For example, Thai-Nge et al.'s work [68] researched using matrix factorization to predict student performance. Michlík and Bieliková [51] used a content-based approach to recommend exercises for test preparation given a limited amount of time to learn. They found that by adapting the algorithm to prioritizing covering as much material as possible instead of only a few topics in-depth, students performed better in a computer adaptive test.

2.1.2 Item Response Theory

Item Response Theory (IRT) is often used to predict the "difficulty" of an item or the "ability" of an individual. One common application of IRT is computerized adaptive testing (CAT). CATs are tests that adjust the difficulty level of the questions according to the ability of the test-taker. These tests require thorough calibration of the questions' difficulty levels using IRT to be successfully administered. CATs have been widely used in practice. Famous examples of applications of the CAT are the Graduate Record Examinations (GRE) and the Graduate Management Admission Test (GMAT) [17].
Although applying IRT may seem fit for adaptive learning, there are two main complications with this idea. First, the calibration process takes a significant amount of time and resources, making it difficult to administer in e-learning platforms where the number of problems to solve can be quite substantial. Furthermore, IRT assumes that the user's skill (ability) stays constant [11, 63], which is not the case in e-learning.

2.1.3 Elo Rating System

An often-used solution to the problem discussed above regarding IRT is the use of the Elo rating system [26], which was devised by Arpad Elo, notorious for its application in chess.
The Elo rating system in its original form is straightforward. Each player is assigned their[2] own Elo rating. The Elo rating system attempts to predict the likelihood of one player winning against another player based on their respective Elo ratings. Once the match's outcome has been determined, each player either gains or loses Elo depending on whether they have won or lost the game. The amount of Elo gained or lost depends on the initial likelihood of the particular player winning the match: the higher the likelihood of winning, the more Elo lost (fewer Elo that are gained) when the player loses (wins) the match. The Elo rating system can be applied to e-learning by viewing a student solving an exercise as a match between said student and the exercise. Each question in the problem set has an Elo rating, as each student

[2]Singular they/their is used throughout this work when applicable.

Table 2.1: The K values used by FIDE (https://ratings.fide.com/).

K values used by FIDE	
K=40	"for a player new to the rating list until they has completed events with at least 30 games"
K=20	"as long as a player's rating remains under 2400"
K=10	"once a player's published rating has reached 2400 and remains at that level subsequently, even if the rating drops below 2400"
K=40	"for all players until their 18th birthday, as long as their rating remains under 2300"

would. The original formula proposed by Elo was as follows:

$$r_n = r_0 + K(S - E(S)) \tag{2.4}$$

where r_n is the new rating, r_0 is the original rating before the match was played, K is a constant (sometimes referred to as the step size), S is the score (outcome) of the match, and $E(S)$ the expected score (outcome) of the match. Originally, $E(S)$ was calculated using the normal distribution. However, most implementations now use a logistic distribution instead due to their preferred characteristics, such as their wider tails. The expected outcome of the match for player A against player B can be calculated as follows:

$$E(S_A) = (1 + 10^{(r_B - r_A)/400})^{-1} \tag{2.5}$$

The same can be calculated for player B by switching B and A in the equation.
The Elo rating system has seen various modifications to achieve a more accurate rating system. For example, the value of K has been the topic of various pieces of literature. The larger its value, the more significant the difference between the initial and final Elo ratings will be. In chess, the value of K depends on the number of games played and the range in which the player's Elo rating falls. For example, FIDE[3] uses the values of K as seen in Table 2.1. In an e-learning context, the value of K can also be calculated using an uncertainty function, as shown in Pelánek's work [63]. Papoušek and Pelánek also propose using different K values depending on whether the question is answered correctly/incorrectly in their Performance Factor Analysis Extended / Elo system [62].

2.2 Explainable AI

2.2.1 What is XAI?

Explainable AI (XAI) (sometimes referred to as *interpretable AI*, especially within machine learning) has recently seen a resurgence in interest due to AI's increased

[3]The international chess federation (https://ratings.fide.com/)

adoption across many industries [4]. The word "black-box" is often attributed to AI, giving the impression that the reason for an output provided by an algorithm is a mystery left to the developer or end-user to decipher. For example, back in 2002, the Wall Street Journal published an article "If TiVo Thinks You Are Gay, Here's How to Set It Straight" [78] discussing the many complications users of TiVo faced due to inexplicably being recommended movies not to their taste. The importance increases even more when dealing with high-risk situations such as those that doctors face in medicine. Clinical decision support systems are tools that healthcare workers use to assist in, amongst other tasks, diagnoses of diseases. Amann et al. [6] discuss the *Clever Hans phenomenon*'s occurrence in medicine [44], where the model classified high-risk patients by which machine was used for the x-ray rather than an underlying medical cause. XAI should allow these problems to be mitigated.
However, research around XAI dates back to the 1970s in the form of *explainable expert systems* (EES). For example, medical expert systems required a form of explanation to be adopted by doctors. Using symbolic reasoning, MYCIN [72] was able to advise physicians concerning diagnoses and treatments and provide explanations and justifications for its conclusions. Although decades apart, the motivations behind using explanations for EES are widely the same for current XAI systems.

As with many popular topics, XAI does not have a single definition used by all researchers. Many authors have tackled this ambiguity in the literature. For example, Arrieta et al.'s survey [7] starts this discussion with a definition of XAI provided by D. Gunning [33]: "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners." Arrieta et al. critique this definition due to only including trust and understanding but failing to acknowledge concepts such as fairness and confidence. The authors also provide a crucial insight arising from the definition of an explanation. An explanation is *audience-dependent*: the reasons for an explanation and whether an explanation is "easy to understand" are entirely dependent on the user using it. Thus, Arrieta et al. define XAI as "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand." Mohseni et al. [54] similarly emphasize the *user groups* (audience) split into three groups:

- **AI novices:** "End-users who use AI products in daily life but have no (or very little) expertise on machine learning systems."
- **Data experts:** "Data scientists or domain experts who use machine learning for analysis, decision-making, or research."
- **AI experts:** "Machine learning scientists and engineers who design machine learning algorithms and interpretability techniques for XAI systems."

Depending on the user group, the design goals and evaluation measures vary accordingly, as visualized in Fig. 2.1. Recently, the importance of the end-user has been pushed further, for example, to distinguish explanations based on personal characteristics [10, 52].

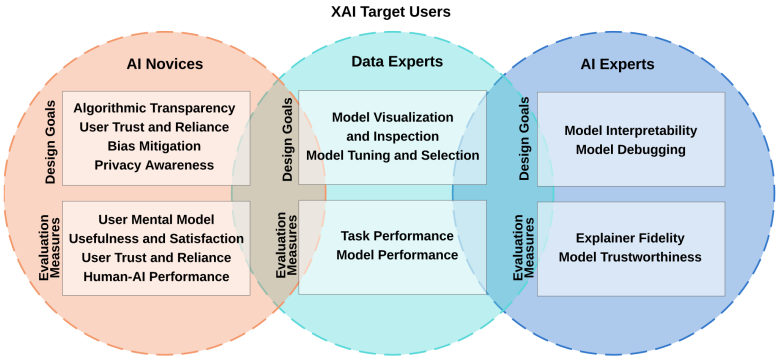


Figure 2.1: XAI target users with their respective design goals and evaluation metrics from Mohseni et al. [54].

It is important to keep in mind that the definitions of various terminology can differ depending on the work read. Some works are dedicated to addressing the various terminology and definitions [46]. Here, we outline the most important definitions needed for this thesis. Other definitions can be read in [54, 46, 6].
Miller [53] defines *interpretability* as "the degree to which an observer can understand the cause of a decision." Miller thus deduces that an explanation is one of the various mediums by which "understanding" can be achieved. Furthermore, as defined by Miller, a "*justification* explains why a decision is good but does not necessarily aim to give an explanation of the actual decision-making process." This definition implies that a justification is an explanation that does not require transparency [69].

XAI should always be accompanied by a goal that the developer wishes to achieve with the explanations. Adadi et al. [4] list the goals justification, control, improvement, and discovery. Mohseni et al.'s [54] framework define AI goals such as algorithmic transparency, user trust and reliance, bias mitigation, and privacy awareness for AI novices.

2.2.2 Explainable Recommender Systems

Interest in explanations for recommender systems began in the early 2000s when recommender systems in e-commerce started to gain traction. One of the earlier works on explanations for recommender systems is that of Herlocker et al. [35]. Various approaches to explaining recommendations were compared to investigate whether end-users had a preference. The authors also looked into whether explanations could increase acceptance as well as the accuracy of the recommendations.

Tintarev and Masthoff's work [70] set a precedent for various literature concerning explainable recommender systems. In their study, they outline four guidelines for designing explainable recommender systems. Here, we briefly discuss the first (and most often cited) guideline: designating a specific goal for using explanations. The

seven goals presented in their work can be seen in Fig. 2.2. The authors note that not

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 use or enjoyment

Figure 2.2: Seven aims of explanations with their respective definitions from [70].

all goals are simultaneously achievable: an inevitable trade-off must usually be made. For example, effectiveness may come at the cost of efficiency. Listing all possible reasons for why and why not to watch a particular movie may allow users to make a correct decision at the cost of them taking longer to process all of the information. Another trade-off can be persuasiveness for effectiveness and trust, where the former does not necessarily take into account the buyer's interests as long as the customer purchases a product.

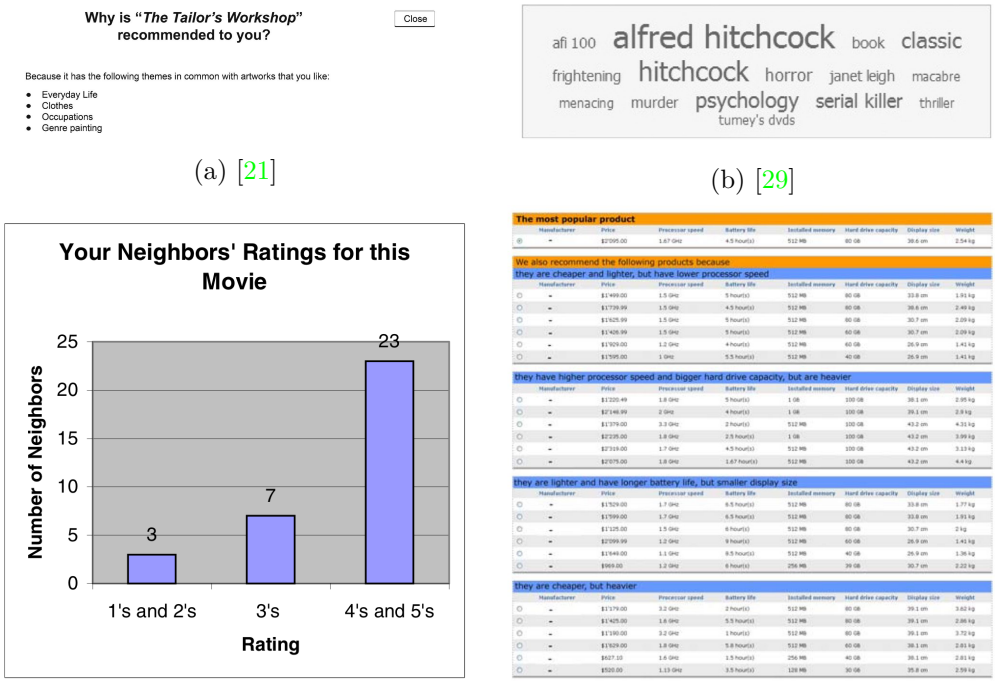
Numerous research for explainable recommender systems focuses on one or more of the goals presented above. Research can also pertain to the context in which the explainable recommender system is used or the interface used to communicate the explanations.
Explanations for recommender systems have been studied in various contexts. The most prevalent examples include e-commerce and media recommendations. Frequently, the applications in which the explainable recommendations (and XAI in general) are implemented are "mocked" instead of developing a real, fully functional application to be able to focus on the explanation aspect of the study [25].
Due to the inherent explainability of the traditional recommender system algorithms, there has also been a focus on how to present explanations in this field. The underlying recommender system algorithm itself usually constrains these explanations [70]. For example, a simple user-based collaborative filtering algorithm would usually not be capable of delivering an explanation based on the recommended item's inherent features. Although explanations come in all shapes and sizes, the domain can mostly be split into two parts: textual and visual explanations.
Textual explanations are prevalent in various commercial applications. Typical explanations include Uber Eats'[4] "People who ordered ... also enjoyed ..." or Netflix's[5] "Because you watched ...". Commercial applications also leverage their user base by explaining recommendations using reviews for the product, such as in Vivino[6]. Facebook[7] explains friend suggestions ("People You May Know") by showing the

[4]https://www.ubereats.com/
[5]https://www.netflix.com/
[6]https://www.vivino.com/
[7]https://www.facebook.com/

number of mutual friends you have ("... mutual friends"). Cramer et al. [21] explained art recommendations by listing the "themes" it had in common with other artworks the user liked (Fig. 2.3a). Tintarev and Masthoff [70] also list a myriad of examples in their work.

Visual explanations leverage some form of visualization in order to convey (a large amount of) information to the end-user understandably and efficiently. This explanation type has been the topic of various comparative studies in the literature. Herlocker et al.'s [35] paper compared various explanation interfaces, both textual and visual, to conclude that the "histogram with grouping" interface (Fig. 2.3c) performed the best in their user study.

Some explanations leverage both textual and visual information. For example, Gedikli et al. [29] used tag clouds (Fig. 2.3b), where the size of a word (tag) is proportional to its relevance. Pu and Chen [64] designed an "organization interface" (Fig. 2.3d) that categorized product recommendations according to their trade-offs with respect to the most popular product.



(a) [21]

(b) [29]

(c) [35]

(d) [64]

Figure 2.3: Various explanations used in the literature.

Explanations cannot be blindly added to an AI application, and developers of such applications must keep in mind the risks/trade-offs in mind when designing explanations/explanation interfaces. Naiseh et al. [56] outline potential risks of using explanations such as over-/under-trust and information overload, leading to incorrect usage of AI systems. Zhao et al.'s [79] research in progress similarly aims

to investigate the effects of over-transparency on user trust.

2.3 User Trust

This section aims to give insight into research concerning trust in technology and HCI. Trust is one of the seven aims Tintarev and Masthoff outline in their work [70]. It is a complex yet essential topic within the domain of HCI. Various studies, for example, show that the adoption of technology (e.g., intention to return, intention to save effort) is correlated to user trust [9, 64]. Some studies further hypothesize that enhancing user trust can increase purchasing behavior in an e-commerce context [31]. Without trust or with too much trust, users may misuse technology, leading to undesired effects [16, 50].

2.3.1 Various Definitions of Trust

Many authors state that trust is a complex construct. It has multiple definitions that depend on the field and context in which it is used. In the early 2000s, researchers questioned whether trust between two people (interpersonal trust) is the same as or similar to trust between a person and a computer (human-computer trust). However, early researchers give evidence [9, 20] as to why this is justified. In more recent literature, Holliday et al.'s study [37] show that similar to interpersonal trust, the rate at which trust is gained is significantly lower than the rate at which it is lost. Recommender systems are often viewed as virtual assistants, and many commercial applications include conversational elements, which personifies technology even more. Therefore, the assumption that human-computer trust is similar to interpersonal trust is usually taken to be true.

A plethora of definitions of trust exists in the literature. This section does not aim to discuss all of these definitions. Instead, a few relevant definitions are presented to observe some common themes. Some definitions of trust in the literature can be seen in Table 2.2.
Competence is a common theme (construct) in many of the definitions, which aligns with Palmer et al.'s [61] statement, "... when a person trusts a system, it is not that a trust threshold has been met but because the person has determined that the system can adequately perform a specific purpose." In fact, perceived competence is sometimes used as a substitute for trust [64].

2.3.2 Time-Dependency of Trust

Aside from the complexities arising from the various definitions of trust, there is also a consensus that trust is not static but rather something that evolves depending on various factors. However, many authors in the literature measure user trust with a single measurement, usually defining it as *initial user trust*. As the name suggests, initial user trust is used when the trustor has no previous knowledge of the trustee. This trust is developed when a user, for example, visits a website or uses

Table 2.2: Some examples of definitions of trust used in the literature, as well as the themes present.

Definitions of Trust		
Grandison and Sloman [32]	"The firm belief in the competence of an entity to act dependably, securely, and reliably within a specified context."	Competence Reliance Security
Chopra and Wallace [20]	"The willingness to rely on a specific other, based on confidence that one's trust will lead to positive outcomes."	Reliance Benevolence
Muir [55]	"Persistence of the natural and moral social orders, technical competence, and carrying out their fiduciary obligations and responsibilities."	Integrity Competence Responsibility
Lee and See [45]	"The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability."	Benevolence Uncertainty Vulnerability
Cramer et al. [21]	"The user's willingness to depend on a system and its recommendations in the specific context of the user and his or her task(s), even though the system might make mistakes."	Reliance Uncertainty
Wang and Benbasat [9]	"An individual's belief in the in an agent's competence, benevolence, and integrity."	Competence Benevolence Integrity

an application for the first time. Authors such as Wang and Benbasat [9] argue for utilizing initial user trust in the context of technology adoption as the barriers built by the uncertainty of technology and intentions are the largest that must be overcome.

Some works have measured user trust over time. Holliday et al. [37], for example, showed that user trust is a "dynamic" attitude that evolves. Here, the "journey" of user trust over time was compared for a group without explanations and a group with explanations. Although initial user trust levels were comparable, user trust increased by using explanations, whereas for the group without explanations, the user trust stayed the same or decreased. User trust, according to Holliday et al.,

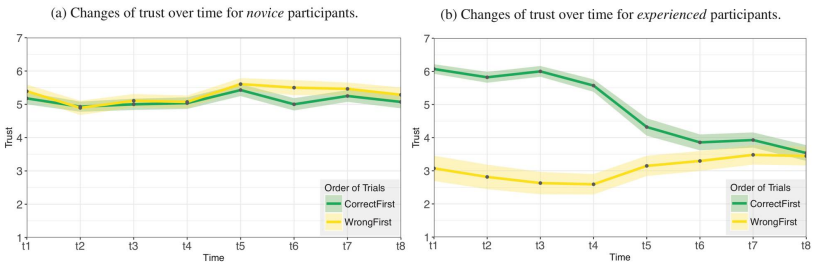


Figure 2.4: Value of trust at various points in time in Nourani et al.'s experiment [59]. The change in trust is apparent for the users with domain knowledge.

should thus be measured as a "journey" over time instead of one static measurement. Nourani et al. [59] also measured user trust at various points in time for their research. They found that user trust in the system and its evolution were influenced by the user's domain knowledge and the first impression of its competence. A graph of their results can be seen in Fig. 2.4.

2.3.3 Measuring Trust

The brief discussion above should convey to the reader the complexity of the topic of "trust". If trust is not defined officially, how is one expected to measure it? As the reader may expect, there is, unfortunately, no standard method of measuring user trust. Thus, readers may find various ways of measuring user trust in the literature due to the lack of a universal standard. It is also important to note that no "absolute" measure of trust exists [61]. This implies that trust must be measured relatively. As mentioned above, trust can, for example, be measured at various points in time. Trust can also be measured through A/B testing (between-subjects): examining whether doing x increases/decreases trust compared to a certain baseline.
In HCI, there are two predominant measurement types: *explicit measurements* and *implicit measurements*. Explicit measurements are usually done through questionnaires and interviews. On the other hand, implicit measurements use techniques such as logging to obtain values. We first discuss explicit measurements, as this is still the dominant way of measuring in this field. Afterward, we outline some implicit measuring techniques.

Explicit Measurements

The literature shows two primary types of user trust, which we refer to from here on out as *one-dimensional* and *multi-dimensional* trust. One-dimensional trust is measured by explicitly asking the end-user whether they trust the system employing, for example, a Likert scale [59, 37, 52]. The literature that uses this type of trust consists primarily of research where trust is not the only concern (and is only one aspect of the full questionnaire). The benefit of this method is, obviously, its ease: only one question must be asked to the participant to obtain a value. However, as can

be expected, this method is far from optimal. Trust is "complex and multidimensional" [20]. Therefore, one measure may not be able to capture the complexities of trust to a satisfactory degree. Furthermore, participants of user studies may interpret trust differently, reflecting the multiple definitions in the literature.
Multi-dimensional user trust attempts to avoid the problem above by measuring the various constructs of trust in some manner. It can, for example, be calculated by splitting "trust" into its multiple *constructs* and evaluating them using a questionnaire. The values obtained (e.g., on a Likert scale) can then be summed to derive a value. Note that in the literature, one may also read about *dimensions* of trust. In this work, constructs and dimensions are considered as synonyms.
McKnight et al. [48] introduced *trusting beliefs* as "the perception that the trustee ... has attributes that are beneficial to the truster." in their influential work. The three constructs are defined as follows:

- **Perceived Competence.** "The ability of the trustee to do what the truster needs."
- **Perceived Benevolence.** "Trustee caring and motivation to act in the truster's interests."
- **Perceived Integrity.** "Trustee honesty and promise keeping."

These constructs frequently return in the literature. Vidotto et al. [73] explain trusting beliefs as "a relevant factor in causing an individual to consider another individual to be trustworthy." More recent works, such as Berkovksy et al. [10], add *perceived transparency* for the end-user to these three constructs to measure user trust. Berkovsky et al. further clarify integrity also to contain unbiasedness and non-discrimination. *Intention to reuse/return* has also often been used as a construct of user trust [65, 10]. Chen [18] developed a trust model for a product recommender system. In this model, competence is further split into "perceived ease of use, perceived usefulness, decision confidence, and enjoyment." Furthermore, reputation was added as a construct to trusting beliefs.
Splitting trust into its various constructs may lead to more reliable and interpretable results than its explicit counterpart due to its granularity. On the other hand, the measurements may require (long) questionnaires, which often may not be optimal. Furthermore, questionnaires are susceptible to many types of bias [19], which may influence the final result. Users may also behave differently in an environment where they know they are being tested [43].

Implicit Measurements

There have been efforts made to measure implicitly to avoid problems such as bias and other complications that accompany explicit measurements. Ghergulescu et al. [30] outlined three implicit metrics (and one explicit metric) to measure motivation in the context of game-based e-learning. These metrics are based on time spent on a particular action, how often actions are performed/repeated, how often assistance is requested, and the user's confidence. Implicit measurements also play an essential

role in the research domain of search engines [28, 24]. For instance, Fox et al. [28] investigated implicit measurements such as time spent on a page, click-through, exit manner (how the user exits a page) for search engine performance and satisfaction. Trust (in general) has been measured implicitly by various means. Ermish et al. [27] conducted an experiment where a participant had a choice to entrust money to a stranger to measure trust between two individuals. Perhaps more interesting, Burns et al. [15] used the *Bona Fide Pipeline* to measure trust towards co-workers implicitly. Palmer et al. [61] outlined various "measure of effectiveness" for attributes (constructs) of trust in the context of autonomous systems. For example, perceived competence could be measured by "percentage of time operator chooses to not override system."
Unfortunately, implicit measurements for trust in the context of explainable recommender systems have not yet been widely applied in the literature. Tintarev and Masthoff [70] suggest the use of *loyalty* to measure trust implicitly. McNee et al. [49] measured loyalty using the number of logins post signup. This measure's success aligns with the consensus that trust must be measured over time. Furthermore, loyalty can be monitored without the users' knowledge, effectively reducing the bias that may stem from a controlled experiment. However, especially in short-term user studies, obtaining such measurements can be challenging.

Placebo Explanations

As mentioned earlier, trust is a relative measure, meaning it must be compared to some baseline value. Recently, questions have been asked on the validity of the baseline that is typically used [54]. It has become the norm in the literature to measure the effects of explanations on, amongst other things, trust by comparing the obtained value to a baseline where no explanations are present. However, there may be evidence to believe that just the presence of words may influence a user's trust, disregarding the explanation's actual content [25]. Langer et al. [43] showed that requests using placebic information were more likely to be accepted as opposed to no explanation at all.
Somewhat related is Nourani et al.'s work [58], where they investigated the effect of meaningless explanations in XAI. Their results showed that the meaningfulness of the explanations significantly affected the participants' perceived accuracy. The authors further state that meaningless or non-understandable explanations may reduce a user's trust in the system.

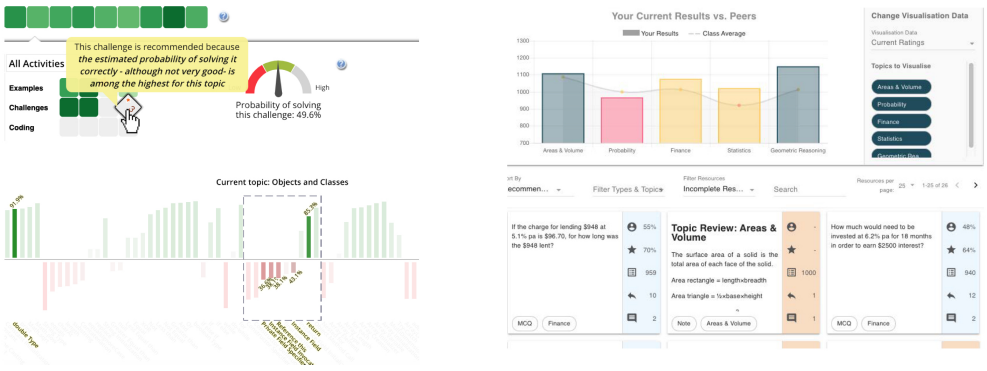
2.4 Related Work and Gaps in Literature

Despite high interest in explainable recommender systems and XAI in general, there seems to be little literature concerning its application in e-learning systems, a view shared by Barria-Pineda [8]. However, some similar approaches, such as Open Learner Models (described later in this section), have been studied.
There also seems to be a lack of XAI/explainable recommender system research and

trust research for the lower end of the age spectrum (individuals under the age of 18). Indeed, the literature primarily focuses on university students and young adults [29, 25, 18, 21, 9, 66, 71, 59, 37, 10, 31, 58]. In Section 2.3.3, we discussed some ways trust has been measured implicitly in the literature. However, many of these approaches are not applicable when measuring trust through users' actions on a website. We have seen that long-term trust in the form of loyalty could be measured implicitly [49]. This approach is not feasible for short-term studies. In general, there seems to be a lack of literature concerning measuring trust implicitly in an online context for explainable recommender systems.
In this section, we scope in on the results of a few pieces of literature that tried to find the effect of using explanations on user trust. The section also outlines this thesis' contributions to address the gaps in the literature mentioned above.

2.4.1 Effect of Explanations on User Trust

Pu and Chen [64] developed the organization interface as an alternative to traditional presentation interfaces such as the Top-N items to address the diversity of recommendations and efficiency. The result was an interface that visually organized a diverse array of products according to their trade-offs compared to the top recommendation. They found that their organization-based interface was better for user trust (measured through perceived competence, intention to return, and intention to save effort). Cramer et al. [21] studied the effect of transparency on user trust in the context of an art recommender system. The results showed that transparency did not lead to higher user trust nor perceived competence in the system. However, acceptance of the recommendations was higher with transparency.
The effect of various recommendation interfaces, including one using explanations, on user trust was studied by Berkovsky et al. [10]. Their results showed that different types of explanations (persuasive, personalized, or IMDb) had varying effects on each construct of trust in the context of a movie recommender system. Trust was measured through trusting beliefs (competence, integrity, benevolence), intention to reuse, and transparency.
Nourani et al. [58] researched how the level of meaningfulness of an explanation affected the perceived accuracy of the system. The study showed that non-meaningful explanations significantly lowered the perceived accuracy of the system.
Inspired by Langer et al. [43], Eiband et al. [25] recently conducted an experiment to investigate the influence of placebic explanations in the context of a nutrition recommender system. Despite the low number of participants, comparing the medians of the responses showed that placebo explanations had higher trust levels than no explanations and similar levels of trust to the real explanation.
Barria-Pineda's [8] work in progress plans to investigate the effects of transparency in educational recommender systems. Amongst others, their work aims to answer "What type of explanatory interfaces would be most appealing and meaningful to students for obtaining information about why specific learning materials were recommended to them?" The target audience in this research is college students. An



(a) [8]

(b) [3]

Figure 2.5: The left image shows the explanation interface used in Barria-Pineda's work in progress [8]. This interface uses both a visual OLM and a textual explanation. The image on the right shows the augmented OLM used in Abdi et al. [3].

initial experiment has been conducted using a rule-based recommender system in an introductory Java course. The explanation interface combined both a visual OLM and a textual description, as can be seen in Fig. 2.5a.

2.4.2 E-Learning Platforms

In this thesis, we build upon an existing e-learning platform, Wiski [60], which was developed for Ooge's thesis. The platform was a proof of concept to study the effects of gamification on Belgian high school students' motivation. However, various other e-learning platforms have been developed for research as well. For the remainder of this section, we briefly list and discuss various other e-learning platforms related to our work.

Math Garden Math Garden [40] is an example of an e-learning platform that successfully implemented the Elo rating system for adaptive learning. Their algorithm is a variant of the Elo rating system that takes into account how quickly the user can answer the question. The Dutch platform was used by over 5000 primary schools in the Netherlands to practice basic arithmetic [12].

Matistikk Matistikk [34, 23] is a simple e-learning platform where students can take tests created and uploaded by a teacher. As an extension to this platform, Dahl and Fykse [23] introduced the idea of combining the Elo rating system with the collaborative filtering algorithm to determine the following question to show to the end-user. This solution can be seen as a two-step algorithm: first, all unsolved questions in a predetermined Elo range are retrieved from the database. Then, the estimated number of tries is calculated for each question using collaborative filtering, and the list of potential questions is sorted according to these values (in ascending

order).
Although the differences in Elo rating between the two groups were not statistically significant, it is essential to note that the number of students was limited (sample size equal to 48). More importantly, first-year university students were used as subjects while the math problems were for 8th graders.

RiPPLE RiPPLE [2] is a relatively novel learning platform that, amongst other algorithms, uses a multivariate variant of the Elo rating system. Such a multivariate system can keep track of a user's knowledge level for multiple, independent courses. Recently, Abdi et al. [3] investigated the effects of augmenting an educational recommender system with an *Open Learner Model* (OLM). The interface developed can be seen in Fig. 2.5b. Intelligent Tutoring Systems contain Learner Models of each user. Such models can be seen as the internal representation/profile the system creates. Traditionally, Learner Models are hidden from the user. However, much research has been done concerning opening the Learner Model to the user (OLM) to see added benefits such as learner reflection and increased trust [14]. Abdi et al. conducted a randomized controlled experiment to test whether adding an OLM to the recommendation interface impacted, amongst others, understanding and trust in the recommendations. Their studies showed that adding the OLM was a statistically significant benefit to the system. The OLM in this context can be seen as a type of explanation interface.

METAL Project The METAL project [13, 22] aims to develop a fully integrated e-learning platform for high school students in France. Apart from recommending (with explanations) exercises to the end-user, it also recommends resources such as books or lectures. Students also have access to visualizations of their level and academic performance. The METAL project also provides functionality for the teachers, such as dashboards to see the users' activities.

2.4.3 Research Goals

At the beginning of this section, we addressed three gaps in the literature. This thesis aims to contribute to the literature by tackling these gaps. An e-learning platform, Wiski, is modified such that it attempts to recommend exercises to the end-user based on difficulty level. Next, an explanation interface for the recommended exercises is developed, and its effect on Belgian high school students' initial user trust in the platform is evaluated. The effect of placebo explanations is studied by using it as a second form of control. The lack of studies surrounding placebo explanations, especially with the research's unique target audience, may offer interesting insights. Finally, we investigate whether there is a correlation between the acceptance of a recommendation (click-through rate) and the initial user trust of the user, which may open up the possibility of using it as an implicit measurement.

Chapter 3

Methodology

The previous chapter outlined the main goals of this research. Based on those goals, we define the following research questions for the thesis.

Research Question 1: Do explanations lead to increased *initial* user trust for Belgian high school students in the context of an e-learning platform?

Research Question 2: What influence do placebo explanations have on *initial* user trust?

Research Question 3: Is there a correlation between the *initial* user trust and the acceptance of the recommendations (measured through click-through rate)?

High school in this thesis is used to refer to grades 9 to 12, whereas middle school refers to grades 7 and 8.
This chapter delineates the appropriate setup needed and the reasoning followed (when applicable) to address the outlined research questions. Although the thesis is in English, Dutch is occasionally used as this is the language used in Wiski and the questionnaires. Translations are provided where they are necessary. The research conducted has been approved by the *Sociaal-Maatschappelijke Ethische Commissie* (SMEC) (File: G-2021-3233-R2(MAR)).

3.1 Wiski

A modified version of Wiski, a platform developed by Ooge in 2019 [60], is used as the medium to conduct the research. Due to the modified version of Wiski and the original version of Wiski sharing many ties, we must often refer back to the original platform developed by Ooge. To distinguish between the two versions, the Wiski developed for this thesis is simply be referred to as Wiski, whereas the Wiski

developed by Ooge in 2019 is called Ooge's Wiski.
Wiski is a free e-learning platform that provides more than one thousand multiple-choice math exercises for students in Belgian secondary education (grades 7 to 12). These math exercises are courtesy of die Keure[1], a renowned publisher in Belgium that offers textbooks for mathematics (van Basis Tot Limiet) widely used by Belgian secondary education. The exercises that can be found on Wiski overlap with those offered on die Keure's e-learning platform, Polpo. However, there are no binding terms attached to the use of these exercises and thus, die Keure are in no shape or form further involved in this work. Some screens from Ooge's Wiski can be seen in Fig. 3.1. Upon registration, users can freely use the website according to their

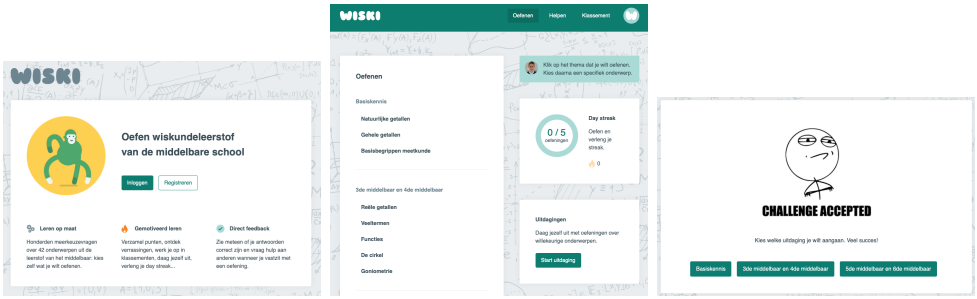


Figure 3.1: Screens from Ooge's Wiski. The website can be visited at www.wiski.be.

needs. Wiski offers exercises from 14 different subjects, further split into 41 different sections. For clarification purposes, a *subject* is, for example, derivatives. A *section* for the derivatives subject can be, for example, chain rule or derivatives of logarithmic functions. Exercises are only offered at the level of a section. Users can choose from which section they wish to practice, and all content is immediately available. All exercises on Wiski are multiple-choice questions, and users must repeatedly make attempts for the exercise until they answer it correctly to move on.

The most important screens of Wiski can be seen in Figs. 3.2 to 3.6. The platform was accessible at sho.wiski.be during the research. Afterwards, the platform was brought offline.

3.2 Participants

As the target population of this research was high school students, the sample needed to reflect this as closely as possible. In this section, we discuss how the participants were recruited as well as their demographics.

3.2.1 Recruitment Process

High schools in Flanders were contacted to inquire whether they would be interested in using Wiski during one of their math classes or assigning its use as homework.

[1] https://www.diekeure.be

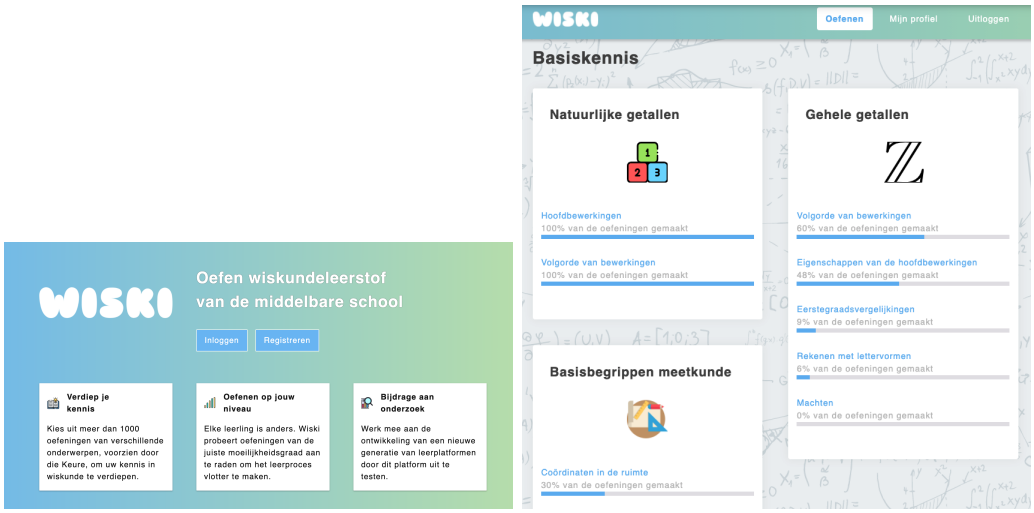


Figure 3.2: The left image shows the landing page of Wiski when sho.wiski.be is visited. The image on the right shows the subject and section selection page. Users can see the percentage of exercises solved for a particular section.

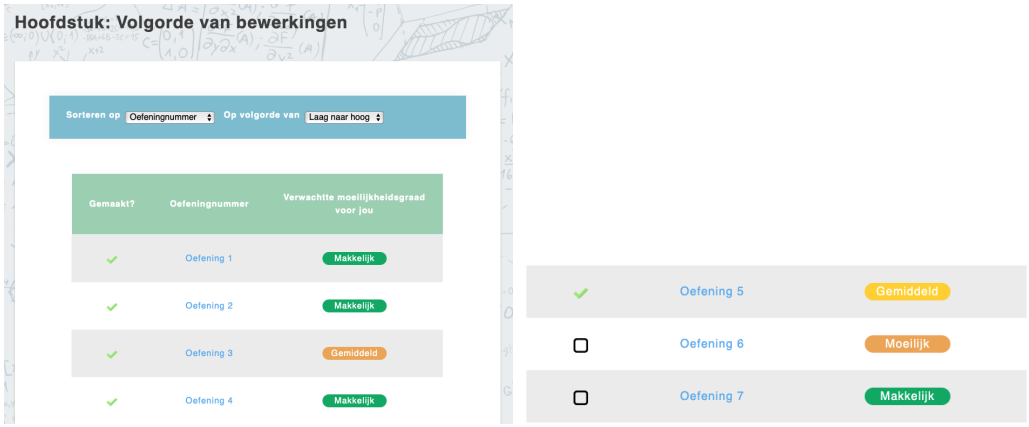


Figure 3.3: The exercise selection page (left) displays whether exercises have been solved or not, the link to the exercise, and the expected difficulty level for the user (based on the difference between the user's Elo rating and that of the exercise). The three difficulty levels are "makkelijk" (easy), "gemiddeld" (medium), and "moeilijk" (hard). The color coding of the difficulty levels can be seen on the right.

Teachers interested in participating then received an information brochure outlining the goals of the research, the process of the research, and extra information relating to how the data is used and stored. Furthermore, it was made clear that it was each individual student's choice to decide whether they wished to participate. If a student did not wish to participate, they had access to an offline alternative similar to the content provided on Wiski. A similar information brochure for parents was also sent to the teachers to forward to the participating students' parents. Students

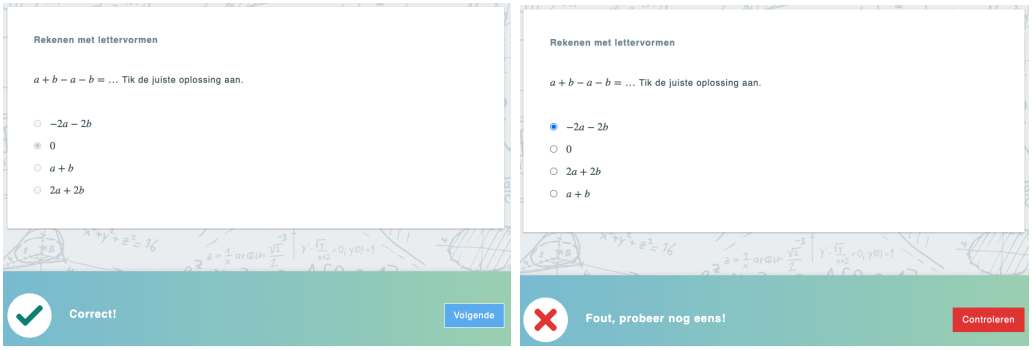
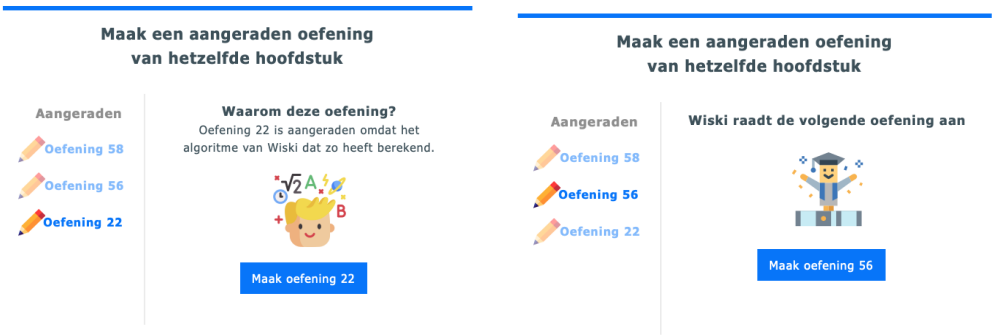


Figure 3.4: Screen the users saw when solving an exercise correctly (left) or incorrectly (right). The "Volgende" (next) button leads to the explanation interface.

(a) IRE

(b) Not Enough Data

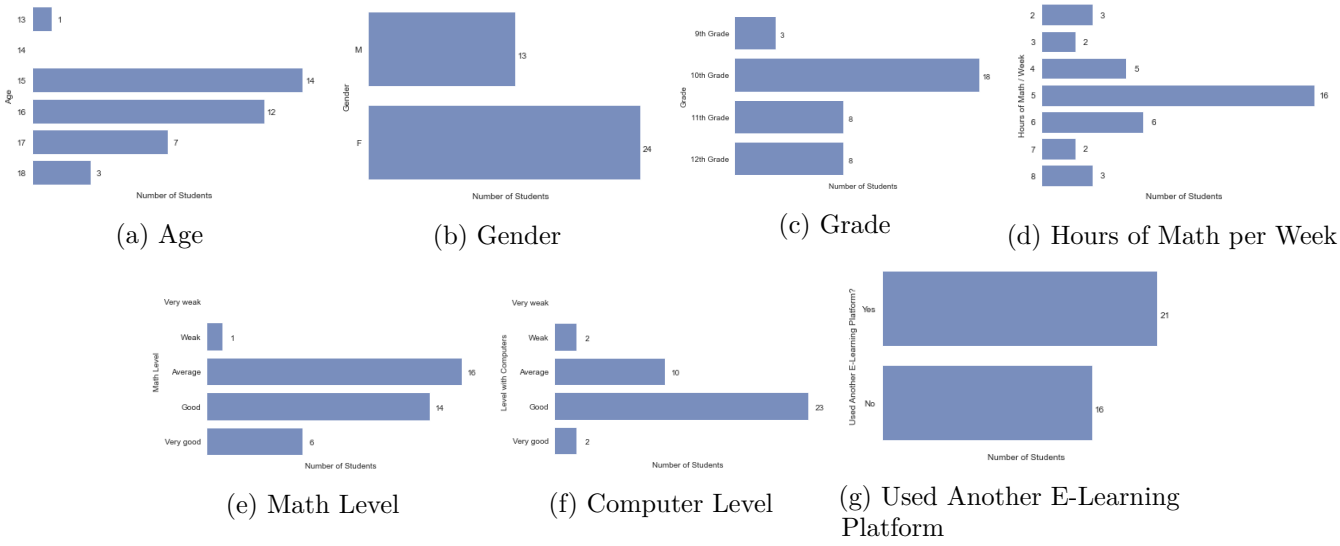
Figure 3.5: The IRE when there is enough data (left) and when there is not enough data (right).



(a) IPE

(b) INE

Figure 3.6: The IPE (left) and the INE (right).



(a) Age (b) Gender (c) Grade (d) Hours of Math per Week

(e) Math Level (f) Computer Level (g) Used Another E-Learning Platform

Figure 3.7: The demographics of the participants of the research. Participants not in high school have been pruned from the data. An extra option "Other" was provided for gender but was never selected.

(and parents for students aged 15 and under) were required to sign an informed consent form to participate in the study. The documents concerning the recruitment process (information brochure for teachers, parents, and informed consent form) can be consulted in Appendix A.1.

3.2.2 Final Participants

37 high school students were recruited from Flanders to participate in the study. Fig. 3.7 shows the demographics of the participants of the final user study. We can observe that the majority of participants are in 10th grade.
The total number of participants is much less than what was expected. Verbal agreements were made with various teachers to use the platform, but unfortunate circumstances led to most teachers not having enough time to participate in the study. Attempts were made to recruit further teachers but to no avail.

3.3 Randomized Controlled Experiment

Three research groups are required to address the research questions: a group that receives a real explanation, a group that receives a placebo explanation, and a group that receives no explanation at all. The evaluation of the explanation interface is conducted using a *randomized controlled experiment*. Randomized controlled experiments (also known as randomized controlled trials) are commonly used to measure the effect of a treatment (in this case, an interface) by defining two or more groups, of which one is used as control. Here, the no-explanation group acts as the control

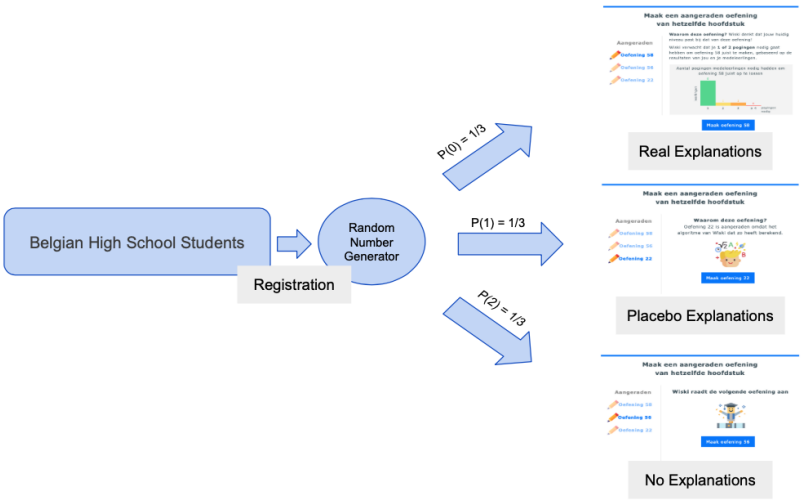


Figure 3.8: Visualization of the randomized controlled experiment. Participants were assigned a group randomly upon registration. The group determined which explanation interface the user saw.

group. Users are assigned to one of the three groups mentioned above randomly upon registering for Wiski. The website's interface and functionality are the same for the three groups, except the explanation interface (which is discussed more in detail in Section 3.6). This way, we can most likely conclude that any observed differences in the results between the three groups stem from the explanation interface itself. A visualization of the randomized controlled experiment can be seen in Fig. 3.8.

One trade-off was made when designing the randomized controlled experiment: how users were assigned to the groups. As the experiments were conducted in real schools, there was a possibility that students could come to the realization that other students received different explanation interfaces. This realization could lead to biases in the results, especially if a user with no explanations saw a student's interface with explanations. The alternative was to assign a random group to each participating school. This way, all students within one school received the same explanation interface, and the likelihood of the above problem would significantly decrease. However, the selection bias introduced using this method would most likely outweigh the one introduced in the former method, as the student's group becomes dependent on where the user lives or the courses offered at the school. Furthermore, due to the Covid-19 pandemic, most schools still operated remotely further decreasing the likelihood of the problem mentioned above. We, therefore, made the conscious decision not to use this option.

3.4 Flow of Study

The final user study consisted of the participants registering for Wiski, using it for a short period, and then filling in a post-study questionnaire asking about the participants' trust in the system. In order to obtain reliable results, the flow that the participants experienced must be uniform. Participants could only solve math exercises on Wiski, limiting the possible flows they experienced on the platform. Furthermore, all participants saw and filled in the questionnaires at the same points in the study. Participants were redirected to the pre-study questionnaire immediately after registration. Upon starting the sixth exercise (and thus having already solved five exercises), participants were redirected to the post-study questionnaire.
The users experienced a flow similar to that visualized in Fig. 3.9. After filling in the final questionnaire, the participants were free to use the platform as they wished. The decision to show the post-study questionnaire after solving five exercises and starting the sixth exercise resulted from the following two factors.

- By showing the post-study questionnaire after a set number of questions rather than a set amount of time, users had the opportunity of seeing the explanation interface the same amount of times before filling in the questionnaire.
- As the use of the platform was courtesy of the participating teachers' time, participants must be capable of completing it in under an hour. Analysis from Wiski's past data provided by Ooge [60] showed that participants solved an average of three quizzes (each with five questions) per hour. However, users did not need to fill in a post-study questionnaire. Thus, six questions seemed a safe number while still allowing the users to see the explanation interface enough times, further accounting for time to set up, clean up, and other deviations.

3.5 Algorithm

As mentioned in the introduction, the main goal of this thesis is not to find the most accurate algorithm for adaptive learning, nor is it to study novel approaches for XAI. A best effort was, nonetheless, made to use an appropriate algorithm. However, complex algorithms such as neural networks were not considered due to lack of training data and interpretability difficulty.

The recommendation algorithm used by Wiski is based on the one presented in Dahl and Fykse [23]. The Elo rating system and the collaborative filtering algorithm are capable of complementing each other well. The Elo rating system can somewhat alleviate collaborative filtering's cold start problem. Even when little to no data is present, exercises can still be recommended by comparing the Elo ratings between the user and the exercise. The Elo rating system also offers the opportunity to track the users' levels and those of the exercises naturally.
Theoretically, the collaborative filtering algorithm may pick up on information unavailable to the Elo rating system, as it can group users that answer specific

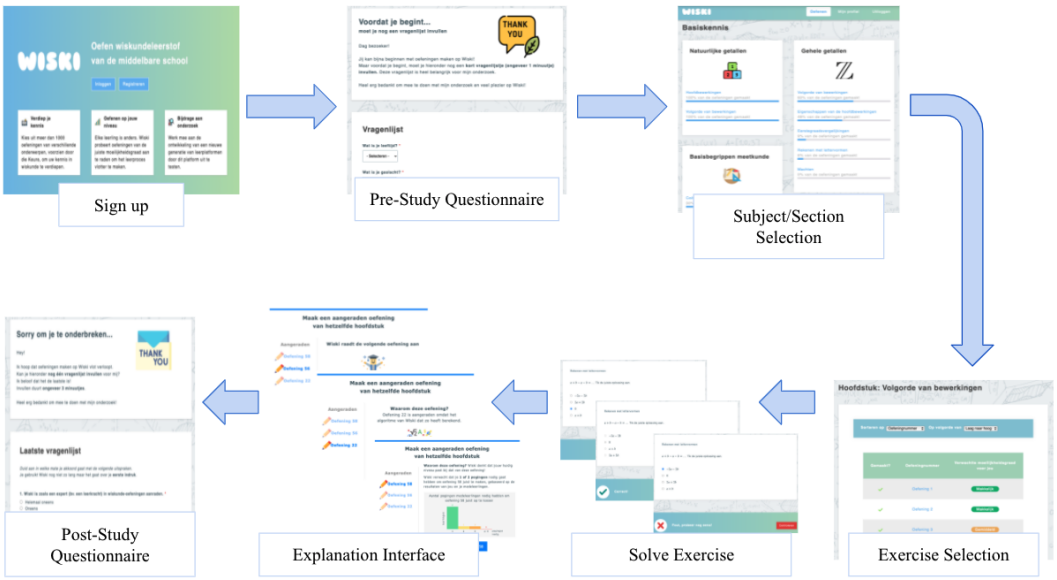


Figure 3.9: Possible flow participant could follow when using Wiski. The post-study questionnaire is shown upon starting the sixth question.

exercises (in)correctly. Furthermore, collaborative filtering can be a significant asset to the platform when users move on to a subsequent section or subject. In this situation, comparing the users' Elo ratings to that of the exercises may not lead to accurate results. However, collaborative filtering is not dependent on the section or subject in which the users find themselves.

3.5.1 Elo Rating System

To use the Elo rating system, we must set the following: the initial Elo ratings for the users and the exercises, the value of K (the step size), and how to calculate the expected score of a "match".

Initial Elo Rating The Elo ratings can be seen as relative measures. Thus, the initial values for the users and exercises can be set to an arbitrary baseline. We set the default value as 1000. However, if we have extra information regarding the difficulty level of the exercises, we can obtain a somewhat better starting value (compared to the baseline Elo rating). Data concerning the number of attempts each user required for a particular exercise was made available from Ooge's original research with Wiski [60]. Therefore, we can leverage this data to provide initial difficulty levels for particular exercises that have already been solved at least once. Wauters et al. [76, 63] state that the proportion of users that answered an exercise correctly ($correct\_attempts/total\_attempts$) forms a reasonable estimation of an exercise's difficulty level in their work. Building upon this statement, we can calculate

an exercise's initial Elo rating with the following reasoning.
The number of attempts needed to solve an exercise can be converted to the "proportion correct" (PC) by only counting answering an exercise correctly on the first attempt as correct (and all other attempts as incorrect). By interpreting the proportion correct as the probability of solving the exercise correctly and setting the initial Elo rating for a user at an arbitrary constant (in this case 1000), we can take the logistic function (Eq. (2.5)) used to calculate the expected score of a match and solve the equation for the exercise's initial Elo rating.

$$Elo_{\text{exercise,initial}} = Elo_{\text{user,initial}} + 400 * \log_{10}(\frac{1}{PC} - 1) \tag{3.1}$$

K value We can calculate the value of K similarly to the approach used by FIDE (see Table 2.1). The calculation of the value of K for exercises and users differs slightly. The users' K values can either be 40, 20, or 10, depending on the same conditions as FIDE's version. The exercises' K values are either only 10 or 20. Using uncertainty functions as in [63] require extensive parameter tuning, which is impossible with the scarce data currently available and the short duration of this research. The simplicity of FIDE's case statements and its wide-scale adoption in chess make it a reasonable choice to implement in Wiski, especially when the algorithm's accuracy is not the main focus of the research.

Expected Score Calculation Similar reasoning as above follows for the calculation of the expected score. The response-time-dependent Elo rating system used in Math Garden [40] can indeed be beneficial to Wiski, especially as analysis of data obtained from the prior use of Wiski shows that most users answer the exercises correctly on their first attempt. However, no information regarding the average time to solve an exercise is present in the data, let alone any data for the majority of the exercises. We thus opt to use the original expected score function, as seen in Eq. (2.5).

We further make two significant modifications to Dahl and Fykse's algorithm to fit our needs better.
Modification 1. Possibly the most significant change to Dahl and Fykse's algorithm is that exercises are recommended until all exercises from the current section are solved. The original algorithm only looked for exercises in a certain Elo range and stopped recommending exercises once no exercises fit this criterion. Wiski suffers from the cold start problem. Therefore, the Elo ratings of most exercises are far from their accurate, converged value. There is thus a risk that users prematurely stop receiving recommendations. As Wiski is a practicing platform, it made sense to recommend all exercises, prioritizing interesting exercises first.
Modification 2. As a consequence of the above modification, the n exercises with Elo ratings closest to $Elo_{user} + 50$ are retrieved. The constant 50 is added to the user's Elo so that there is a preference for harder exercises as opposed to easier ones. For

simplicity, we refer to this value ($Elo_{user} + 50$) as Elo_U.

3.5.2 Collaborative Filtering

Section 2.1.1 outlined the basics of collaborative filtering. We opt to use the k-nearest neighbors (KNN) algorithm with baseline for Wiski. The baseline algorithm should theoretically offer more accurate results than the basic KNN algorithm used by Dahl and Fykse. The added computational cost further does not pose a problem for the platform. The similarity measure used is the Pearson correlation. The *number of attempts needed* for an exercise are used as the "rating" the user gives for the exercise.

3.6 Explanation Interface

The explanation interface explains to the end-user why a specific exercise has been recommended and extra information that allows them to make an informed selection. The user is shown this interface upon solving an exercise correctly. The explanation interface is central to the research questions and has as its primary purpose in this thesis to increase the initial user trust.
There are three different explanation interfaces in Wiski: the interface for real explanations (Fig. 3.5), the interface for placebo explanations (Fig. 3.6a) and the interface with no explanations (Fig. 3.6b). The research group assigned after registration decides which interface the participant experiences.

3.6.1 Interface for Real Explanation

The explanation in the interface for real explanations (IRE) (Fig. 3.5a) can be seen as a combination of three parts: the why-explanation, the justification-explanation, and the histogram explanation. The translations of the text used in the explanations can be seen in Section 4.5.

Why-explanation. Upon inspecting the explanation interface, it is quite clear that the why-explanation is vague and generic: the explanation only communicates to the user that the recommendations are made based on the user's level and the difficulty level of the exercise. However, this is a decision that was made consciously. As explained earlier, the algorithm uses a combination of collaborative filtering and the Elo rating system to recommend exercises to the end-user. Exposing the Elo rating system to the end-users can have unintended effects on the study. Doing so can be seen as a (very) simple Open Learner Model, which may lead to added outside influences [3]. Furthermore, users may be confused by the mismatch between the difference in Elo ratings and the estimated number of tries the user needs to solve the recommended exercise. Recall that all exercises in a particular section are recommended, and thus the final questions in a chapter may be too easy / too hard. There is a mismatch in Elo ratings between the user and the recommended exercise

in this situation.

Justification-explanation. The justification-explanation provides users information that allows them to make informed decisions when selecting a recommended exercise. The justification displays the estimated number of tries the algorithm expects the user to need for the recommended exercise. It also further communicates that the estimated number of tries is calculated using the user's data and the data of their fellow students.

Histogram explanation. The histogram is based on the visual explanation introduced in Herlocker et al. [35]. Here, the number of neighboring students is plotted against the number of attempts needed to solve the exercise.

When the collaborative filtering algorithm does not have enough information to give an informed recommendation (either when the user solves an exercise on Wiski for their first time, or when no users have solved the exercise yet), the explanation is transparent to the end-user by showing the screen in Fig. 3.5b. The words on the interface translate to "Wiski does not yet have enough information to support this recommendation. Do you want to solve this exercise such that Wiski can collect more information?". The conducted think-aloud studies showed that using transparency in this form may positively influence user trust in the system. A few participants of the final think-aloud study remarked that it adds confidence in not believing that the recommendations are random. These observations align with Tintarev and Masthoff's thoughts where they state "users may also appreciate when a system is "frank" and admits that it is not confident about a particular recommendation." [70]

It is essential to try and capture a balance when giving explanations to the end-user. For example, Zhao et al. [79] hypothesize that it is possible to give too much transparency to the end-user. As discussed in Section 2.2, it is critical to keep the audience in mind when providing explanations. The explanation interface, in this case, has also been interactively developed to suit the target audience's needs as much as possible. This is discussed in more detail in Section 4.1.

3.6.2 Interfaces for Placebo and No Explanation

Placebo explanations [25, 54] can be used to indicate the effectiveness of the IRE. The user trust obtained through the interface for placebo explanation (IPE) (Fig. 3.6a) can be compared to that of the IRE to investigate if there is a difference. The IPE can thus be seen as a second control group for the randomized controlled experiment. We use a placebo explanation similar to the one used by Eiband et al. [25], adjusted to the current context. The explanation seen in Fig. 3.6a can be translated to "Exercise 22 is recommended because this is what Wiski's algorithm calculated." Apart from the placebo explanation, we also add an image to the interface to fill space, so users are not weary of a relatively empty interface. The interface with no explanation

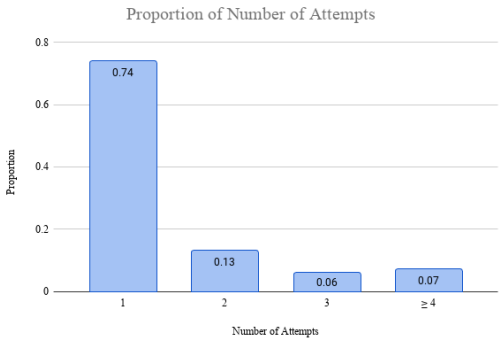


Figure 3.10: The proportion of the number of attempts needed retrieved from Ooge's Wiski's data.

(INE) (Fig. 3.6b) does not contain an explanation. An extra title and an image are added for the same reasons as mentioned above.

3.6.3 Cold Start Problem

As do all recommender systems, Wiski also suffers from the cold start problem. Although exercises can always be recommended to the end-user thanks to the Elo rating system (all exercises and users are assigned an initial Elo rating), the explanation interface still suffers as the histogram explanation and the justification-explanation are dependent on the collaborative filtering algorithm. This is problematic for the research as some participants may only see the interface indicating that not enough information is present for the recommender system to offer an informed recommendation (Fig. 3.5b). Thus, certain users in the real explanation group may go through the user study without seeing the real explanation interface a single time.
"False attempts" were added to the initial training data of the recommender system to solve this problem. Although this is not ideal, the benefits outweigh the disadvantages. This way, all users in the real explanation group can experience the real explanation interface before filling in the questionnaire.
The false attempts were appended in a controlled manner. First, prior data from Ooge's Wiski was analyzed to calculate the distribution of attempts needed to solve an exercise. The proportions can be seen in Fig. 3.10. Ten fake accounts (user ids) were then created. As it is suspicious that the histogram always has the same number of users, each fake account has a 50% probability of attempting an exercise. The number of attempts the fake account needs is calculated based on the probability distribution found earlier.
These false attempts influence the ordering of the recommended exercises. However, the end-users should not experience a significant difference than if there were no false attempts added. If absolutely no prior data is present for a section, the recommendation algorithm only recommends exercises in order of exercise number.

3.7 Questionnaires

3.7.1 Pre-Study Questionnaire

The pre-study questionnaire was used to gain insight into general information about a participant. Questions ranged from asking about demographics to proficiency with computers. The questions asked in this questionnaire can be seen in Appendix B, Table B.1.

3.7.2 Post-Study Questionnaire

Section 2.3.3 depicts the various ways of measuring user trust in the literature. In this thesis, trust was measured through a questionnaire. This section outlines what questions were used and why these questions were selected. It is important to note that the emphasis of the thesis is on initial user trust, and thus we did not measure trust at various points in time.
Various questionnaires exist in the literature to measure user trust. In this research, trust (multi-dimensional) was measured through trusting beliefs, intention to return, and perceived transparency. One question also explicitly asked about the user's trust (one-dimensional trust) in receiving recommendations from the system.
Wiski can be seen as a recommender agent that recommends math exercises to the end-user. Measuring trust through these constructs aligns with how other recommender systems/agents are evaluated [9, 18, 21, 10, 29] in comparison to the more XAI oriented questionnaires [36].
As shown earlier, trusting beliefs are an amalgamation of competence, benevolence, and integrity. Trusting beliefs can thus be measured by evaluating each of these constructs individually and combining the results. The questionnaire used to measure competence, benevolence, and integrity was based on the one used in Wang and Benbasat [9]. Questions regarding intention to return, transparency, and explicit trust were constructed from scratch (along the lines of those in other questionnaires). All questions were measured on a 7-point Likert scale. The questions were written in Dutch to match the language of the participants of the research. Appendix B, Table B.2 shows the questions used in the post-study questionnaire in Dutch. Appendix B, Table B.3 presents the questionnaire translated to English.

Ideally, to allow for comparison with existing studies, the original questionnaire used in Wang and Benbasat could also be applied in this thesis. However, as also done in the literature (e.g., [10, 31]), a few changes had to be made to make the questionnaire fit the current context. Two significant modifications were made to the trusting beliefs questionnaire used by Wang and Benbasat to apply it in this research. First, as the recommendation agent in Wang and Benbasat's research recommended digital cameras, the questionnaire needed to be adjusted to evaluate an e-learning platform that recommends math exercises. Furthermore, as stated earlier, the users of the platform were Belgian high school students in Flanders. Thus, the questionnaire must be translated into Dutch. As a consequence of the students' ages,

certain vocabulary from the original questionnaire were also required to be simplified to allow the participants to understand the questions. Despite these changes, the questions do continue to focus on the respective constructs of trusting beliefs. The original questions from Wang and Benbasat's questionnaire and remarks concerning the modifications can be seen in Appendix B, Table B.4. For example, virtual advisor was replaced with Wiski to fit the context better. Product recommendations / digital cameras was replaced with math exercises for the same reason. Note that the section titles (competence, benevolence, ...) were not present in the questionnaire.

3.8 Analysis of Results

To analyze the results, we must decide between using parametric or non-parametric statistics. As we are mainly dealing with Likert scales, non-parametric statistics are usually the primary choice (although arguments can be made for using parametric tests [57]). We, therefore, utilize non-parametric tests, similar to other work in this domain such as that of Cramer et al. [21]. One commonly used test is the *Mann-Whitney U test*. The following information concerning the Mann-Whitney U test stems from Laerd Statistics' article [42]. The Mann-Whitney U test is a non-parametric test that can be used when the data is ordinal (Likert scale), there are "two categorical, independent groups" (e.g., explanations vs. no explanations), and there is "independence of observations". We must further make an assumption regarding the distribution of the underlying data to interpret our results correctly. If we assume that the underlying distribution of the data is equal, the result of the Mann-Whitney U test can be interpreted as testing for a difference of medians. However, if we cannot make this assumption, the result must be interpreted as a difference in distributions.
We can choose between Kendall's τ and Spearman's ρ to test for correlations in non-parametric statistics. According to Abdi et al. [1], Kendall's τ can be interpreted "as the difference between the probability for this[sic] objects to be in the same order ... and the probability of these objects being in a different order ... ." Kendall's τ is known to have lower values compared to Spearman's ρ and also is usually preferred when working with smaller sample sizes. We thus opt to use Kendall's τ for correlation analysis, both for ordinal-ordinal values and ordinal-continuous values.
To assist with the analysis of our results, we use a combination of Python, Pandas [47], and SciPy [74]. For the visualization of our results, we use a combination of matplotlib [39] and seaborn [75].

Chapter 4

Development

This chapter covers the iterative development process of the platform and an overview of Wiski's backend. Wiski has undergone a total of four complete iterations, from the first low-fidelity prototype to the final proof of concept (Fig. 4.1).

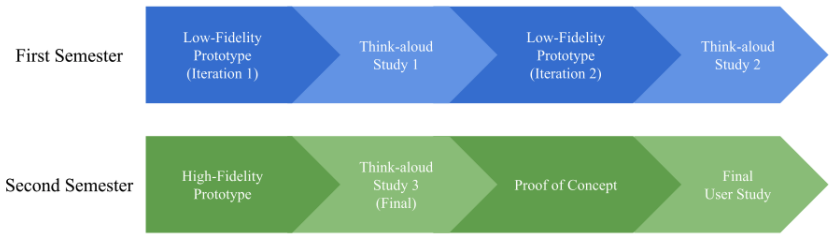


Figure 4.1: Four stages of Wiski and the accompanying think-aloud studies / user study.

4.1 Iterative Development Process

Wiski has been developed using the *User-Centered Design process*. As the name suggests, the user and their needs play a central role in the application's development. Wiski's target audience is Belgian high school students, as the exercises on Wiski align with their curriculum. As one may expect, this target audience is very diverse. It is easy to fall out of touch with these students during the development process, especially due to the possible discrepancies in values and prior knowledge. *Usability tests* can be conducted to alleviate this problem. These tests give the developer a chance to catch mistakes early on and incorporate important feedback into the platform. Wiski has undergone four complete iterations, from the first low-fidelity prototype to the final proof of concept. Between each iteration, a usability test was conducted. For Wiski, we opted to use think-aloud studies for the usability test at each iteration. *Think-aloud studies* are one of the most convenient yet effective ways of conducting a usability test. Participants are asked to do a series of tasks using

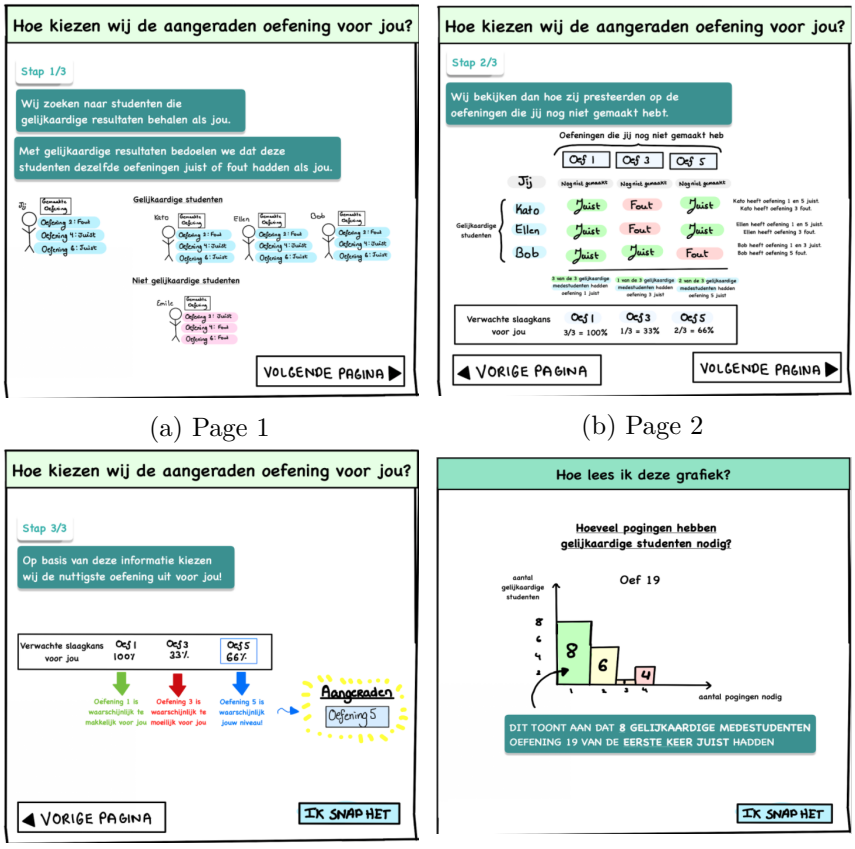
the prototype while "thinking aloud". The observer (developer) listens and watches the participant's reasoning and actions closely. This information is used as feedback for the prototype.

4.2 Low-Fidelity Prototype

Low-fidelity prototypes establish high-level concepts of an application. These prototypes, usually nothing more than sketches on paper, are made early on in the design process and force the developer to think of the application's flow and general look. These prototypes can then be used in usability tests. The low-fidelity prototype has gone through a total of two iterations. Throughout the development of the first low-fidelity prototype, four university students have been consulted to test basic usability principles. Between iterations, past participants of think-aloud studies have been consulted to approve of changes and additions. The prototype (second iteration) can be seen in Figs. 4.2 to 4.4. As this implementation builds upon Ooge's Wiski, certain design elements have been recycled from their website (Fig. 4.3 and Fig. 4.5a).

Think-aloud studies were conducted for each iteration of the low-fidelity prototype to understand the students' needs and product usability. The questions asked during the think-aloud studies can be split into three categories: task-oriented questions, understanding-oriented questions, and feedback-oriented questions. Task-oriented questions focused on the flow of the website, testing whether given tasks on the website can be efficiently completed. Understanding-oriented questions gauged how well users understood certain elements on the website. These questions primarily focused on the explanation interface to see if high school students comprehended the given explanations. Finally, feedback-oriented questions allowed users to provide general feedback about the prototype. They further provided insight into their values. The questions used for the think-aloud studies can be seen in Appendix C, Table C.1. The most relevant feedback obtained from the think-aloud studies can be seen in Appendix C, Table C.2.
Five participants (1 teacher, 3 high school students, 1 middle school student) were recruited for the first think-aloud study. The main problems (as expected) came from the explanation interface and the transparency pages. The first think-aloud study showed the importance of explicit language. Participants suggested adding extra information to clarify concepts displayed on the screens.
The second think-aloud study had seven participants (4 middle school students and 3 high school students). A first glance at the feedback matrix shows little to no improvements made. However, many of the understanding-related problems were only present with middle school students. For example, the high school students had no problem understanding the histogram. As Wiski's target audience is high school students, we decided to leave the histogram as is. The following think-aloud study (high-fidelity prototype) that comprised of participants at least in high school showed that none of the participants had a problem interpreting the histogram.
Little to no problems were observed for the main flow of Wiski, which is mostly likely

due to it not differing much from the one used in Ooge's Wiski. Their thesis [60] can be referred to for the usability tests conducted to develop their version of Wiski.



(a) Page 1

(b) Page 2

(c) Page 3

(d) Histogram Interpretation

Figure 4.2: The transparency pages used to explain how collaborative filtering works to the users as well as the explanation for how to read a histogram.

4.3 High-Fidelity Prototype

The high-fidelity prototype translates the low-fidelity prototype and the information obtained from the accompanying studies to a digital, functional prototype. This prototype should resemble the final product closely to be able to catch any potential (significant) shortcomings of the application before release.
One final think-aloud study with three high school students and one university student was conducted to test the usability of the high-fidelity prototype.
The questions asked, marked with HF, can be seen in Appendix C, Table C.1. Further questions were also asked regarding the wording of the post-study questionnaire. The feedback matrix for the think-aloud study can be seen in Table C.3. Some of the feedback regarding the questionnaire can be seen as remarks in Table B.2. Although

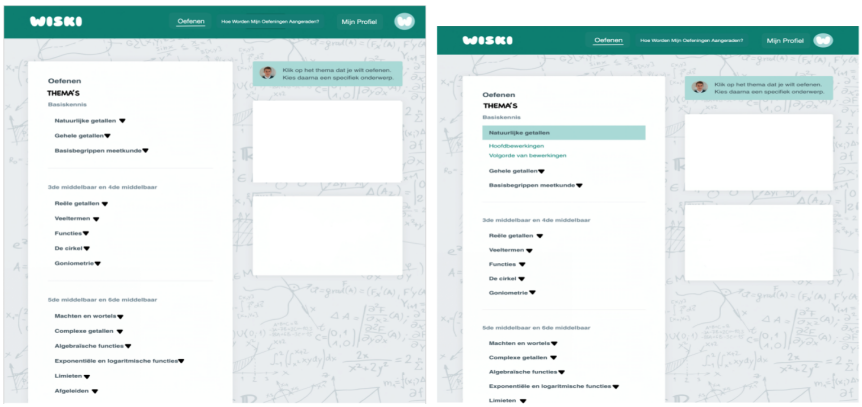
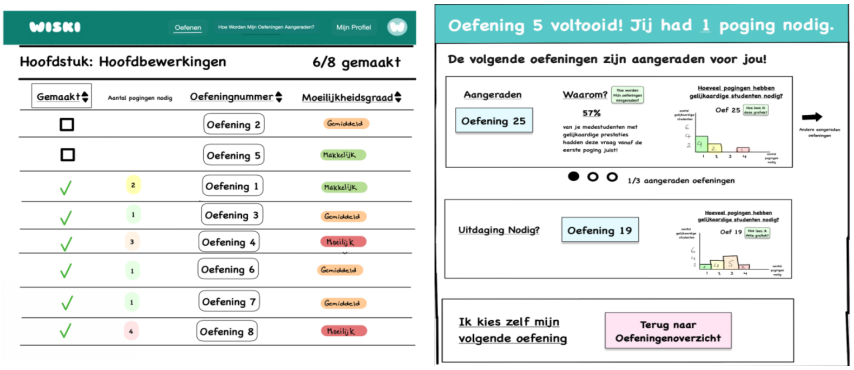


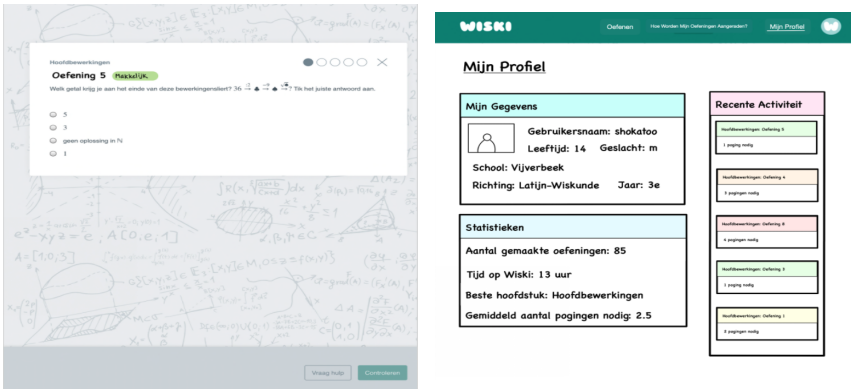
Figure 4.3: Page with subjects and sections in the low-fidelity prototype. Sections only appear after clicking on the subject. This page was reused from Ooge's Wiski.



(a) Exercise List

(b) Explanation Interface

Figure 4.4: The list of exercises for a particular section and the explanation interface.



(a) Exercise Interface

(b) Profile

Figure 4.5: The screen seen when solving an exercise and the profile page. The screen in Fig. 4.5a is built upon on the one used in Ooge's Wiski.

the sample size used for this think-aloud study was small, it showed that the flow of the website and its usability were satisfactory. All users could navigate through the website without a problem, and the feedback was mostly concerned with small details.

4.4 Significant Changes

A close look at the figures of the low-fidelity prototype and the final deployed website show a few significant changes/omissions. This section is devoted to explaining these changes.

Collaborative Filtering Transparency Pages. One notable omission in the final version of Wiski is the (for the lack of a better word) "transparency pages" (Figs. 4.2a to 4.2c) explaining the collaborative filtering algorithm to the end-user. During the low-fidelity prototype stage, a significant amount of time was put into designing a comprehensible explanation of the algorithm to provide extra transparency to the end-user. However, even after the second iteration, it was clear that not all students understood the content on the transparency pages. Select students also communicated during the think-aloud studies that this extra transparency did not matter much to them. Some students appreciated its presence, but they too admitted they would probably not look at it. Finally, it may be possible to give too much transparency [79, 58]. These factors led to the decision to remove the transparency pages from the final prototype.

How to Read a Histogram. Another omission is a separate page that explained how to interpret the histogram (Fig. 4.2d). The histogram only formed a problem for middle school participants in the second think-aloud study. High school participants of the think-aloud studies had little to no trouble interpreting the figure, as long as an explicit title was shown above it.

Justification-explanation. The justification-explanation has also been modified from the low-fidelity prototype (Fig. 4.4b). The original explanation communicated the percentage of fellow students that solved the exercise on the first attempt. This explanation can be deduced from the histogram explanation and is thus redundant. As the recommender system's output is the *estimated number of tries*, it made more sense to use this as an explanation instead. One past participant of the think-aloud studies mentioned that it was difficult to relate to a percentage. Three past participants of the think-aloud studies were asked which explanation they preferred. All three preferred this explanation as it gave "extra" information compared to the past version. One user stated they liked it as they could "confirm" whether the predictions were accurate or not. This emotion was further shared by two participants of the final think-aloud study.

Profile Page. We decided to omit a full profile page (seen in Fig. 4.5) due to (*i.*) lack of time and (*ii.*) to focus more on the main goals of the research.

4.5 Explanation Interface

The explanation interface plays a central role in this thesis. This section explains the reasoning behind the final explanation interface. Fig. 4.6 shows a marked explanation interface which we use to refer to specific elements of the explanation interface.
As stated by Mohseni et al. [54], it is essential to identify the type of audience when developing an XAI system. Wiski is a platform where high school students can practice mathematics. Therefore, the audience is AI novices. Furthermore, it is also crucial to keep the main goal in mind. According to Tintarev and Masthoff's outlined goals [70], we primarily aim to increase user trust. Going by Adadi et al.'s goals [4], we primarily aim for justification (why may this exercise be a good match) and transparency (the exercises are recommended based on the user's level).
One main pattern that can be observed throughout this interface is the use of explicit language. The various usability tests have shown that high (and middle) school students require (and want) information to be communicated explicitly to comprehend the material faster and better. This is achieved at the expense of adding redundant information throughout the interface. The following shows the translation of various components in the explanation interface and the reasoning behind the choice of words.

Component a
Translation: Solve a recommended exercise from the same section.
Remarks: Feedback from three participants of the final think-aloud study showed that it was unclear from what section the recommended exercise came.

Component b
Translation: Recommended
Remarks: Observing the participants' behavior during the final think-aloud study showed that the participants understood that there are three exercises recommended after reading the "Aangeraden" (Recommended) title. Participants of the final think-aloud study would typically read the title aloud (during the think-aloud study) and communicate that they think the three exercises listed here are recommended. No official A/B testing was conducted to test the effectiveness of the title. However, all participants agreed that it adds value to the interface.

Component c
Translation: Why this exercise? Wiski thinks that your current level matches that of this exercise.
Remarks: The subtitle "Why this exercise" was essential for the participants to understand that this is an explanation for why the exercise has been recommended.

Component d

Figure 4.6: Marked version of the explanation interface, delineating the various elements contained.

Translation: Wiski expects you to need 1 or 2 attempts to solve exercise 21 correctly, based on your results and those of your fellow students.
Remarks: This explanation gives students insight into the estimated number of tries they need to solve the exercise and therefore allows users to make informed decisions about the difficulty level of an exercise. 2 of the 4 participants of the final think-aloud study communicated that accountability was an added benefit: users could confirm how accurate the algorithm is by comparing the estimated number of tries and the actual number of attempts. Participants from both the low- and high-fidelity prototype think-aloud studies also claimed that the transparency level of the explanation was satisfactory.
The exercise number is also stated explicitly as user studies with the low-fidelity prototype showed that some students mistakenly interpreted these numbers to be for the exercise they just solved.

Component e
Translation: Number of attempts fellow students needed to solve exercise 21 correctly.
Remarks: By making the title for the histogram more explicit, the extra page describing how to read the histogram could be avoided. Once again, the exercise number is explicitly stated here as user studies with the low-fidelity prototype showed that some students thought that the histogram indicated how students did on the

exercise the user just solved. Some participants of the think-aloud studies appreciated how much information could be transferred efficiently and the visual appeal it added.

4.6 Takeaways

We briefly summarize the main takeaways observed from the development process. These are observations made from the think-aloud studies conducted during this research and thus may not be applicable in a general context. However, it may be interesting to keep the following in mind when designing a similar application for a similar target audience.

- We observed that being explicit is very important as it allows the users to process the information quickly and easier. Redundancy may thus be advantageous for this audience.
- Quite a significant difference between high school students and middle school students was observed. Therefore, these user bases should not be grouped into one category.
- Participants of the think-aloud study often stated that they would not go out of their way to view extra pages concerning how the platform recommends exercises or how to interpret the histogram. Thus, developers should think twice before adding such pages to their platform.

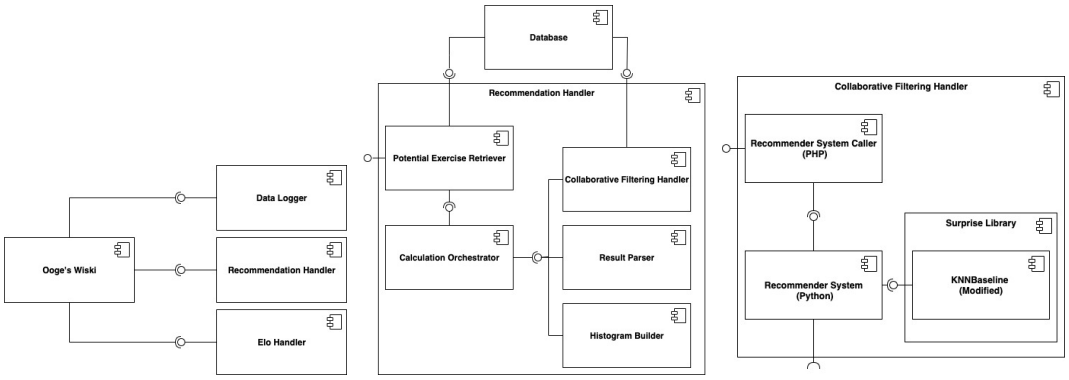
4.7 Technical Implementation

This section gives an overview of the technical implementation of Wiski. Wiski is developed in Drupal7[1]. Ooge's thesis [60] can be referred to for information regarding how Wiski utilizes Drupal7. The languages used are HTML, CSS, and JavaScript for the frontend and PHP (with a small amount of Python) in the backend. Here, we take a look at Wiski's backend architecture, highlighting the various components contributing to the added features.

4.7.1 Backend Architecture

Fig. 4.7a gives a visualization of the current Wiski's architecture. Three main components have been added to the backend of Ooge's Wiski to obtain the current version of Wiski: the Elo Handler, the Recommendation Handler, and the Data Logger. The Elo Handler is responsible for the updating of the Elo ratings. The Recommendation Handler handles the various steps needed to recommend exercises to the end-user. Finally, the Data Logger stores both data used by the recommender system and the data used for later analysis in this work.

[1] https://www.drupal.org/drupal-7.0



(a) High-level architecture (b) Recommendation Handler (c) CF Handler

Figure 4.7: A basic overview of Wiski's architecture.

Elo Handler

The Elo Handler comprises methods concerning the updates of the users' and exercises' Elo ratings. The Drupal Rules module calls the Elo Handler and is configured such that when an answer is submitted for a multiple-choice problem (the event), the Elo handler is called (the action). As users must repeat an exercise until it is correctly answered, Elo is only awarded to the user (and deducted from the exercise) if the exercise is answered correctly on their first attempt. On the other hand, users continually lose Elo (and exercises gain Elo) if exercises are repeatedly answered incorrectly.

Recommendation Handler

The Recommendation Handler is responsible for recommending the subsequent exercises to the end-user. This process is based on the algorithm used in Dahl and Fykse [23]. The sub-components of the Recommendation Handler can be seen in Fig. 4.7b. This handler works as follows:

1. *Potential next exercises* are retrieved from the current section (the section of the exercise the user just solved). Potential next exercises are n exercises that (*i.*) have not yet been solved by the user, and (*ii.*) have an Elo rating closest to Elo_U (see Section 3.5). A *max-heap* is used to find n exercises with Elo ratings closest to Elo_U. This way, the n elements in the max-heap at the end of the computation are the n exercises that satisfy requirement 2.
2. The *Calculation Orchestrator* receives the potential next exercises. This orchestrator passes each component's results to the following component returning the final result to the application's frontend.
3. The orchestrator passes the data to the *collaborative filtering handler* (Fig. 4.7c), which consists of a PHP component and a Python component (script). The Python script is called from the PHP backend to execute collaborative filtering.

The collaborative filtering algorithm is implemented by Surprise [38], a library that provides a convenient way of implementing recommender systems. A MySQL connector is used to connect to Wiski's database. Pandas [47] uses this connection to read a table that contains the number of attempts the user with *user-id* needed to solve an exercise with *node-id* as a Pandas dataframe. The collaborative filtering algorithm is trained on this dataframe, and the estimated number of tries the user needs for each potential exercise is returned. The Surprise library's KNNBaseline algorithm has been modified to return the neighbor's attempts for the potential exercise. This information is later used in the explanation interface.

4. The *Result Parser* extracts and interprets the data from the Python script. Once the data is parsed, the n potential exercises are sorted by the estimated number of tries in ascending order. The *Histogram Builder* transforms the users' attempts into a histogram-compatible form.

Chapter 5

Results

This chapter presents the results of the user study. In total, 37 high school students participated in the research. 12 students received real explanations, 12 students received placebo explanations, and 13 students received no explanations. This is a relatively small sample size and should be taken into account throughout the chapter.

5.1 Responses

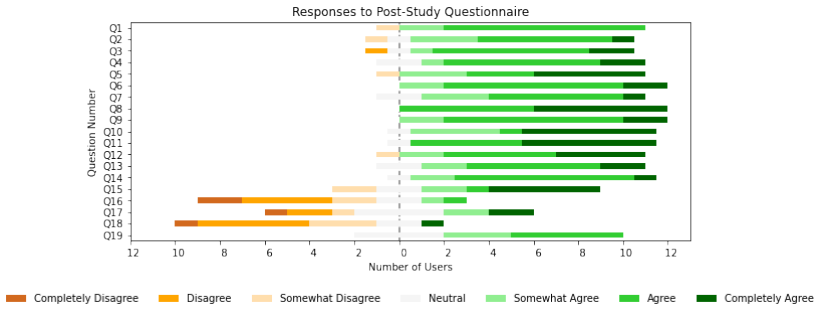
The responses for each question of the post-study questionnaire (Table B.2 (Dutch), Table B.3 (English)) can be observed in Fig. 5.1. Box plots of the responses can be seen in Fig. 5.2. The mapping from constructs and measures to question numbers are displayed in Table 5.1, alongside a reference to their respective box plots.

Table 5.1: Table displaying the the trust constructs and measures mapped to their respective question numbers. A further reference is given to the box plot displaying their results.

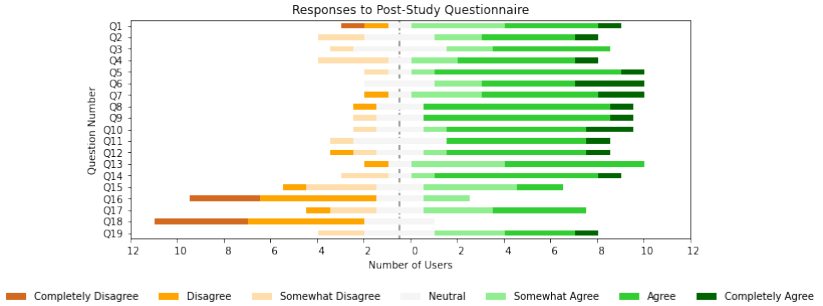
Construct / Measure	Question Number	Box Plot
Competence (C)	Q1-Q5	Fig. 5.2a
Benevolence (B)	Q6-Q8	Fig. 5.2b
Integrity (I)	Q9-Q11	Fig. 5.2c
Trusting Beliefs (TB)	C+B+I	Fig. 5.2f
Intention to Return (ITR)	Q13-Q14	Fig. 5.2d
Perceived Transparency (PT)	Q15	Fig. 5.2e
1D Trust	Q12	Fig. 5.2g
MD Trust	TB+ITR+PT	Fig. 5.2h

5.2 Explanation Interface

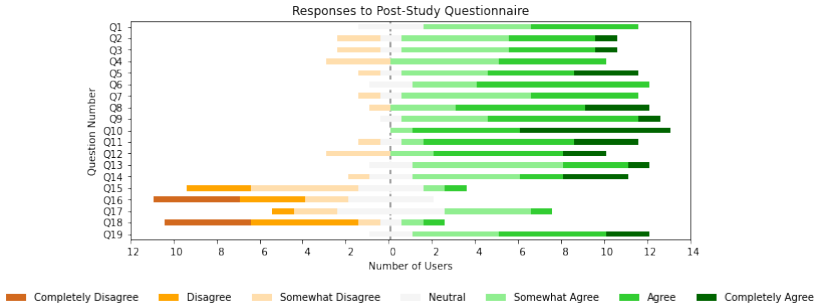
Real Explanation vs. No Explanation We use the one-sided Mann-Whitney U test to compare the results obtained from the IRE group to those from the INE



(a) Interface for Real Explanation Group



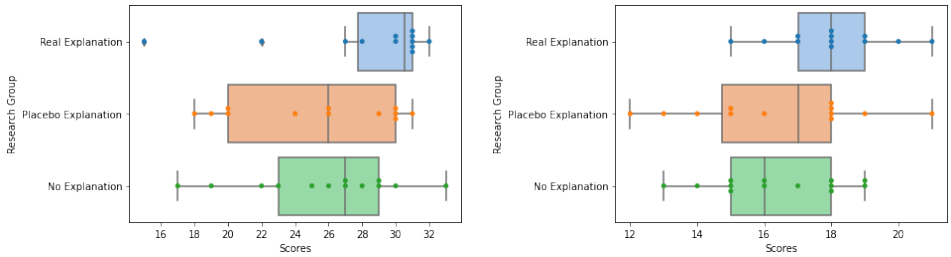
(b) Interface for Placebo Explanation Group



(c) Interface for No Explanation Group

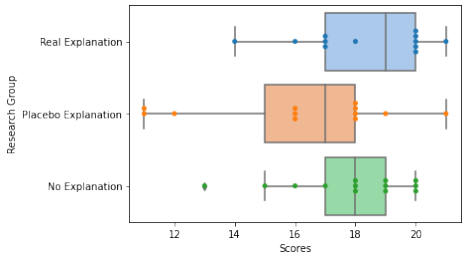
Figure 5.1: Diverging bar chart for the responses to the post-study questionnaire. The question numbers map to those in Table B.2 (Table B.3 for English). Competence refers to Q1-Q5, benevolence to Q6-Q8, integrity to Q9-Q11, intention to return to Q13-Q14, perceived transparency to Q15, and one-dimensional trust to Q12.

group. The one-sided test checks for a shift in a specific direction. As we are looking for whether the IRE performs better than the INE situation, a one-sided test is appropriate. We further make the significant assumption that the underlying distributions are equal to one another. The null hypothesis and the alternative hypothesis for the one-sided Mann-Whitney U test are as follows [42]:

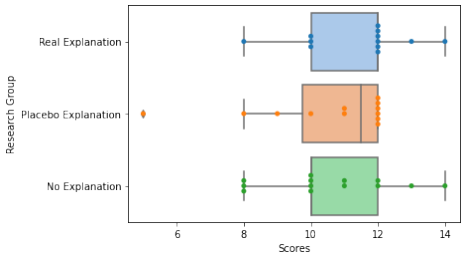


(a) Competence

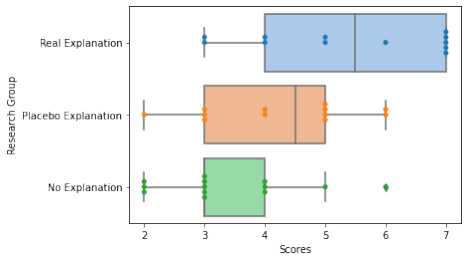
(b) Benevolence



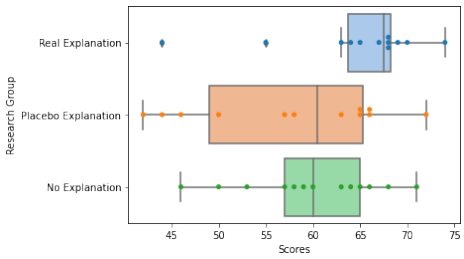
(c) Integrity



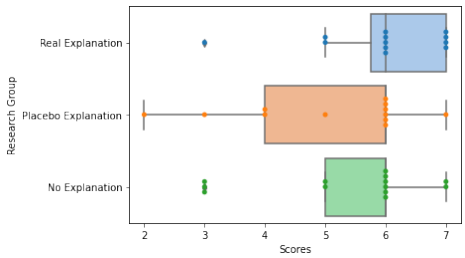
(d) Intention to Return



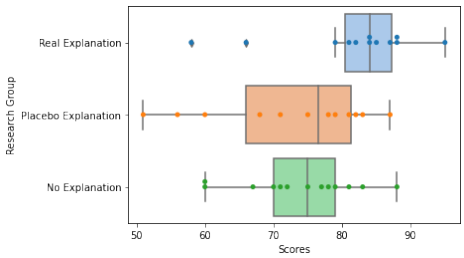
(e) Perceived Transparency



(f) Trusting Beliefs



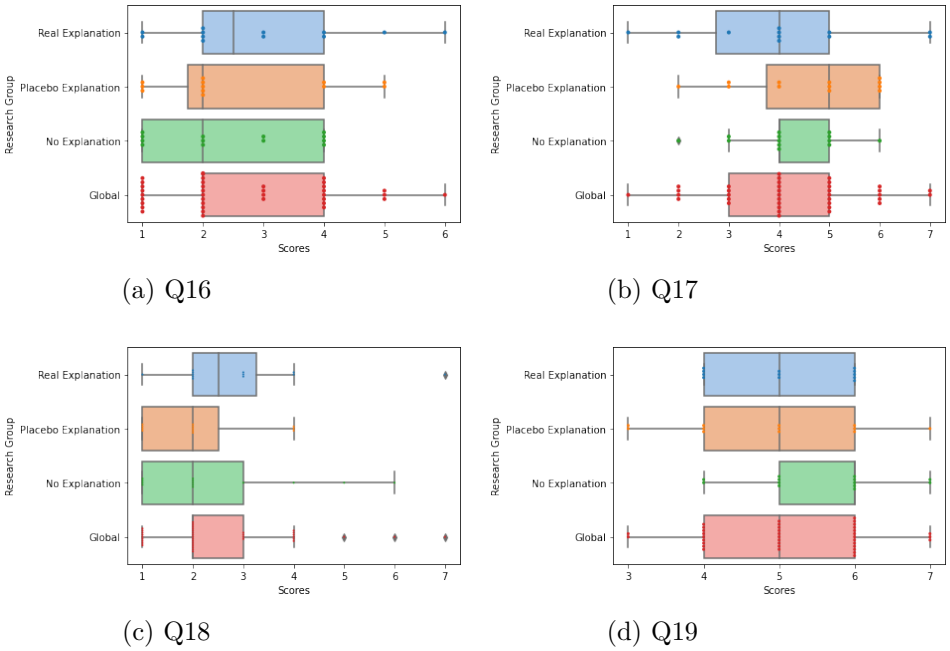
(g) 1D Trust



(h) MD Trust

Figure 5.2: Box plots of the responses from the post-study questionnaire related to trust.

Hypothesis H_0. $X_{\text{IRE}} \sim X_{\text{INE}}$ ("*the distributions of the two groups are equal*")
Hypothesis H_1. $\text{med}_{\text{IRE}} > \text{med}_{\text{INE}}$ (*the median of the IRE is larger than the INE*)



(a) Q16 (b) Q17

(c) Q18 (d) Q19

Figure 5.3: Box plots of the responses from the post-study questionnaire for questions 16 to 19. "Global" indicates the results from the three research groups combined.

The results of the Mann-Whitney U test can be seen in Table 5.2. Alongside the p-value, the table also shows the U test statistic and the Common Language Effect Size[1]. The results are statistically significant ($p < 0.05$) (we can reject the null hypothesis) for competence, benevolence, trusting beliefs, perceived transparency, and multi-dimensional trust. Only perceived transparency is significant for $p < 0.01$. There is no statistically significant evidence that using explanations increases one-dimensional trust. The box plots in Fig. 5.2 indicate that this is mostly due to the high values from the IPE and INE group rather than low results from the IRE group.

Real Explanation vs. Placebo Explanation The one-sided Mann-Whitney U test is once again utilized to investigate the effects of the IRE compared to the IPE. We are once again interested in whether the IRE obtains higher trust scores than the IPE. The null and alternative hypotheses are similar to those of the IRE vs. the INE. The results of the test are shown in Table 5.3. We can observe that the results are statistically significant ($p < 0.05$) for competence, trusting beliefs, perceived transparency, and multi-dimensional trust. No results are significant for $p < 0.01$. Furthermore, a particularly large p-value for one-dimensional trust can be observed.

[1] "*The probability that a randomly selected score from the one population will be greater than a randomly sampled score from the other population.*"[77]

Table 5.2: Results of the one-sided Mann-Whitney U test for the group with the IRE vs. the group with the INE.

	p-value	U value	Common Language Effect Size
Competence	0.029*	113.0	0.724
Benevolence	0.030*	112.5	0.721
Integrity	0.261	90.0	0.577
Trusting Beliefs	0.038*	111.0	0.712
Intention to Return	0.109	100.5	0.644
Perceived Transparency	0.002**	130.5	0.837
One-Dimensional Trust	0.137	97.5	0.625
Multi-Dimensional Trust	0.014*	119.0	0.763

$^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.3: Results of the one-sided Mann-Whitney U test for the group with the IRE vs. the group with the IPE.

	p-value	U value	Common Language Effect Size
Competence	0.023*	37.5	0.740
Benevolence	0.074	47.0	0.674
Integrity	0.054	44.0	0.694
Trusting Beliefs	0.030*	39.0	0.729
Intention to Return	0.139	54.0	0.625
Perceived Transparency	0.041*	42.0	0.708
One-Dimensional Trust	0.937	47.5	0.330
Multi-Dimensional Trust	0.013*	33.0	0.771

$^{**}p < 0.01$, $^{*}p < 0.05$

Placebo Explanation vs. No Explanation The box plots in Fig. 5.2, show that these two groups have quite similar values, except for integrity and perceived transparency. In fact, we observe that the median integrity for the INE group is higher than that of the IPE group.
We utilize a two-sided Mann-Whitney U test to analyze the results for the IPE group against those of the INE group. A two-sided test is more appropriate in this case, as we are not interested in a shift in a particular direction. The null hypothesis and alternative hypothesis are as follows [42]:

Hypothesis H_0. ($X_{\text{IPE}} \sim X_{\text{INE}}$) ("*the distributions of the two groups are equal*")
Hypothesis H_1. $\text{med}_{\text{IPE}} \neq \text{med}_{\text{INE}}$ ("*the medians of the two groups are not equal*")

The results of the two-sided Mann-Whitney U test reflect the observations from the box plot (the table can be observed in Appendix D, Table D.1). No results are

significant for $p < 0.05$. The two smallest p-values are 0.099 and 0.143 for perceived transparency and integrity respectively. The other values are all above 0.696 further indicating that the two distributions are similar to one another.

General Observations Interestingly, the spread of the responses for many questions in the IPE group is relatively high compared to those of the other two research groups. Competence and trusting beliefs for the INE group and perceived transparency for the IRE group also have high spread for their responses.
Intention to return seems to be nearly identical amongst the three groups according to Fig. 5.2d.
The box plots of the general questions (Q16-Q19) (Fig. 5.3) show that the responses to these questions were fairly similar across the three groups. The red box plot in the figures mentioned above shows the responses of the three groups combined. The median participant

- *disagrees* (2) that they do not want to receive explanations when using Wiski.
- is *neutral* (4) concerning thinking that receiving explanations for why a math exercise has been recommended is more important than receiving an explanation for why a movie has been recommended.
- *disagrees* (2) that they were not happy with the level of exercises they have been recommended.
- *somewhat agrees* (5) that they, in general, would like to receive explanations when something is recommended.

5.3 Qualitative Data

The participants were required to provide some form of explanation for their response to Q15 (perceived transparency). They were further free to give explanations to their responses for each construct and give general remarks at the end of the questionnaire. Here, we present some of the interesting textual responses received. As the responses are in Dutch, an accompanying translation (as literal as possible and to our best efforts) is provided next to the original statements. Furthermore, we add what the users answered for Q15 next to the respective responses (completely disagree = 1 - completely agree = 7).

IRE

Some users were positive about the explanations for why a certain exercise had been recommended. This was reflected in responses such as

- *"Zo kan ik mezelf ook beter inschatten."* (7)
  (This way, I can also estimate myself better.)

- *"De uitleg die Wiski gaf, vond ik wel kloppen en voldoende."* (7)
  (I found the explanation that Wiski gave correct and satisfactory.)

- *"Ik kreeg voldoende uitleg waarom ik deze oefeningen kreeg. En had hier veel aan."* (6)
  (I received enough explanation as to why I received these exercises. And I found it really useful.)

However, some participants were maybe not satisfied with the explanations and may have wanted a different type of explanation. This can be seen in the following responses:

- *"Er staat toch gewoon hoeveel pogingen Wiski denkt dat je zal doen om het juiste antwoord te vinden. Het legt niet specifiek uit."* (3)
  (Doesn't it just state how many tries Wiski thinks I would need to find the correct answer. It doesn't explain specifically.)

- *"Tot nu toe is de reden duidelijk, deze oefening is van een zelfde moeilijkheidsgraad. Ik weet niet of het complexere uitleg kan geven op basis van bijvoorbeeld regelmatige fouten. Daarvoor heb ik te weinig oefeningen gemaakt."* (5)
  (The reason is clear up until now, this exercise is of the same difficulty level. I don't know if it can give more complex explanations based on for example reoccurring mistakes. For this, I solved too few exercises.)

There is also evidence that some users did not require the explanations. This observation can be inferred from responses such as

- *"Ik heb de uitleg niet echt gelezen ..."* (4)
  (I didn't really read the explanation ...)

- *"Ik heb niet echt opgelet wat er me aangeraden werd ..."* (5)
  (I didn't really pay attention to what was recommended to me ...)

- *"(Te) grote nadruk op waarom ik een oefening krijg ..."* (5)
  (Large (too large of an) emphasis on why I received an exercise ...)

There was also one participant giving a contradictory response, selecting "Somehwat disagree" for Q15 but responding "goede uitleg" (good explanation).

IPE

The responses within this group were also quite mixed. Some students indeed did not perceive the placebo explanations as explanations, giving responses such as

- *"Wiski zegt gewoon, kijk hier is een oefening die je kan maken."* (2)
  (Wiski just says, look here is an exercise that you can solve.)

- *"Wiski zegt gewoon 'berekend door het algoritme van..."* (3)
  (Wiski just says calculated by the algorithm of ...)

However, we also see that some students were satisfied with the placebo explanation. This is indicated by responses such as

- *"zo weet ik waarom ik dat krijg"* (6)
  (This way I know why I received it)
- *"Het is duidelijk waarom ze voor de volgende oef kiezen"* (5)
  (It is clear why they chose for the next exercise)
- *"Wiski zegt dat het algoritme de volgende oefening aanraadt dus ik vertrouw het algoritme."* (6)
  (Wiski says that the algorithm recommends the next exercise thus I trust the algorithm.)

One participant stated that they did not require extra transparency by stating *"Ik denk niet dat er meer uitleg over waarom de oefening wordt aangeraden nodig is."* (5) (I don't think that there needs to be more explanation as to why an exercise has been recommended).
However, one respondent did communicate that they would have wanted a better explanation, as they responded *"het zou fijn zijn voor een uitgebreide uitleg waarom het beter is om die oefening te maken."* (3) (it would be nice for an extensive explanation as to why it is better to solve this exercise).

INE

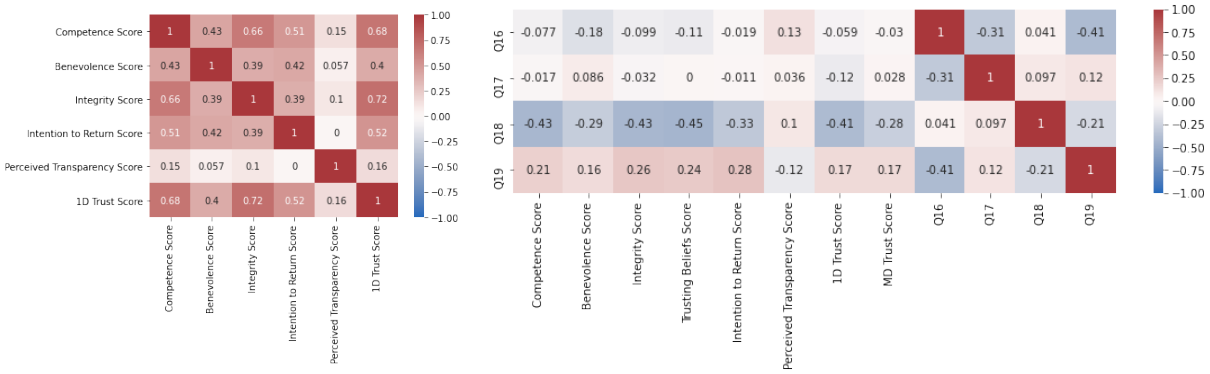
The responses within this group were reasonably consistent. Close to all users who gave responses indicated that they did not see an explanation or maybe missed it. One user gave the following response at the end of the post-study questionnaire: *"Ik vind het een heel gebruiksvriendelijke site, ik kan er goed mee omgaan en het verloopt volt[sic]. Ik vind het wel spijtig dat er niet staat waarom een bepaalde oefening aangeraden wordt. Het is fijn om te weten waarom die oefening bij jou past, maar er moet ook niet te veel info in staan want dan is het niet meer leuk om te lezen."* (3) (I find it a very user-friendly site, I can use it well and it runs smoothly. I find it unfortunate that it does not explain why an exercise was recommended. It is nice to know why an exercise is recommended for you, but there should also not be too much information as then it would not be fun to pleasant to read.)

One interesting perspective comes from the following response: *Ik had bij de vorige vragenlijst aangeduid dat ik mezelf 'gemiddeld' vind in wiskunde, en ik heb (daardoor denk ik maar weet ik niet zeker) enkel 'gemiddeld' oefeningen gekregen. Misschien dat het beter zou zijn als je bijvoorbeeld net begonnen bent met studeren dat je enkele makkelijke oefeningen kan maken om het 'op te frissen' of 'met een goed gevoel te beginnen' of andersom, dat je graag jezelf uitdaagt en een moeilijke oefening wil proberen (als de aanbevelingen van gemiddeld niks te maken hadden met wat ik had aangeduid op die vorige enquete dan mag je deze uitleg gewoon negeren :))* (3) (I filled in that I find myself 'average' in math, and I (this is what I think

but I'm not certain) only received 'average' exercises. It may be better if you, for example, just start studying that you solve a few easy exercises to 'warm up' or to 'start with a good feeling' or on the other hand, that you want to challenge yourself and attempt a difficult exercise (if the recommendations with average did not have anything to do with what I selected in the previous questionnaire then you may just ignore this explanation:)).

Finally, there were two users from this group that thought they received explanations according to their responses to Q15. These are also the two users that gave the two highest values (somewhat agree and agree) for this question. One user wrote "*Als je een nieuwe oefening wilt maken is het handig dat je weet waarom deze oefening aangeraden wordt, dit doet de website goed.*" (6) (If you want to solve a new exercise, it is useful that you know why this exercise is recommended, the website does this well). The other participant stated "*Ja ik vind dat er genoeg uitleg is.*" (5) (Yes I find that there is enough explanation).

5.4 Correlations



(a) Trust Correlations

(b) General Question Correlations

Figure 5.4: Correlation matrices

Fig. 5.4a shows Kendall's τ between the various constructs of trust and one-dimensional trust. The matrix indicates that competence and integrity are the two constructs that are correlated the most with one-dimensional trust. On the other hand, we see that perceived transparency has a very weak correlation with one-dimensional trust. In fact, perceived transparency in general has little to no correlations with any of the trust constructs.
One-dimensional trust is strongly correlated with trusting beliefs (0.673) and multi-dimensional trust (0.624), which shows that our trust measures are related to one another.

We do observe a moderate correlation between the satisfaction with the level of recommended exercises (Q18) and the various constructs and measurements of trust (Fig. 5.4b). The median participant (as shown in Section 5.2) did disagree that they were not satisfied with the level of exercises they were recommended, and we could observe that the responses amongst the three groups were fairly similar. Perceived accuracy should have, therefore, not played a significant role in biasing the results.

Finally, we also see a moderate correlation between wanting explanations from Wiski (Q16) and wanting explanations in general (Q19).

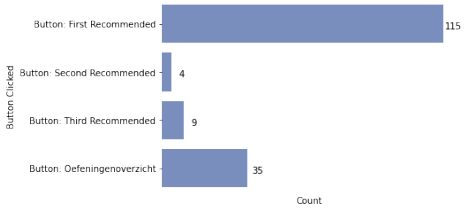
5.5 Recommendation Clicks

As mentioned earlier, Wiski logs whether the user clicks on a provided recommendation or not. The distributions of the first five interactions can be seen in Fig. 5.5a. Only the first five interactions are used as this is theoretically the number of times the participant sees the explanation interface before answering the post-study questionnaire. We define the acceptance of a recommendation as clicking on one of the three recommended exercises. *Button: Oefeningenoverzicht* indicates that the user explicitly chose not to select a recommended exercise and hence did not accept the recommendation. The figure shows that the first recommended exercise is clicked the most, followed by not accepting a recommendation. Interestingly, the second and the third recommendations are rarely chosen.
Fig. 5.5b depicts the proportion of times users from each research group accepted a recommendation for their first five (or four[2]) interactions. We removed all users that had less than four interactions.
The INE group tends to accept the recommendation less than users from the other two research groups. This is also confirmed by the one-sided Mann-Whitney U test, where (once again assuming equal underlying distributions) the median of both the IRE group ($p = 0.007$, U=67.0, CLES=0.827) and the IPE group ($p = 0.039$, U=72.0, CLES=0.727) are significantly higher.
One of the users that was not satisfied with the placebo explanation (as shown in their qualitative response) also had a low acceptance value.
Table 5.4 shows Kendall's τ between the acceptance of the recommendations with each of the constructs and various measures of trust. Little to no correlations can be observed in the table.

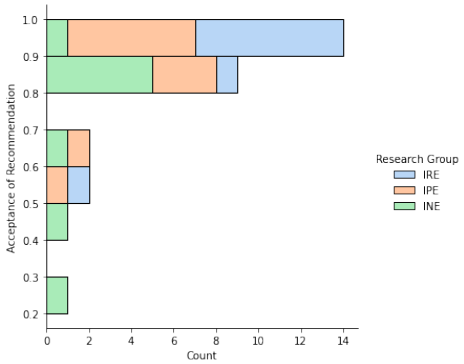
Table 5.4: Kendall's τ between acceptance of recommendations and various trust constructs and measures.

	Competence	Benevolence	Integrity	Trusting Beliefs	Intention to Return	Perceived Transparency	One-dimensional Trust	Multi-dimensional Trust
Acceptance	0.057	-0.017	-0.024	0.030	-0.043	0.253	-0.069	0.129

[2]It is important to note that there are limitations regarding the logging, which are discussed later in the thesis (Section 6.4). The exact values should thus be interpreted cautiously.



(a) Distribution of Clicks



(b) Proportion of Acceptance

Figure 5.5: Plots visualizing the logged data for the recommendation clicks. The left plot shows the distribution of the clicked buttons. The plot on the right depicts the acceptance rate (proportion of times the user accepted a recommendation).

Chapter 6

Discussion

This chapter discusses and analyzes the results outlined in the previous chapter. We further answer the three research questions presented in Chapter 3. Finally, the chapter outlines the study's limitations and how these limitations (may) have influenced the results. It is important to note that although some results may seem promising, we must be very cautious in their interpretation due to the small sample size, especially for quantitative data.

6.1 Effect of Real Explanations

Our first research question is concerned with whether using explanations can increase initial user trust in this specific context and target audience. The results show that using explanations has a statistically significant effect on increasing initial user trust (both in terms of trusting beliefs and multi-dimensional trust) compared to using no explanations and using placebo explanations. This is consistent with other related work concerning using explanations to increase user trust. However, we see no statistically significant evidence that one-dimensional trust is increased by using explanations, both when compared to the INE and IPE group. This result puts our findings for trusting beliefs and multi-dimensional trust into perspective. The discrepancy between the obtained results for multi-dimensional trust and one-dimensional trust relates back to the discussion held in Section 2.3.3. On one hand, this discrepancy may be further indication that multi-dimensional trust provides a more balanced view than its one-dimensional counterpart. On the other hand, the relatively high-valued responses obtained for the one-dimensional trust question across the three research groups may indicate that the explanation interface is not the most significant factor in determining whether to trust the system to recommend exercises in the short term. Users in the INE group and IPE group may base their initial trust upon other factors of the website, such as the perceived accuracy of the recommender system, quality of the exercises, and/or the platform's appearance. This statement is further backed by the observations from the correlation matrix. Perceived transparency as well as Q16, Q17 and Q19 have little to no correlation with

the one-dimensional trust score. Instead, constructs such as integrity and competence show a higher correlation. Therefore, using explanations for recommendations should be seen as a method to increase competence, which in turn affects initial user trust (according to multi-dimensional trust). This interpretation is similar to that of Berkovsky et al. [10]. The importance of perceived competence resembles many of the definitions we presented in Section 2.3.1.

We can also observe that perceived transparency of the IRE is somewhat controversial. This observation aligns with what was found in the qualitative data: not all users in this age group perceive the utility of the explanation interface the same way. The box plots in Figs. 5.3a, 5.3b and 5.3d further support this argument, where the opinions on the importance of explanations were quite divided amongst all of the participants. The qualitative responses also indicate that some users may have their own perception of what a good explanation is. Whereas this recommender system and explanation interface are solely focused on the difficulty level, some users may seek other forms of explanations. The solution to this problem may be to give users more control over the type of explanations they receive (which indirectly also changes the type of recommendation), or over whether they would like to see any explanations at all. Another solution can be to customize the interface according to personal characteristics [52, 10].

The values for intention to return are very similar across the three research groups, whereas we do see an increase in perceived competence. This result is different from what Pu and Chen state in their work [65]: "... the most remarkable benefit of the competence-inspired trust is its positive influence on users' intention to return. Accordingly, we regard the "intention to return" as an important criterion to judge the trust achievement of explanation-based recommendation interfaces." The discrepancy may stem from the users' age group, the context in which the recommender system was used, and/or the user study setup. The participants in Pu and Chen were mainly 20-30 years old, and the context in which the recommender system was used was e-commerce (notebooks and digital cameras). Intention to return may be influenced more by explanations when receiving assistance to make a significant purchase, as opposed to an e-learning platform where the importance of selecting the next exercise may not carry much weight. Furthermore, their user study only evaluated an explanation interface, whereas our research implements a fully functional e-learning platform. Thus, recommendation accuracy and/or outside factors (as mentioned before, e.g., exercise quality, appearance) may have made up for the lack of an explanation interface and, indirectly, the lower perceived competence score for the INE and IPE group. The discrepancy shows that it could be interesting to research how the differences between various contexts such as e-learning and e-commerce influence (initial) user trust.

Finally, the further importance of an explanation interface is still highlighted through the response of one user from the INE group. This participant formed their own idea as to how the system works, as they thought that the exercise that was

recommended depended on their answer to the pre-study questionnaire. Explanations mitigate such ambiguities in the platform.
In a similar way, the two users from the INE group that stated that they received explanations (Section 5.3) may have believed that the labels from the exercise selection page (Fig. 3.3) were the explanations. The two users could have also just not paid attention when filling in the post-study questionnaire.

Due to the discrepancy between the results obtained for multi-dimensional trust and one-dimensional trust, it is difficult to answer the first research question in a conclusive manner. We thus answer it as follows: *Based on the observations from the user study, the explanation interface can increase trusting beliefs as well as multi-dimensional trust (as defined in this work) for Belgian high school students in this context. However, one-dimensional trust is not significantly affected, and also shares a low correlation with perceived transparency. This may be an indication that the explanation is not the most important factor compared to other factors such as perceived recommendation accuracy or the quality of the exercises. On the other hand, it may also show the more balanced, extra perspective multi-dimensional trust offers compared to one-dimensional trust.*

6.2 Effect of Placebo Explanations

The second research question aimed to gain insight into placebo explanations and their influence on initial user trust. Contrary to the results of Eiband et al. [25], we see little to no difference between the initial user trust for the group that received the IPE and the group that received the INE. In fact, the median integrity score is lower compared to that of the INE group, although the one-sided Mann-Whitney U test indicates no significant differences between the distributions. Participants that saw through the placebo explanation may have felt that they were cheated by the platform, as can be inferred from how heavily impacted Q10 (“Wiski is honest”) is. The discrepancy with Eiband et al.’s results can be due to a wide array of factors. The most obvious explanation is that both their research as well as this user study have a very low sample size. Their study similarly consisted of 30 participants split evenly amongst three groups. Furthermore, there are quite significant differences between the contexts of the research and the questions asked. Eiband et al.’s research, for instance, do not explicitly take into account constructs such as competence, benevolence, and integrity. Our results thus give a different perspective concerning placebo explanations. The authors see the possibility of using placebo explanations as a placeholder when not enough information is present. However, our research shows that using placebo explanations in this manner may give the users the impression that the platform is less competent (and maybe less integrous).

Perceived transparency for the IPE is somewhat controversial: half of the participants in this group at least somewhat agree that the platform gives enough explanations. Therefore, some participants are somewhat satisfied with the placebo

explanation, while others are dissatisfied, as also seen by the written responses in the previous chapter. This may be explained by, for example, the fact that some participants did not pay much attention to the explanation interface or to the post-study questionnaire. Another explanation, as alluded to in the previous section, may be that some participants may not require much or any transparency. It is interesting to observe a discrepancy between the participants' views on placebo explanations. These observations show that extra information (e.g., "Does this user really need transparency?") can be collected compared to using a no-explanation baseline.

Based on the observations from the user study, we can answer the second research question in the following manner: *Quantitatively, there are little to no differences between using placebo explanations and providing no explanations on initial user trust. However, using placebo explanations in user studies may provide an extra dimension of information compared to only using a no-explanation baseline.*

6.3 Recommendation Clicks

We see little to no correlation between the acceptance of the recommended exercises and the initial trust level. Fig. 5.5 shows that the first recommended exercise is clicked the most. Therefore, most users may just accept recommendations due to the convenience rather than the trust.
We do, however, observe that users from the INE group tend to accept a recommendation less than those from the IRE and IPE group (even though it is mostly one less time out of five). These results are similar to those of Cramer et al. [21] and may be explained as follows: users do not know how the recommendations are provided, and thus may turn to the labels on the exercise selection page.
We can combine this observation and the fact that users mostly click on the first recommended exercise to infer the following insight: when users know (or think they know) why an exercise is recommended (in this case, difficulty level), they most likely select the first recommended exercise due to its convenience. However, when the users are left in the dark as to how or why an exercise is recommended, they may feel more comfortable selecting an exercise themselves. This way (when using Wiski), the users at least know they are selecting an exercise according to its difficulty level.

These observations have interesting implications. For example, giving multiple recommendation options may not be necessary for a context similar to that of Wiski, due to how often the first recommended exercise is chosen. Recall that the second and third recommended exercise are rarely chosen. Note that the possibility exists that the majority of users did not see the second and third recommended exercise. However, the final think-aloud study showed that all four users knew that there were three recommended exercises, and thus we can most likely conclude that this is not the reason.
The results further support our argument concerning the relative importance of explanations in this context: users may be more interested in solving as many exercises

as possible rather than making informed decisions about which exercise to solve next, if they know that the algorithm recommends exercises according to difficulty level. It also shows that the first recommended exercise (out of multiple recommendations) should be the best or most important exercise for the given user, as this is, most likely, the exercise the user clicks on the most.

It is nonetheless important to be cautious about the generalizability of these implications. Using a different type of explanation (interface) may give different results. Furthermore, we must take into account that there is little to no risk involved when accepting a recommendation in this context. Participants are conscious of the fact that this is an experiment of short duration. Therefore, when there are little to no repercussions for accepting a bad recommendation, it may be easier for participants to opt for convenience in these situations.

Based on the obtained results, we answer the third research question in the following manner. *There is no correlation observed between acceptance of recommendations measured through click-through rate and initial user trust in this thesis.*

6.4 Limitations

The research conducted is not void of limitations. This section aims to outline these limitations and how they may have influenced our results.

1. Unfortunately, only a small number of participants were able to be recruited for the final user study. With only 37 participants in total, the results should be interpreted cautiously. However, we do present valuable data points for users in this specific age group. Our results can thus be used as starting points for future research.

2. There are a few limitations that should be taken into account when it comes to the platform Wiski.
   First, the algorithm used in the thesis is fairly basic compared to those that are state-of-the-art. For example, we use the Elo rating system in its simplest form, and there is only one rating per user. Thus if this platform is used in a long-term study, a better solution would consist of using a multi-dimensional Elo rating system (e.g., Elo ratings per subject or section), similar to Abdi et al. [2]. The algorithm's accuracy has also not been measured, and no parameter tuning has been conducted to optimize it. However, with the available (or rather the lack of) prior data, making an improvement here would have been quite challenging.
   Furthermore, the exercises on the platform are quite basic. Some participants have communicated this when filling in the post-study questionnaire. If solving a math exercise takes an insignificant amount of time, the importance of picking a good exercise becomes smaller. This is another outside factor that could have influenced the results. The research should thus be conducted with harder

math exercises to investigate whether the results stay the same.
Next, to guarantee that all participants that receive the real explanation see the explanation interface with the explanation and histogram, fake attempts had to be added to the data. It is important to note that the end-users did not know that fake attempts were added and displayed in the histogram. Nonetheless, this could have impacted the accuracy of the recommendations, which in turn may have influenced the users' trust levels. However, we believe that not doing so and risking users not seeing the explanation interface would have influenced the results even more.
Finally, the platform itself does not explain why a certain answer-option is incorrect, let alone does it show a step-by-step solution for any exercise. Various participants have expressed how this is a feature that is missed in the platform, and it may have therefore influenced the results. As this feature was not present for all three groups, its effect should also be uniformly present across the groups. However, if this is not the case (e.g., the IRE group missing it less due to receiving explanations in another form), the obtained results become less reliable.

3. Due to the cold-start problem of the algorithm, a trade-off had to be made between risking recommending too few exercises or recommending all of the exercises for a particular section. We opted for the latter, and therefore, users can be recommended exercises that lie outside of their Elo range. A solution to this problem can be to show the remaining exercises with a disclaimer that it may be too difficult or too easy.
Furthermore, due to the lack of data in the beginning, participants towards the end of the study may have received "more accurate" recommendations, making the comparisons between the first and last participants possibly biased.

4. The logging of the interactions with the explanation interface is not perfect. For example, the logging feature does not work well depending on the browser used. Participants were notified at various steps before participation and registration to use the appropriate browsers. However, as information regarding the participants' used browser was not stored, we cannot be 100% confident in the logged interactions with the explanation interface.

5. Although the post-study questionnaire questions for trusting beliefs are based on the ones used by Wang and Benbasat [9], modifications had to be made to use them. Future work can consist of validating the used questionnaire.

Chapter 7

Conclusion

This thesis tackled the complex topic of initial user trust for explainable recommender systems. Trust has been classified as an essential goal in the domain of HCI due to the vast array of benefits it can provide. This research investigated the effects of accompanying recommendations with an explanation interface in an e-learning platform for high school students.
We first augmented an e-learning platform, Wiski [60], with a recommender system that uses a combination of the Elo rating system and collaborative filtering [23]. An accompanying explanation interface was designed following the user-centered design principle.
A user study (randomized controlled experiment) consisting of 37 high school students from Flanders was then conducted to research the effects of the explanation interface on initial user trust. We further investigated the influence of placebo explanations and whether a correlation exists between initial user trust and the acceptance of recommendations.

The results show that the explanation interface is successful in increasing certain aspects of initial user trust. The interface mainly increases perceived competence for the end-user. However, we did not find significant evidence that the interface had a substantial effect on increasing initial user trust when asked explicitly. This result leads us to either believe that other factors such as perceived accuracy or the website's appearance may have made up for the difference in perceived competence, or that one question cannot capture the multi-faceted nature of trust.
We further found that placebo explanations did not offer any significant differences quantitatively compared to using no explanations. However, the divisive nature of the qualitative responses gives the belief that placebo explanations can be used to inquire extra information from a user study.
Finally, we found no correlation between acceptance of recommendations and initial user trust. However, we observed that users who received no explanations tended to accept a recommendation less than the users from the other two research groups.
One significant limitation of the study is the small sample size, and thus, the results should be interpreted cautiously. Nonetheless, using explanations in an e-learning

platform seems to be an asset for high school students. Our results show that many users of this target audience wishes to receive explanations when using Wiski or a recommender system in general. Accompanying recommendations with explanations should, therefore, definitely be considered when implementing a similar application.

Future Work

Our research opens up the opportunity for various future work.

- One obvious direction of future work is to study the effects of using different explanation interfaces and explanations. Our results point towards a hypothesis that different users may desire different forms or levels of transparency and explanations. It may be interesting to give users more control by allowing them to select the type or level of explanation themselves, or take personal characteristics into account, similar to Millecamp et al.'s work [52]. The use of Open Learner Models is also an interesting direction for future work with Wiski, especially as the platform partially uses the Elo rating system to make recommendations. Therefore, future work can use a combination of textual explanations as well as an Open Learner Model (similarly to the work in progress by Barria-Pineda [8]).

- In our research, we showed that placebo explanations could offer an extra dimension of information. According to our knowledge, only Eiband et al.'s [25] work uses placebo explanations as an extra alternative to a no-explanation baseline. Further work can thus look into the effects of placebo explanations on a much broader scale and scope in on their applications in user studies.

- The thesis limited us to a short-term study, whereas trust is an entity that evolves [59, 37]. Our results showed that the intention to return amongst the three groups was nearly identical. However, it may be interesting to perceive how trust evolves in a long-term study, and whether users from a particular research group, in reality, come back to use the platform more often than another group. Such a long-term study also opens up the possibility of measuring trust implicitly using loyalty [70, 49].

Appendix A

User Study Recruitment Documents

A.1 Recruitment Documents

This part of the appendix consists of the various documents used to recruit participants for the final user study. The first document shows the information brochure given to the teachers of the participating students. A second, similar document was sent to the parents of the participating students. Finally, participants were asked to fill out an informed consent form, which is the last document in this part of the appendix.

Informatiebrochure Wiski

Beste leerkracht,

Bedankt voor uw interesse in het wetenschappelijke onderzoek voor mijn masterthesis. Hier zal u lezen hoe het onderzoek precies zal verlopen.

Doel van het onderzoek
Elk leerling is anders. Een oefening die voor sommige leerlingen makkelijk is, kan moeilijk zijn voor andere leerlingen en omgekeerd. 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. Wat er speciaal is aan Wiski is dat het oefeningen probeert aan te raden die bij de leerling zijn/haar niveau passen. De leerling kan dus kiezen om een oefening zelf te kiezen, of een aangeraden oefening te maken. Ik zou met dit onderzoek te weten willen komen of de leerling de website vertrouwt.

Hoe verloopt het onderzoek?
Toestemmingsformulier:
Elke leerling moet een *toestemmingsformulier* invullen voor het deelnemen van de studie.
(zie: https://www.kuleuven.be/english/research/ethics/committees/smec/faq)
Het is hier van uiterst belang de leerling in kwestie niet onder druk gezet wordt om deel te nemen aan de studie. Indien de leerling niet wilt meedoen, zal de leerkracht een gelijkwaardig alternatief voorzien. Dit kunnen bijvoorbeeld oefeningen op papier zijn. Verder kunnen leerlingen op elk moment en zonder reden hun deelneming stopzetten, zonder enig nadeel te ondervinden.
Voor leerlingen die 15 jaar of jonger zijn, moet het formulier ingevuld worden door zijn/haar ouders. De leerling zelf moet het formulier ook ondertekenen. Leerlingen van 16 jaar en ouder mogen zelf het formulier invullen. Hierbij hebben de ouders wel het recht om geïnformeerd te worden over de studie. Leerkrachten kunnen bijvoorbeeld de informatiebrochure doorsturen naar de ouders via Smartschool.

Gebruik van Wiski:
Elke leerling moet zich eerst registreren voor de website. Daarbij moet de leerling een kort vragenlijstje invullen. Na de registratie kan de leerling zoveel oefeningen maken als hij/zij wenst. Al hun digitale activiteiten omtrent het kiezen en oplossen van oefeningen worden op de achtergrond bijgehouden en later door mij geanalyseerd. Wanner de leerling een zesde oefening wilt maken, moet hij/zij terug een korte vragenlijst moeten invullen. Het zou dus fijn zijn moesten de leerlingen **minstens 6 oefeningen kunnen maken op Wiski als huiswerk of in de les**. ***Hun antwoorden op deze vragenlijst zijn heel belangrijk voor mijn onderzoek.***

Wat gebeurt er met de gegevens?
De gegevens worden veilig bewaard op de servers van KU Leuven en zijn alleen toegankelijk voor mij. Bij de registratie zal de leerling zijn/haar e-mailadres moeten ingeven. Deze wordt alleen gebruikt door mij om hem/haar te contacteren op het einde van de onderzoek, en zal dus ook verwijderd worden hierna. Er worden verder geen persoonlijke gegevens gevraagd die gebruikt kunnen worden om de leerling direct te identificeren.

Hebt u nog vragen?
Indien er onduidelijkheden zijn of indien u nog vragen/feedback/opmerkingen hebt, kan u steeds contact opnemen met mij op
shotallo.kato@student.kuleuven.be

Informatiebrochure Wiski

Beste ouders,

Ik ben Shotallo en voor mijn masterthesis voer ik een studie uit in middelbare scholen. De school van uw kind werkt mee aan mijn onderzoek en uw kind kan kiezen om hieraan mee te doen. Hier zal u lezen hoe het onderzoek precies zal verlopen.

Doel van het onderzoek
Elk leerling is anders. Een oefening die voor sommige leerlingen makkelijk is, kan moeilijk zijn voor andere leerlingen en omgekeerd. 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. Wat er speciaal is aan Wiski is dat het oefeningen probeert aan te raden die bij de leerling zijn/haar niveau passen. De leerling kan dus kiezen om een oefening zelf te kiezen, of een aangeraden oefening te maken. Ik zou met dit onderzoek te weten willen komen of de leerling de website vertrouwt.

Hoe verloopt het onderzoek?
Toestemmingsformulier:
Elke leerling moet een *toestemmingsformulier* invullen voor het deelnemen van de studie.
(zie: https://www.kuleuven.be/english/research/ethics/committees/smec/faq)
Het is hier van uiterst belang de leerling in kwestie niet onder druk gezet wordt om deel te nemen aan de studie. Indien de leerling niet wilt meedoen, zal de leerkracht een gelijkwaardig alternatief voorzien. Dit kunnen bijvoorbeeld oefeningen op papier zijn. Verder kunnen leerlingen op elk moment en zonder reden hun deelneming stopzetten, zonder nadeel te ondervinden.
Voor leerlingen die 15 jaar of jonger zijn, moet het formulier ingevuld worden door zijn/haar ouders. De leerling zelf moet het formulier ook ondertekenen.
Leerlingen van 16 jaar en ouder mogen zelf het formulier invullen.

Gebruik van Wiski:
Elke leerling moet zich eerst registreren voor de website. Daarbij moet de leerling een kort vragenlijstje invullen. Na de registratie kan de leerling zoveel oefeningen maken als hij/zij wenst. Al hun digitale activiteiten omtrent het kiezen en oplossen van oefeningen worden op de achtergrond bijgehouden en later door mij geanalyseerd. Wanner de leerling een **zesde oefening** wilt maken, moet hij/zij terug een korte vragenlijst moeten invullen. ***Hun antwoorden op deze vragenlijst zijn heel belangrijk voor mijn onderzoek.***

Wat gebeurt er met de gegevens?
De gegevens worden veilig bewaard en zijn alleen toegankelijk voor mij. Bij de registratie zal de 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 leerling direct te identificeren.

Hebt u nog vragen?
Indien er onduidelijkheden zijn of indien u nog vragen/feedback/opmerkingen hebt, kan u steeds contact opnemen met mij op
shotallo.kato@student.kuleuven.be

Geïnformeerde toestemming

Titel van het onderzoek:
Practicing the Right Math: Automatically Adapting Digital Learning Environments to Learners

Naam + contactgegevens promotor en onderzoeker(s):
Kato Shotallo shotallo.kato@student.kuleuven.be
Ooge Jeroen jeroen.ooge@kuleuven.be
Verbert Katrien katrien.verbert@kuleuven.be

Doel en methodologie van het onderzoek:
Het doel van het onderzoek is om het niveau van vertrouwen te meten in een e-learning platform "Wiski". Wiski is een website die probeert om oefeningen van het correcte niveau aan te raden aan de gebruiker. De gebruiker zal eerst moeten registreren voor de website en zal daarbij een paar vragen beantwoorden (bijvoorbeeld leeftijd, school, of hij/zij sterk is in wiskunde, …). Daarna zal hij/zij vrij oefeningen maken op Wiski. Nadat de gebruiker gedaan heeft met oefeningen maken, zal hij/zij een korte vragenlijst invullen over zijn/haar ervaringen op Wiski.

Duur van het experiment:
Het experiment zal worden uitgevoerd in verschillende scholen. De student zelf zal ongeveer een uur besteden aan het experiment.

- Ik begrijp wat van mij verwacht wordt tijdens dit onderzoek.

- Ik weet dat ik zal deelnemen aan volgende proeven of testen:
  Beantwoorden van vragenlijsten.
  Maken van oefeningen op Wiski.

- Ik weet dat er risico's of ongemakken kunnen verbonden zijn aan mijn deelname:
  Niet van toepassing.

- 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.

  <u>Alternatief 1</u> (wanneer de studie in opdracht van overheden is of resultaten publiek zullen gemaakt worden)
  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, anonimiteit en de vertrouwelijkheid van de gegevens is 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:
  shotallo.kato@student.kuleuven.be

Opgemaakt in tweevoud.

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 data protection officer: dpo@kuleuven.be

- Voor eventuele klachten of andere bezorgdheden omtrent ethische aspecten van deze studie kan ik contact opnemen met de Sociaal-Maatschappelijke Ethische Commissie van KU Leuven: smec@kuleuven.be

- Ik weet dat ik bij onderstaande terecht kan indien ik na het onderzoek ongemakken of moeilijkheden ervaar als gevolg van de thema's die in het onderzoek aan bod kwamen:
  Shotallo Kato shotallo.kato@student.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.

Belangrijk: als je 15 jaar of jonger bent, dan moeten je ouders dit document verplicht ondertekenen.

Datum:
Naam en handtekening proefpersoon

Naam en handtekening onderzoeker
Kato Shotallo

Naam en handtekening(en) ouder(s) **(verplicht voor leerlingen 15 jaar en jonger)**

Opgemaakt in tweevoud.

Appendix B

Questionnaires

This part of the appendix presents the questionnaires (both the pre-study and post-study questionnaire) used in the final user study. Both the Dutch questions and English questions are provided. A further table showing the modifications made to Wang and Benbasat's original trusting beliefs questionnaire [9] can be seen at the end of the section.

B.1 Pre-Study Questionnaire

Table B.1: The questions asked in the pre-study questionnaire. The possible answer choices can be seen in the second column.

Pre-Study Questionnaire Questions	
Wat is je leeftijd? (What is your age?)	<13/14/15/16/17/18/18<
Wat is je geslacht? (What is your gender?)	Boy/Girl/Other
In welk leerjaar zit je? (In what grade are you?)	3/4/5/6/Other
Hoeveel uur wiskunde krijg je per week? (How many hours of math do you receive a week?)	<2/2/3/4/5/6/7/8/8<
Hoe goed ben je, volgens jezelf, in wiskunde? (How good are you in math according to yourself?)	Very Weak/Weak/Average/Good/Very Good

Hoe goed ben je, volgens jezelf, met het gebruik van computers? (How good are you at using computers according to yourself?)	Very Weak/Weak/Average/Good/Very Good
Gebruikte je al eens een andere website om online oefeningen te maken? (bijvoorbeeld voor wiskunde, Nederlands, ...) (Have you ever used another website to solve exercises? (For example, for math, Dutch, ...))	Yes/No
Gebruik je websites waar producten/films/... worden aangeraden? (zoals Netflix, Amazon, Bol, ...) (Do you use websites where products/movies/... are recommended? (such as Netflix, Amazon, Bol, ...))	Yes/No

B.2 Post-Study Questionnaire (Dutch)

Table B.2: The original questions used in the post-study questionnaire (in Dutch). Appendix B, Table B.3 shows the questionnaire translated to English. All questions were evaluated on a 7-point Likert scale.

Post-Study Questionnaire	
Competence	
Q1.	Wiski is zoals een expert (bv. een leerkracht) in wiskunde-oefeningen aanraden.
Q2.	Wiski heeft de expertise (kennis) om mijn wiskundeniveau te kunnen inschatten.
Q3.	Wiski kan mijn wiskundeniveau inschatten.
Q4.	Wiski begrijpt de moeilijkheidsgraad van wiskunde-oefenigen goed.
Q5.	Wiski houdt rekening met mijn wiskundeniveau om oefeningen aan te raden.
Benevolence	
Q6.	Wiski zet op de eerste plaats dat ik vorderingen maak in wiskunde.
Q7.	Wanneer Wiski oefeningen aanraadt, doet Wiski dat zodat ik vorderingen maak in wiskunde.
Q8.	Wiski wilt mijn wiskundeniveau goed inschatten.
Integrity	
Q9.	Wiski raadt oefeningen op een zo correct mogelijke manier aan.
Q10.	Wiski is eerlijk.
Q11.	Wiski maakt oprechte aanbevelingen.
Trust (One-dimensional)	
Q12.	Ik vertrouw Wiski om mij wiskunde-oefeningen aan te raden.
Intention to Return	
Q13.	Als ik nog eens online wiskunde-oefeningen maak, dan kies ik voor Wiski.
Q14.	Als ik nog eens wiskunde-oefeningen aangeraden wil krijgen, dan kies ik voor Wiski.

Transparency	
Q15.	Ik vind dat Wiski genoeg uitleg geeft over waarom een oefening aangeraden is.
General Questions	
Q16.	Wanneer ik Wiski gebruik, wil ik GEEN uitleg over waarom een oefening wordt aangeraden.
Q17.	Ik vind uitleg krijgen over waarom een oefening wordt aangeraden belangrijker dan waarom een film wordt aangeraden.
Q18	Ik ben NIET blij met het niveau van de oefeningen die Wiski aanraadde.
Q19.	In het algemeen vind ik het belangrijk om uitleg te krijgen wanneer iets (oefening/film/product/...) wordt aangeraden.

B.3 Post-Study Questionnaire (English)

Table B.3: The questionnaire participants answered at the end of the user-study, translated to English. All questions were evaluated on a 7-point Likert scale.

Post-Study Questionnaire	
Competence	
Q1.	Wiski is like an expert (for example, a teacher) for recommending math exercises.
Q2.	Wiski has the expertise (knowledge) to estimate my math level.
Q3.	Wiski can estimate my math level.
Q4.	Wiski understands the difficulty level of math exercises well.
Q5.	Wiski takes my math level into account when recommending exercises.
Benevolence	
Q6.	Wiski prioritizes that I improve in math.
Q7.	Wiski recommends exercises so that I improve in math.
Q8.	Wiski wants to estimate my math level well.
Integrity	
Q9.	Wiski recommends exercises as correctly as possible.
Q10.	Wiski is honest.
Q11.	Wiski makes integrous recommendations.
Trust (One-dimensional)	
Q12.	I trust Wiski to recommend me math exercises.
Intention to Return	
Q13.	If I want to solve math exercises again, I will choose Wiski.
Q14.	If I want to be recommended math exercises again, I will choose Wiski.
Transparency	
Q15.	I find that Wiski gives enough explanation as to why an exercise has been recommended.
General Questions	
Q16.	I do NOT want any explanations about why an exercise has been recommended when I use Wiski.

Q17.	I find receiving an explanation about why an exercise has been recommended more important than an explanation for why a movie has been recommended.
Q18.	I am NOT happy with the level of math exercises Wiski recommended.
Q19.	I find it important to receive explanations when something (exercise/movie/product/...) has been recommended.

B.4 Modifications to Trusting Beliefs Questionnaire

Table B.4: The original questions used by Wang and Benbasat [9] to measure trusting beliefs, along with remarks concerning what was modified to obtain the questions used in the post-study questionnaire (Table B.2 (Dutch), Table B.3 (English)).

Post-Study Questionnaire	
Competence	
Original Question.	The virtual advisor is like a real expert in assessing digital cameras
Remarks Q1.	An example "a teacher" was added to the question as user studies showed that it was not clear what an expert was to the participants.
Original Question.	The virtual advisor has the expertise to understand my needs and preferences about digital cameras.
Remarks Q2.	The word knowledge has been added to the question in parentheses for clarification purposes. "Needs and preferences about digital cameras" has been modified to "my ability (literally: level) in mathematics".
Original Question.	This virtual advisor has the ability to understand my needs and preferences about digital cameras.
Remarks Q3.	The word ability has been translated to "can" for easier understanding. "Needs and preferences about digital cameras" has been modified to "my ability (literally: level) in mathematics".
Original Question.	This virtual advisor has good knowledge about digital cameras.
Remarks Q4.	"Good knowledge about digital cameras" has been changed to "understands the difficulty level of math exercises" to fit the context better.
Original Question.	This virtual advisor considers my needs and all important attributes of digital cameras.
Remarks Q5.	"Considers" is expanded to "takes into consideration to recommend math exercises" for clarification purposes. "My needs and all important attributes of digital cameras" has been modified to "my ability (literally: level) in mathematics".
Benevolence	
Original Question.	The virtual advisor puts my interests first.
Remarks Q6.	"My interests" has been modified to "progress in mathematics".
Original Question.	The virtual advisor keeps my interests in mind.

Remarks Q7.	"My interests" has been modified to "progress in mathematics". "Keeps in mind" has been expanded to "when recommending exercises" for clarification purposes.
Original Question.	The virtual advisor wants to understand my needs and preferences.
Remarks Q8.	"Needs and preferences" has been translated to "ability (literally: level) in mathematics" to fit the context. "Understand" has been modified to estimate.
Integrity	
Original Question.	The virtual advisor provides unbiased product recommendations.
Remarks Q9.	"Unbiased" has been translated to "as correct as possible" to fit the participants' vocabulary.
Original Question.	The virtual advisor is honest.
Remarks Q10.	\
Original Question.	I consider this virtual advisor to possess integrity.
Remarks Q11.	The decision was made to scope in on the recommendations for this question as we wanted the users to focus on the recommender aspect of Wiski (and not, for example, whether Wiski gives incorrect answers to questions).

Appendix C

Think-Aloud Study Information

This part of the appendix shows the tasks and their respective goals used in the think-aloud studies. The feedback matrices consisting of the most important feedback acquired from the think-aloud studies can be seen afterward.

C.1 Tasks and Goals

Table C.1: The questions asked during the think-aloud studies. Questions indicated with *HF* were also asked during the think-aloud study during the high-fidelity prototype stage.

Tasks and Goals for Think-Aloud Studies
Solve an Exercise on Wiski
Task 1. *HF* Maak voor mij oefening 5 van het thema natuurlijke getallen van het hoofdstuk hoofdbewerkingen. **Translation.** Solve exercise 5 from the subject natural numbers in the section basic operations. **Goal:** Observe whether the participant can navigate through the main flow of the website without any issues.
Task 2. *HF* Kan je mij zeggen wat je op de "voltooid" pagina ziet? **Translation.** Can you tell me what you see on the "completed" page? **Goal:** Observe what elements of the page the participant picks up on and check whether they interpret them correctly.
Task 3. *HF* Leg de "waarom" uitleg uit. **Translation.** Explain the why explanation. **Goal:** Observe whether the participant understands the why explanation.

Task 4. *HF* Leg de interpretatie van de grafiek (histogram) uit.
Translation. Explain the interpretation of the histogram.
Goal: Observe whether the participant understands the visual explanation.

Question. *HF* Hoeveel oefeningen zijn er aangeraden?
Translation. How many exercises are recommended?
Question. *HF* Wat zou je klikken om de andere aangeraden oefeningen te zien?
Translation. What would you click to see the other recommended exercises?
Question. *HF* Welke van de 2 aangeraden oefeningen is waarschijnlijk moeilijker?
Translation. Which of the 2 recommended exercises do you expect to be harder?
Question. Wat zou je klikken als je een moeilijke oefening wilt krijgen?
Translation. What would you click if you want a difficult exercise?
Goal: Further examine how well the participant understands the explanation interface.

Task 5. *HF* Stel je vindt de aangeraden oefeningen niet leuk. Wat zou je doen?
Translation. What would you do if you do not like the recommended exercises?
Goal: Observe whether the participant understand the alternative to selecting a recommended exercise.

Transparency Page + Histogram Tutorial

Task 1. Lees de volgende pagina's aandachtig en probeer met eigen woorden uit te leggen wat je begrepen hebt.
Translation. Carefully read the following pages and try to explain in your own words what you understood.
Goal: Observe whether the participant understands the transparency pages + histogram.

Question. Waarom is Emile geen gelijkaardige student?
Translation. Why is Emile not a similar student?
Question. Wat is de interpretatie van de tweede bar op de grafiek?
Translation. What is the interpretation of the second bar in the histogram?
Question. Waarom worden er oefeningen die 100% slaagkansen hebben niet aangeraden?
Translation. Why are exercises with 100% passing rate not recommended?
Goal: Further examine whether the participants understand the transparency pages + histogram.

Exercise List

Question. *HF* Hoe zie je of een oefening gemaakt is?

Translation. How do you see that an exercise has already been solved?
Question. *HF* Kan je de oefeningen sorteren op gemaakt, moeilijkheidsgraad en oefeningnummer?
Translation. Can you sort the exercises by solved, difficulty level, and exercise number?
Goal: Observe whether the participant understands the features on this page.

Feedback Questions

Question. *HF* Wat vind je van pagina ... ?
Translation. What do you think of the ... page?
Goal: Obtain feedback of the various pages.

Question. *HF* Snap je de waarom uitleg? Is het nuttig?
Translation. Do you understand the why explanation? Is it useful?
Goal: Gain insight into what the participant thinks of the explanation.

Question. Vind je de "tutorial" nuttig? Waarom wel/niet? Zou je het lezen?
Translation. Do you find the tutorial useful? Why (not)? Would you read it?
Goal: Gain insight into what the participant thinks of the transparency pages.

Question. Hoe in detail wil je weten wat er achter de schermen gebeurt? (voor de aanbevelingen)
Translation. How much in detail would you want to know what is happening behind the recommendations?
Goal: Gain insight into how much transparency the participant values.

C.2 Feedback Matrices

Table C.2: The main problems for the first and second think-aloud studies, along with their frequencies and possible changes.

	Frequency Think-Aloud Study 1	Change	Frequency Think-Aloud Study 2	Change
Transparency Pages				
Did not understand transparency page completely.	3/5	Use more explicit wording and only show page at explanation interface.	5/7	Remove transparency page.
Explanation Interface				
Interprets histogram incorrectly.	4/5	Make title more explicit instead of just "grafiek".	4/7	Only middle schoolers answered incorrectly.
Number of recommended exercises not clear.	2/5	Add title stating that following exercises are recommended.	4/7	Look for more intuitive toggle methods.
Difficulty browsing through recommended exercises.	2/5	Use explicit wording by the arrows.	2/7	Look for more intuitive toggle methods.
Interpreted explanation incorrectly.	1/5	Wait for results of second think-aloud study.	4/7	Only middle schoolers answered incorrectly.
How do users experience the flow?				
Difficulty navigating to exercise.	/	/	3/7	Display the sections immediately.

Table C.3: The main problems stemming from the think-aloud study during the high-fidelity prototype stage.

	Frequency	Changes
Explanation Interface		
Not clear what section recommended exercise is from.	3/4	Add that recommended exercise comes from same section.
What data is used to calculate the estimated number of tries.	1/4	Add that it is based on the user's data and that of their fellow students.
Colors make me think that last problem is a challenging problem.	1/4	Change color of recommended exercises to single color.
Exercise List		
Not clear that difficulty level is personal.	3/4	Write "Expected difficulty level for you".

Appendix D

Results

This part of the appendix shows the results that were not explicitly necessary in in the main part of the thesis.

D.1 IPE vs. INE

Table D.1: Results of the two-sided Mann-Whitney U test for the group with the IPE vs. the group with the INE.

	p-value	U value	Common Language Effect Size
Competence	0.978	78.0	0.5
Benevolence	0.978	78.0	0.5
Integrity	0.143	51.0	0.327
Trusting Beliefs	0.703	70.5	0.452
Intention to Return	0.696	85.5	0.548
Perceived Transparency	0.099	108.0	0.692
One-Dimensional Trust	0.728	71.5	0.458
Multi-Dimensional Trust	0.978	78.0	0.5

$^{**}p < 0.01$, $^{*}p < 0.05$

Bibliography

[1] H. Abdi. Kendall Rank Correlation Coefficient. *The Concise Encyclopedia of Statistics*, pages 278–281, 2008.

[2] S. Abdi, H. Khosravi, S. Sadiq, and D. Gasevic. A multivariate elo-based learner model for adaptive educational systems. *arXiv*, (1), 2019.

[3] S. Abdi, H. Khosravi, S. Sadiq, and D. Gasevic. Complementing educational recommender systems with open learner models. *ACM International Conference Proceeding Series*, pages 360–365, 2020.

[4] A. Adadi and M. Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), 2018.

[5] C. C. Aggarwal. *Recommender Systems The Textbook*, volume 39. 2016.

[6] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):1–9, 2020.

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

[8] J. Barria-Pineda. Exploring the Need for Transparency in Educational Recommender Systems. *UMAP 2020 - Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 376–379, 2020.

[9] I. Benbasat and W. Wang. Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems*, 6(3):72–101, 2005.

[10] S. Berkovsky, R. Taib, and D. Conway. How to Recommend? User Trust Factors in Movie Recommender Systems. pages 287–300, 2017.

[11] M. Brinkhuis. *Tracking Educational Progress*. PhD thesis, Psychology Research Institute (PsyRes), Amsterdam, 2014.

[12] Brinkhuis, Matthieu, Cordes, Wessel, and Hofman, Abe. Governing games Adaptive game selection in the Math Garden. *ITM Web Conf.*, 33:3003, 2020.

[13] A. Brun, G. Bonnin, S. Castagnos, A. Roussanaly, and A. Boyer. Learning analytics made in France: The METALproject. *arXiv*, pages 1–16, 2019.

[14] S. Bull and J. Kay. Categorisation and Educational Benefits of Open Learner Models. 17(2008):2009, 2009.

[15] C. Burns and S. M. Conchie. Measuring implicit trust and automatic attitude activation. *Handbook of Research Methods on Trust: Second Edition*, (November 2011):292–301, 2015.

[16] A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, (October):160–169, 2015.

[17] C. M. Chen, H. M. Lee, and Y. H. Chen. Personalized e-learning system using Item Response Theory. *Computers and Education*, 44(3):237–255, 2005.

[18] L. CHEN. User Decision Improvement and Trust Building in Product Recommender Systems. (August 2008):257, 2008.

[19] B. C. Choi and A. W. Pak. A catalog of biases in questionnaires. *Preventing Chronic Disease*, 2(1):1–13, 2005.

[20] K. Chopra and W. A. Wallace. Trust in electronic environments. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, HICSS 2003*, 2003.

[21] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga. *The effects of transparency on trust in and acceptance of a content-based art recommender*, volume 18. 2008.

[22] J. Daher, A. Brun, and A. Boyer. A Review on Explanations in Recommender Systems. pages 12–16, 2018.

[23] O. H. Dahl and O. Fykse. *Combining Elo Rating and Collaborative Filtering to improve Learner Ability Estimation in an e-learning Context.* PhD thesis, Norwegian University of Science and Technology, 2018.

[24] O. Dan and B. D. Davison. Measuring and predicting search engine users' satisfaction. *ACM Computing Surveys*, 49(1), 2016.

[25] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann. The impact of placebic explanations on trust in intelligent systems. *Conference on Human Factors in Computing Systems - Proceedings*, 2019.

[26] A. E. Elo. *The rating of chessplayers, past and present.* Arco Pub., 1978.

[27] J. Ermisch, D. Gambetta, H. Laurie, T. Siedler, and S. C. N. Uhrig. Measuring People's Trust. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 172(4):749–769, feb 2009.

[28] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve Web search. *ACM Transactions on Information Systems*, 23(2):147–168, 2005.

[29] F. Gedikli, D. Jannach, and M. Ge. How should i explain? A comparison of different explanation types for recommender systems. *International Journal of Human Computer Studies*, 72(4):367–382, 2014.

[30] I. Ghergulescu and C. H. Muntean. *Measurement and Analysis of Learner's Motivation in Game-Based E-Learning*, pages 355–378. Springer New York, New York, NY, 2012.

[31] M. Gorgoglione, U. Panniello, and A. Tuzhilin. In CARS We Trust : How Context-Aware Recommendations Affect Customers ' Trust And Other Business Performance Measures Of Recommender Systems Michele Gorgoglione Umberto Panniello Alexander Tuzhilin. 2011.

[32] T. Grandison and M. Sloman. A survey of trust in internet applications. *IEEE Communications Surveys & Tutorials*, 3(4):2–16, 2000.

[33] D. Gunning and D. W. Aha. DARPA's explainable artificial intelligence program. *AI Magazine*, 40(2):44–58, 2019.

[34] J. N. R. Haldor Myre, Sondre Oldervoll. *Developing a web application for creating, solving, assessing and collecting data from interactive mathematical tasks in Python/Django, HTML5, Javascript, CSS and MySQL.* PhD thesis, NTNU, 2017.

[35] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 241–250, 2000.

[36] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for Explainable AI: Challenges and Prospects. *arXiv*, pages 1–50, 2018.

[37] D. Holliday, S. Wilson, and S. Stumpf. User trust in intelligent systems: A journey over time. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 07-10-Marc(164):164–168, 2016.

[38] N. Hug. Surprise: A python library for recommender systems. *Journal of Open Source Software*, 5(52):2174, 2020.

[39] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007.

[40] S. Klinkenberg, M. Straatemeier, and H. L. Van Der Maas. Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, 57(2):1813–1824, 2011.

[41] Y. Koren. Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Transactions on Knowledge Discovery from Data*, 4(1):1–24, 2010.

[42] Laerd Statistics. Assumptions of the Mann-Whitney U test.

[43] E. J. Langer, A. Blank, and B. Chanowitz. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology*, 36(6):635–642, 1978.

[44] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, Mar 11 2019.

[45] J. D. Lee and K. A. See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1):50–80, 2004.

[46] Z. C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):35–43, 2018.

[47] W. McKinney. {D}ata {S}tructures for {S}tatistical {C}omputing in {P}ython. In S. van der Walt and J. Millman, editors, *{P}roceedings of the 9th {P}ython in {S}cience {C}onference*, pages 56–61, 2010.

[48] D. H. McKnight, V. Choudhury, and C. Kacmar. Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3):334–359, 2002.

[49] S. M. MCNEE, S. K. LAM, J. A. KONSTAN, and J. RIEDL. Interfaces for eliciting new user preferences in recommender systems. In *Lecture notes in computer science*, pages 178–187, Berlin, 2003. Springer.

[50] S. M. Merritt, H. Heimbaugh, J. Lachapell, and D. Lee. I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, 55(3):520–534, 2013.

[51] P. Michlík and M. Bieliková. Exercises recommending for limited time learning. *Procedia Computer Science*, 1(2):2821–2828, 2010.

[52] M. Millecamp, K. Verbert, S. Naveed, and J. Ziegler. To explain or not to explain: The effects of personal characteristics when explaining feature-based recommendations in different domains. *CEUR Workshop Proceedings*, 2450:10–18, 2019.

[53] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[54] S. Mohseni, N. Zarei, and E. D. Ragan. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. 1(1), 2018.

[55] B. M. Muir. Trust between humans and machines. *International Journal of Man-Machine Studies*, 27:327–339, 1987.

[56] M. Naiseh, N. Jiang, J. Ma, and R. Ali. Explainable Recommendations in Intelligent Systems: Delivery Methods, Modalities and Risks. *Lecture Notes in Business Information Processing*, 385 LNBIP(March):212–228, 2020.

[57] G. Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5):625–632, 2010.

[58] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan. 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, 2019.

[59] M. Nourani, J. T. King, and E. D. Ragan. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. 2020.

[60] J. Ooge. Het personaliseren van motivationele strategieën en gamificationtechnieken mbv recommendersystemen. 2019.

[61] G. Palmer, A. Selwyn, and D. Zwillinger. *The "Trust V": Building and Measuring Trust in Autonomous Systems*, pages 55–77. Springer US, Boston, MA, 2016.

[62] R. Pelánek. Modeling Students' Memory for Application in Adaptive Educational Systems. *Proceedings of the 8th International Conference on Educational Data Mining*, pages 480–483, 2015.

[63] R. Pelánek. Applications of the Elo rating system in adaptive educational systems. *Computers and Education*, 98:169–179, 2016.

[64] P. Pu and L. Chen. Trust building with explanation interfaces. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 2006(January 2006):93–100, 2006.

[65] P. Pu and L. Chen. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542–556, 2007.

[66] P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. *RecSys'11 - Proceedings of the 5th ACM Conference on Recommender Systems*, pages 157–164, 2011.

[67] P. Pu, L. Chen, and R. Hu. Evaluating recommender systems from the user's perspective: Survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22(4-5):317–355, 2012.

[68] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme. Recommender system for predicting student performance. *Procedia Computer Science*, 1(2):2811–2819, 2010.

[69] N. Tintarev. Explainable AI is Not Yet Understandable AI. In *Research Challenges in Information Science*, chapter Abstracts, pages xv–xvi. 2020.

[70] N. Tintarev and J. Masthoff. Designing and Evaluating Explanations for Recommender Systems. In *Recommender Systems Handbook*, pages 479–510. 2011.

[71] N. Tintarev and J. Masthoff. Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization. *User Modeling and User-Adapted Interaction*, 22(4-5):399–439, 2012.

[72] W. van Melle. MYCIN: a knowledge-based consultation program for infectious disease diagnosis. *International Journal of Man-Machine Studies*, 10(3):313–322, 1978.

[73] G. Vidotto, D. Massidda, S. Noventa, and M. Vicentini. Trusting beliefs: A functional measurement study. *Psicologica*, 33(3):575–590, 2012.

[74] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[75] M. L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[76] K. Wauters, P. Desmet, and W. den Noortgate. Item Difficulty Estimation: an Auspicious Collaboration Between Data and Judgement. *Computers & Education*, 58:1183–1193, 2012.

[77] K. L. Wuensch. CL: The Common Language Effect Size Statistic, 2015.

[78] J. Zaslow. If TiVo Thinks You Are Gay, Here's How to Set It Straight, 2006.

[79] R. Zhao, I. Benbasat, and H. Cavusoglu. Do users always want to know more? Investigating the relationship between system transparency and user's trust in advise-giving systems. *Proceedings of the 27th European Conference on Information Systems*, pages 1–13, 2019.