Explaining Recommendations in E-Learning: Effects on Adolescents' Trust

Jeroen Ooge*
jeroen.ooge@kuleuven.be
KU Leuven
Leuven, Belgium

Shotallo Kato*
shotallo.kato@student.kuleuven.be
KU Leuven
Leuven, Belgium

Katrien Verbert
katrien.verbert@kuleuven.be
KU Leuven
Leuven, Belgium

ABSTRACT

In the scope of explainable artificial intelligence, explanation techniques are heavily studied to increase trust in recommender systems. However, studies on explaining recommendations typically target adults in e-commerce or media contexts; e-learning has received less research attention. To address these limits, we investigated how explanations affect adolescents' initial trust in an e-learning platform that recommends mathematics exercises with collaborative filtering. In a randomized controlled experiment with 37 adolescents, we compared real explanations with placebo and no explanations. Our results show that real explanations significantly increased initial trust when trust was measured as a multidimensional construct of competence, benevolence, integrity, intention to return, and perceived transparency. Yet, this result did not hold when trust was measured one-dimensionally. Furthermore, not all adolescents attached equal importance to explanations and trust scores were high overall. These findings underline the need to tailor explanations and suggest that dynamically learned factors may be more important than explanations for building initial trust. To conclude, we thus reflect upon the need for explanations and recommendations in e-learning in low-stakes and high-stakes situations.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → **E-learning**.

KEYWORDS

teenagers, education, interpretability, explainability, XAI

ACM Reference Format:
Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3490099.3511140

*These authors contributed equally.

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03…$15.00
https://doi.org/10.1145/3490099.3511140

1 INTRODUCTION

People are increasingly relying on recommender systems that suggest relevant items, for example movies and music, tailored to their needs and interests. However, people are often left in the dark when it comes to why something has been recommended. In the scope of *explainable artificial intelligence* (XAI), many researchers agree that accompanying recommendations with explanations is often desirable because it can, for example, increase appropriate trust in the recommender [4, 53, 66], which in turn can increase people's willingness to adopt technologies and their outcomes [7]. Therefore, XAI and trust have become prominent research topics in human-computer interaction.

However, the degree to which results of previous research on explaining recommender systems can be generalized is limited because of three reasons. First, studies are mostly framed in application contexts like media recommending [e.g., 8, 27, 51, 67] and e-commerce recommending [e.g., 7, 60, 61]. Other contexts such as education are explored less [6]. Second, most study participants are university students or adults, resulting in scarce results for adolescents (ages 11–19 [25]). Third, on a methodological level, most XAI research measures the effect of explanations by comparing recommender systems with and without explanations. However, this comparison could be unfair as recent studies suggest that the mere presence of *placebo explanations* (i.e., explanations without any meaningful content) can already increase someone's trust in an intelligent system [22].

To address these limitations, we investigated how explanations affect adolescents' trust in an e-learning platform that recommends mathematics exercises, and added placebo explanations as an extra baseline. In particular, we had two research questions:

RQ1. Can explanations increase adolescents' initial trust in an e-learning platform that recommends exercises?
RQ2. How do placebo explanations influence adolescents' initial trust in such an e-learning platform?

Our research contribution is threefold. First, we show that explaining recommendations can significantly increase initial trust in an e-learning platform if trust is measured multidimensionally. However, when measuring trust one-dimensionally, the increase is not significant, which suggests that mainly dynamically learned factors grow initial trust. Second, by comparing our explanation interface with a placebo baseline, we reveal that adolescents have different needs for transparency, so tailoring explanations is essential. Third, we present unique data on how adolescents trust and interact with our e-learning platform, which we share publicly in the spirit of open science[1]. In sum, we hope our work inspires other

[1] https://github.com/JeroenOoge/explaining-recommendations-elearning

researchers to more often target adolescents and study the impact of tailored explanations in e-learning.

2 BACKGROUND AND RELATED WORK

This section discusses some challenges of explaining artificial intelligence, and particularly recommender systems. Then, it zooms in on trust in automated systems and previous studies on the trust effects of explaining recommendations.

2.1 Explainable Artificial Intelligence

Ever since the resurgence of artificial intelligence, there has been a call for algorithmic transparency. Sophisticated algorithms are namely often 'black-boxes': it is unclear how they precisely process vast amounts of input data to obtain an output. Not explaining algorithms' outcomes may suffice for low-stakes applications such as movie recommendation but becomes unacceptable in high-stakes contexts such as healthcare and e-learning. *Explainable artificial intelligence* (XAI) is an umbrella term for techniques that try to explain the logic behind algorithmic decision-making, such that people can understand it, grow appropriate trust in the algorithm, and detect potential biases [32]. A substantial challenge is that XAI encompasses many intertwined topics including trust, fairness, bias, causality, accountability, privacy, and human reasoning [3]. As a consequence, it is hard to find all-embracing definitions for XAI and concepts like 'explainability', 'interpretability', 'understandability' and 'intelligibility' [20, 28, 45].

Because of its broadness, the XAI problem can be approached from different angles. Researchers in artificial intelligence follow an *algorithmic* approach: they develop model-specific and model-agnostic techniques to investigate the local and global behavior of machine learning models and their robustness against data perturbations [4, 5, 31]. In contrast, researchers in human-computer interaction follow a *human-centered* approach: they often draw on the social sciences [21, 52] and let human reasoning processes inform XAI techniques [74]. In short, this led to the understanding that there is no such thing as a one-size-fits-all explanation. Instead, design requirements for explanations depend on the application context [18, 71] and the target audience's goals and personal characteristics [8, 51, 53]; and explanations can be evaluated according to several metrics [35, 53].

2.2 Explaining Recommendations

A lot of XAI research builds upon earlier research with recommender systems [64]. For example, Herlocker et al. [33] compared several explanation designs for collaborative filtering recommenders to increase acceptance of recommendations. Today, explaining recommender systems is still a hot research topic [e.g., 19, 37, 40, 69], generating lively reciprocity with the wider XAI domain.

In general, explanations for recommendations come in three representational forms [57]. First, *textual explanations* use natural-language phrases. Many commercial applications already employ these kinds of explanations, following patterns like "*People who liked X also liked Y*" for collaborative filtering recommenders, and "*You will like X because it has Y and Z*" for content-based recommenders. Second, *visual explanations* use (interactive) visualizations to efficiently convey a lot of information. For example, Herlocker et al. [33] used histograms to show how neighboring users rated a recommended movie; Tsai and Brusilovsky [68] explained similarity-based recommenders amongst others with radar charts and Venn diagrams; and Bostandjiev et al. [9] visualized a music recommending process with an interactive pathway chart. Third, *hybrid explanations* leverage both textual and visual information. For example, Gedikli et al. [27] used tag clouds in which word size encodes relevance, and Szymanski et al. [63] combined a partial dependence plot with text on how to interpret the visual information.

Designing explanations for recommendations brings challenges concerning *what* and *how* to explain [23]. Usually, the recommendation algorithm constrains the explanation type [66]. For example, collaborative filtering recommendations cannot be explained by their inherent features. Furthermore, designing explanations involves making several trade-offs [41]. Tintarev and Masthoff [65, 66] discussed this in detail and outlined seven goals for explanations which are not all simultaneously satisfiable: transparency, scrutability, effectiveness, persuasiveness, efficiency, satisfaction, and trust.

2.3 Trust in Automated Systems

Trusting automated systems has been found essential for adopting them [7, 61]. At the same time, trust research is somewhat controversial [17] because optimizing systems' designs to grow trust might lead to inappropriate trust, which can entail undesirable effects like misusing technology [11, 50]. In addition, trust is a complex topic. On the one hand, it has been defined in many different ways, depending on the field or context [46] and entailing different themes such as competence, benevolence, and reliance [7, 13, 14, 30, 44, 54]. On the other hand, it has been recognized that trust is not static but evolves [36, 56, 59]. Thus, measuring trust in automated systems is challenging and researchers have proposed explicit and implicit measuring techniques.

Explicit measuring techniques ask people about their trust perceptions in questionnaires or interviews. *One-dimensional* approaches measure trust with a single Likert-type question [36, 51, 56]. Although this method is quick and easy, it is susceptible to people interpreting 'trust' differently. Therefore, *multidimensional* approaches use Likert scales to measure trust as an ensemble of multiple constructs. For example, McKnight et al. [48] introduced the concept of *trusting beliefs* [73], consisting of the constructs *competence*, *benevolence*, and *integrity*. Later research added more constructs, including *perceived transparency* and *intention to return* [8, 62]. Overall, while a multidimensional approach is more nuanced than its one-dimensional counterpart, it requires longer questionnaires and is therefore more time-consuming.

Implicit measuring techniques avoid the self-reporting bias in explicit measurements by measuring trust through an intermediary. Examples are: loyalty measured by the number of logins after sign-up [49, 66], acceptance rate for recommendations [14], time spent on a page, click-through rate, and page-exiting manner [26]. In the context of explaining recommender systems, implicit measurements for trust have not yet been widely adopted, possibly because intermediaries like loyalty require long(er)-term studies.

2.4 Trust in Explained Recommendations

Previous research has shown that providing explanations for recommendations can increase the acceptance of recommendations [14, 33], and increase people's trust in the recommender system [8, 61]. While previous studies typically focused on recommenders for movies or e-commerce [e.g., 42], research in an e-learning context is limited [5, 15]. This is unfortunate as Abdi et al. [2] recently demonstrated the potential of a transparent educational recommender system: an Open Learner Model [10] improved understanding of and trust in recommendations for learning materials.

As trust is a relative measure, it must be compared to some baseline. Studies on the effects of explanations typically include a baseline with no explanations. However, a lesser applied baseline are *placebo explanations*. These 'pseudo explanations' are semantically insensible [43], i.e., they do not reveal any information about why something was recommended, for example "*This has been recommended to you because this is what the algorithm calculated.*" Surprisingly, Eiband et al. [22] found that placebo explanations can invoke similar trust levels as real explanations. However, Nourani et al. [55] found conflicting results outside the domain of recommender systems: placebo explanations lowered the perceived accuracy of an image recognition system.

2.5 Underexplored Research Areas

Our literature overview shows that XAI re-nourishes the interest in explaining recommender systems and how that affects trust in recommendations. However, we see two underexplored areas. First, research on trust and explaining recommender systems primarily focuses on university students or adults and often neglects adolescents. Second, while e-learning platforms increasingly adopt recommendation algorithms [2, 16, 39, 47, 72], they lack explanations for their recommendations. Our research addresses both shortcomings: we design hybrid explanations for an exercise recommender on an e-learning platform and investigate their effects on adolescents' *initial* trust (i.e., trust based on their first impressions of the platform).

3 MATERIALS AND METHODS

This section presents our e-learning platform with explanations for recommended exercises and our overall study design. Our research was approved by the ethical committee of KU Leuven (reference number G-2021-3233-R2(MAR)).

3.1 E-learning Platform with an Exercise Recommender

For our study, we built upon an existing e-learning platform called Wiski [58], which was developed in Drupal 7 and contains over 1000 multiple choice exercises on mathematics topics in the Belgian high school curriculum. To estimate the difficulty level of exercises for each student, we set up an *Elo rating system* [24] for students and exercises: if a student correctly solves an exercise, their Elo score rises and the exercise's Elo score drops, and vice versa.

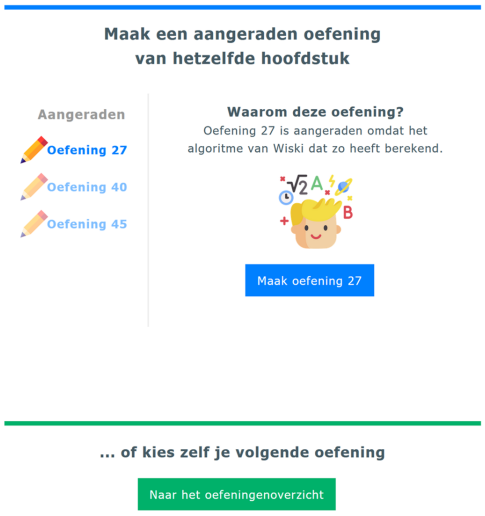
We used the Elo rating in two ways. First, students could see the estimated difficulties while browsing exercises (see Figure 1d) to manually pick exercises suited for their level of mastery. Second, inspired by Dahl and Fykse [16], we recommended exercises with an algorithm implemented in Python 3.8.5. When students solved an exercise on a certain topic, they received three suggestions for follow-up exercises on the same topic. Broadly, our recommender system combines Elo ratings and collaborative filtering: it looks for candidate exercises based on a student's Elo rating and recommends those that the student is most likely to answer correctly. More specifically, to recommend exercises about topic T for student A, our algorithm follows three steps. First, the 7 exercises about topic T with an Elo score closest to the value $\text{Elo}_A + 50$ are selected as candidates. We added the constant 50 to promote recommendations that slightly exceed students' level of mastery [76]. Then, for each candidate exercise E, the algorithm estimates with nearest-neighbors how many attempts A may need to solve E: it looks for students who solved E, selects at most 40 of them close to A in terms of attempts for previously solved exercises (Pearson similarity), and takes a weighted average of their number of attempts for E. Finally, the three candidate exercises with the lowest average number of attempts are recommended in ascending order.

3.2 Explanations for Recommendations

To accompany the recommended exercises, we designed three explanation interfaces, following a user-centered design process. Specifically, we iteratively refined an initial design during three rounds of think-aloud studies with 16 participants (1 teacher, 5 middle school students, 9 high school students, 1 university student). In these think-alouds, participants executed predefined tasks that tested the usability of our interfaces and answered additional questions related to usability, transparency, and explanations in general. We wrote down all relevant remarks and afterwards grouped them thematically to identify the most frequent issues. Based on the collected feedback, we dropped initial designs for transparency pages that explained collaborative filtering, and made the role of certain components in our explanation interfaces more explicit such that students could process them quicker. More details can be found in Kato's Master's thesis [38].

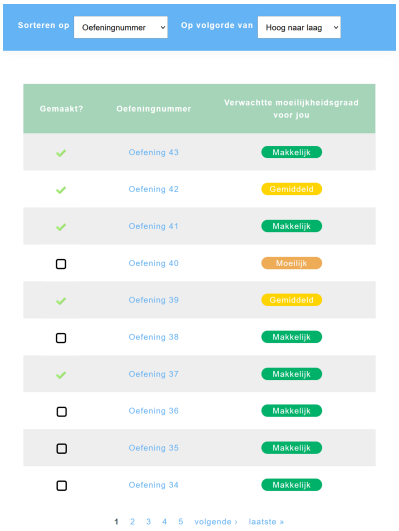
Figure 1 presents our three final explanation interfaces. The first interface (Figure 1a) contains a real explanation, consisting of three parts [*English translation in brackets*]: 1 a why-statement which indicates that the exercise was recommended based on both the student's level of mastery and the exercise's difficulty [*Why this exercise? Wiski thinks your current level matches that of this exercise!*]; 2 a justification-statement with the student's estimated number of tries needed to solve the exercise [*Wiski expects that you will need 1 or 2 attempts to answer exercise X correctly, based on your results and that of your peers*]; 3 a histogram of how many tries similar students required for the exercise, inspired by Herlocker et al. [33] [*Number of attempts peers needed to solve exercise X correctly*]. To avoid students seeing (nearly) empty histograms at the experiment's cold start, we pre-populated the data set with mock data based on logging data from a past experiment on Wiski that used identical exercises [58]. The second interface (Figure 1b) contains the placebo explanation "*Exercise X is recommended because this is what Wiski's algorithm calculated*", which indeed conveys no information about how our recommendation algorithm works. Finally, the third interface (Figure 1c) simply states that the exercise was recommended, without further clarification.

(a) A real explanation for the REAL group with 1 a why-statement, 2 justification-statement, and 3 histogram.



(b) A placebo explanation for the PLACEBO group with a why-statement that the exercise is recommended by an algorithm.

(c) No explanation for the NONE group, only a statement that the exercise is recommended.



(d) Exercise list: rows contain an indication of being solved, a link to the exercise, and a difficulty label (easy, average, hard).

Figure 1: The three explanation interfaces in our randomized controlled experiment (a–c). In each interface, the top part (blue) shows real, placebo, or no explanations. The bottom part (green) allows students to return to the exercise overview (d).

3.3 Participant Recruitment

We contacted teachers of 18 high schools in Belgium (Flanders) and invited them and their students to participate in our research. Teachers and students received an information leaflet that described the research process, stressing that students could not be coerced into participating and would receive an equivalent substitute task if they did not wish to participate. Interested students then gave informed consent and students under the age of 16 also required signatures from their parents. In addition, we recruited extra participants through snowball sampling.

3.4 Study Design

To assess the effects of our explanation interfaces on initial trust, we conducted a randomized controlled experiment [29] with three research groups: *REAL*, *PLACEBO*, and *NONE*, corresponding to the explanation interfaces in Figure 1a to 1c, respectively. Following the steps in Figure 2, all participants (1) registered on our platform and were randomly assigned a research group; (2) answered a pre-study questionnaire with questions related to their demographics, experience with computers and e-learning platforms, mathematical background, and self-perceived mastery in mathematics; (3) solved

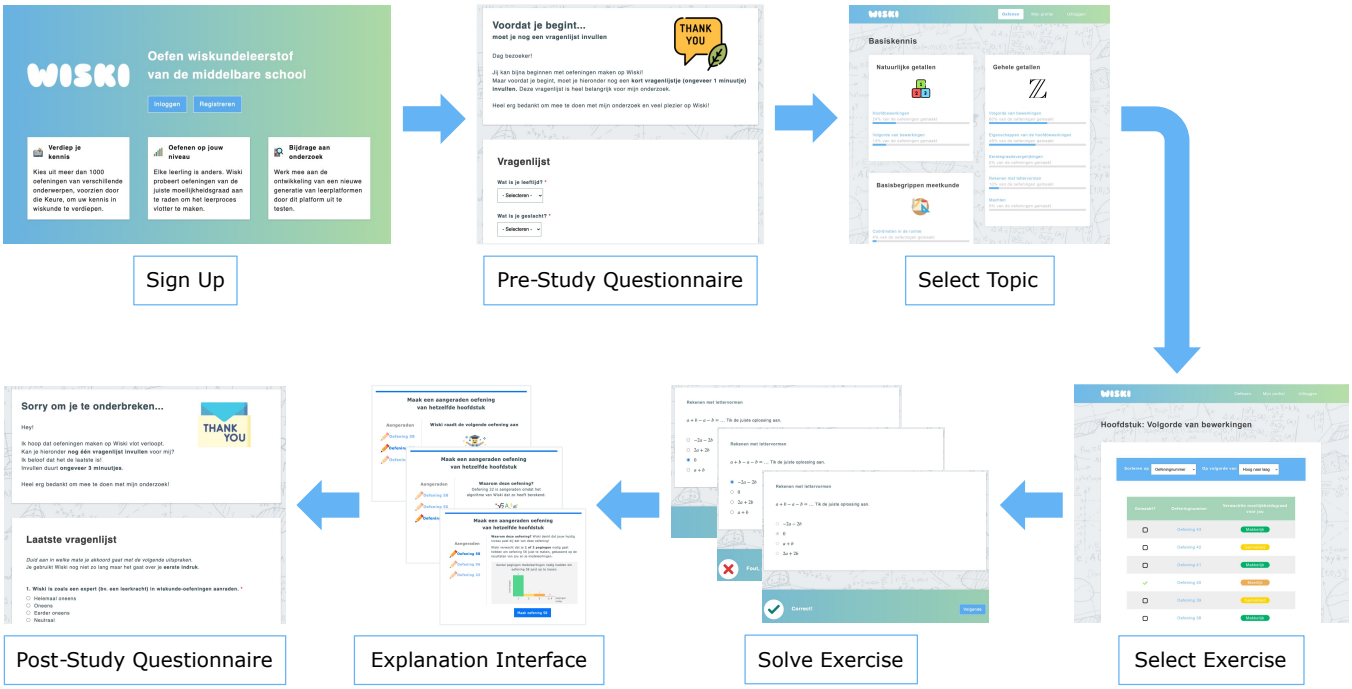


Figure 2: Flow chart of our study: sign up, pre-study questionnaire, solving exercises and interacting with an explanation interface five times, and post-study questionnaire.

five exercises and interacted with their research group's explanation interface after each exercise; (4) answered the post-study questionnaire in Table 2 with questions on trust; and (5) optionally used the platform freely until the end of the study. Thus, participants' experience on our platform only differed in the explanation interface shown after solving exercises. In the background, we also logged whether participants selected recommended exercises.

We decided to let participants answer the post-study questionnaire after five exercises because (a) they then all interacted with an explanation interface equally often, and (b) they often participated during a mathematics period at school and needed to finish in under an hour. The post-study questionnaire itself contained nineteen 7-point Likert-type questions divided into seven groups (see Table 2). We measured trusting beliefs, consisting of *Competence* (Q1–Q5), *Benevolence* (Q6–Q8), and *Integrity* (Q9–Q11) with a validated questionnaire by Wang and Benbasat [7]. To fit the original questions in the scope of Wiski, we translated them to Dutch and made them easier to understand for adolescents by simplifying some vocabulary. The average of the scores for trusting beliefs, *Intention to return* (Q13–Q14), and *Perceived transparency* (Q15) yielded a multidimensional trust score. In contrast, *Trust* (Q12) assessed one-dimensional trust by explicitly asking about trust in Wiski's recommendations. Finally, *General questions* (Q16–Q19) collected extra information about how participants perceived explanations. Furthermore, after each question group, we added a text field in which participants could motivate their Likert-type responses. In the end, we thematically analyzed these written qualitative data to gain further insights into participants' rationale for picking a specific quantitative score. Measuring trust through the above-mentioned constructs aligns with how other recommender systems are evaluated in the literature [7, 8, 12, 14, 27].

3.5 Statistical Analysis

We analyzed our data with Pingouin 0.3.11 [70] in Python 3.8.5. We used non-parametric statistics to avoid normality assumptions, similar to other studies involving Likert-type data [e.g., 2, 14]. More specifically, we tested for significant differences between research groups with Mann-Whitney U and used Kendall's τ to test for correlations. To interpret the former as a test for difference in medians, we assumed equal data distributions in our research groups.

4 RESULTS

In total, 37 students (ages 13–18, 13 male, 24 female) participated in our research: 3 students were from 9th grade, 18 from 10th grade, 8 from 11th grade, and 8 from 12th grade. Figure 7 shows their distribution over the three research groups: 12 in REAL, 12 in PLACEBO, and 13 in NONE. Figures 3 and 4 plot their responses to the post-study questionnaire.

4.1 Effects of Real Explanations

Table 1a and 1b depict the outcomes of one-sided Mann-Whitney U tests, comparing REAL to NONE, and REAL to PLACEBO. Median competence, trusting beliefs, perceived transparency, and multidimensional trust were significantly higher in REAL ($p < 0.05$). However, there was no significant increase in integrity, one-dimensional trust or intention to return. For benevolence, there was only a significant increase ($p < 0.05$) when comparing REAL to NONE.

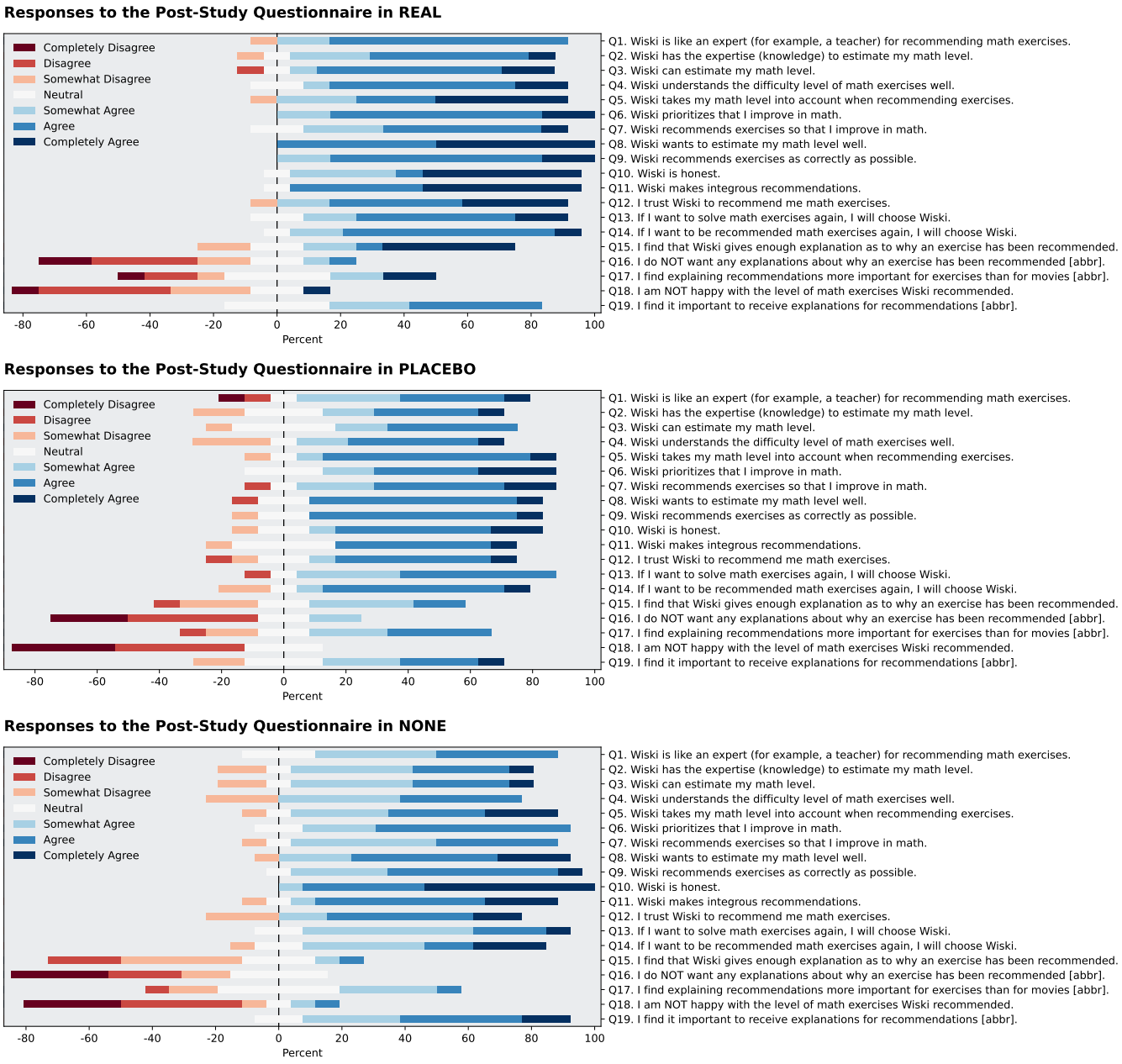


Figure 3: Diverging bar charts of the responses to the post-study questionnaire in Table 2 for each research group.

The qualitative responses[2] on Q15 showed that perceived transparency was somewhat controversial in REAL. Some participants were positive about the explanations: "*I found the explanation that Wiski gave correct and satisfactory.*" Other participants did not seem to be satisfied with the explanations and may have wanted a different type of explanation: "*Doesn't it just state how many tries Wiski thinks I would need to find the correct answer. It doesn't explain specifically.*" Finally, there was also evidence that some participants did not require explanations: "*I didn't really read the explanation…*"

4.2 Effects of Placebo Explanations

Two-sided Mann-Whitney U tests did not reveal any significant difference ($p < 0.05$) between PLACEBO and NONE: the smallest p-values were 0.099 (perceived transparency) and 0.143 (integrity); all other values were above 0.327. Still, it is interesting that in

[2]We translated the original Dutch responses to English as literally as possible.

Table 1: Results of one-sided Mann-Whitney U tests comparing the research groups. The common language effect size is the probability that a random value from the first group is greater than a random value from the second group.

(a) REAL vs. NONE

	p-value	U-value	CLES
Competence	0.030*	113.0	0.724
Benevolence	0.030*	112.5	0.721
Integrity	0.261	90.0	0.577
Trusting beliefs	0.048*	109.0	0.699
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
Multidimensional trust	0.002**	131.0	0.840

*$p < 0.05$, **$p < 0.01$, CLES = common language effect size

(b) REAL vs. PLACEBO

	p-value	U-value	CLES
Competence	0.023*	106.5	0.740
Benevolence	0.074	97.0	0.674
Integrity	0.054	100.0	0.694
Trusting beliefs	0.026*	106.0	0.736
Intention to return	0.139	90.0	0.625
Perceived transparency	0.041*	102.0	0.708
One-dimensional trust	0.071	96.5	0.670
Multidimensional trust	0.013*	111.0	0.771

*$p < 0.05$, CLES = common language effect size

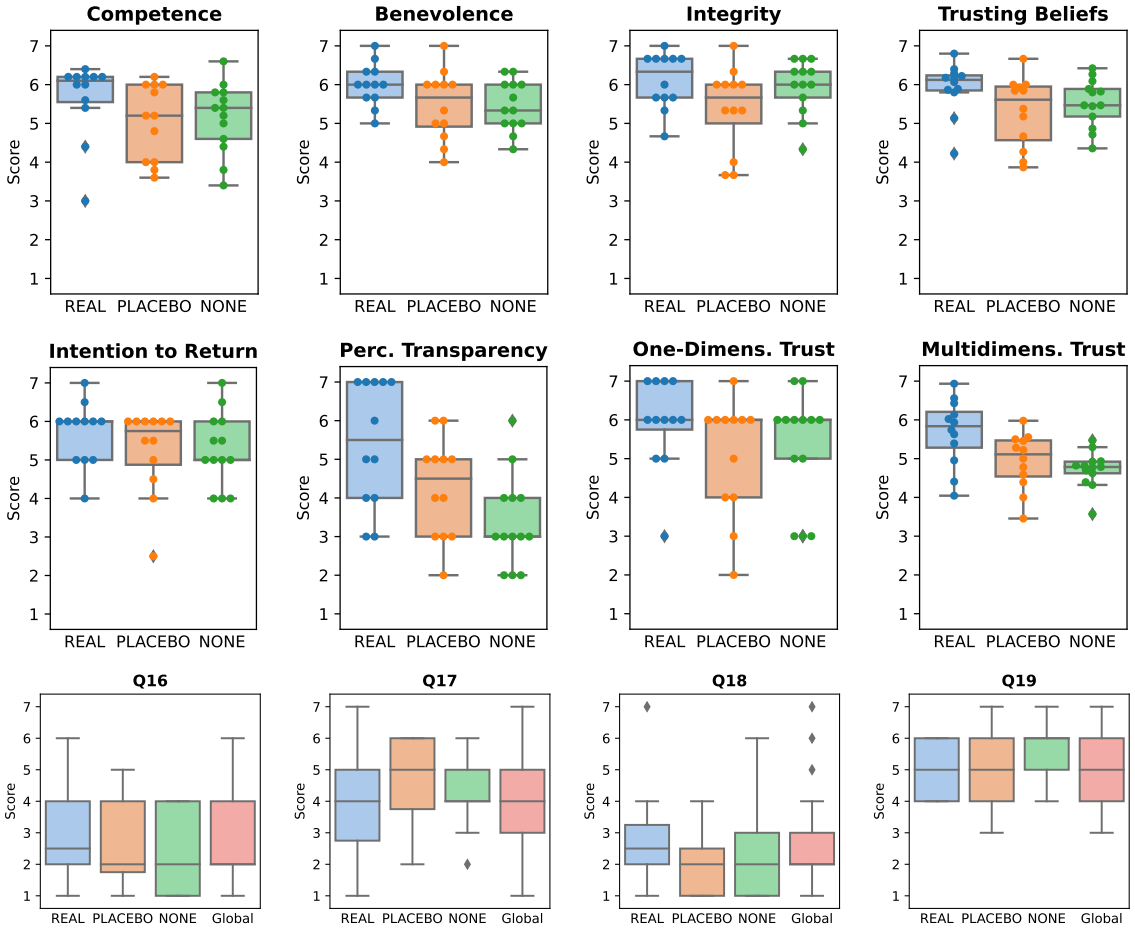


Figure 4: Box plots of the responses to the post-study questionnaire in Table 2 for each research group.

our sample PLACEBO got the lowest median for competence and integrity (see Figure 4).

As in REAL, the qualitative responses concerning perceived transparency (Q15) showed very different sentiments in PLACEBO. On the one hand, some participants did not perceive the placebo explanations as real explanations, as seen in responses like "*Wiski just says calculated by the algorithm of…*" and "*It would be nice for an extensive explanation as to why it is better to solve this exercise.*" On the other hand, several participants found the explanation satisfactory, stating: "*Wiski says that the algorithm recommends the next exercise thus I trust the algorithm*" and "*I don't think that there needs to be more explanation as to why an exercise has been recommended.*"

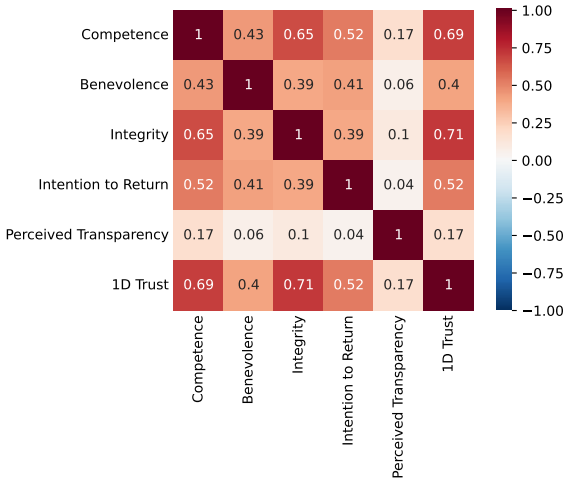


Figure 5: Kendall's τ correlations between trust constructs and one-dimensional trust.

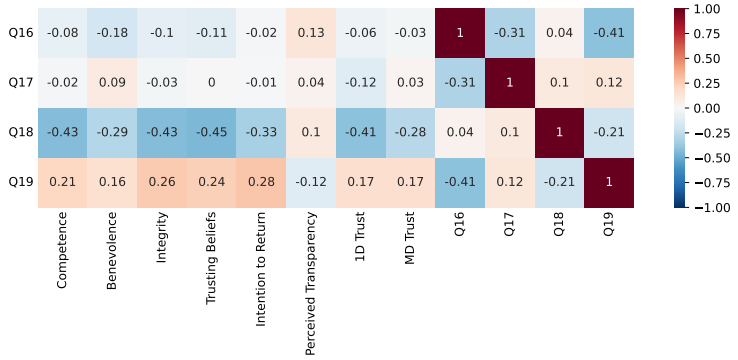


Figure 6: Kendall's τ correlations between trust constructs and questions on the need for explanations (Q16–Q19).

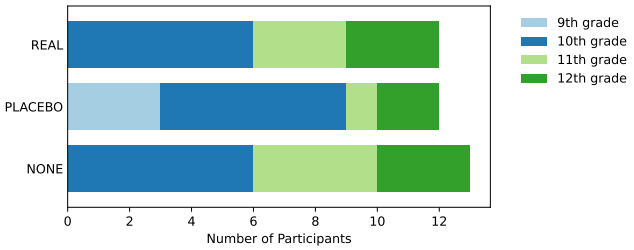


Figure 7: Distribution of the 37 participating students over the three research groups.

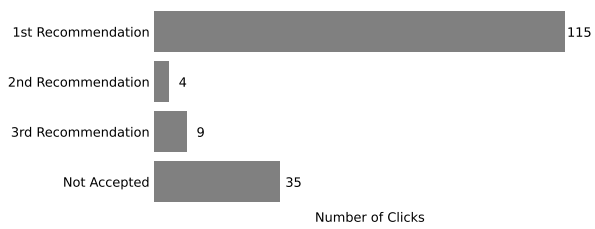


Figure 8: Distribution of how often each option in the explanation interface was clicked.

4.3 Effects of No Explanations

The qualitative responses on Q15 were quite consistent within NONE: close to all participants who gave a meaningful response indicated that they did not see an explanation or missed it. For example, one participant stated: "*I find it unfortunate that [Wiski] does not say why a certain exercise was recommended. It is nice to know why this exercise fits you, but there should also not be too much information as then it would not be fun to read.*" Yet, surprisingly, two participants seemed to believe they *did* receive explanations: "*If you want to solve a new exercise, it is useful that you know why this exercise is recommended, the website does this well*" and "*Yes I find that there is enough explanation.*" Finally, one participant formed a particular mental model of our recommender system: they believed the recommendations depended on the self-reported mastery level of mathematics in the pre-study questionnaire.

4.4 Correlations

Figure 5 shows the correlations between the various trust constructs and one-dimensional trust: competence ($\tau = 0.69$) and integrity ($\tau = 0.71$) are correlated the most, whereas perceived transparency ($\tau = 0.17$) the least. In fact, perceived transparency has little to no correlation with any of the trust constructs. Figure 6 shows how all trust scores and questions Q16–Q19 are correlated. Especially notable is the moderate correlation between satisfaction with the level of recommended exercises (Q18) and most trust scores. We also found that one-dimensional trust is moderately correlated with trusting beliefs ($\tau = 0.68$) and multidimensional trust ($\tau = 0.52$). The latter two constructs are in their turn correlated too ($\tau = 0.56$).

4.5 Recommendation Clicks

Recall from Section 3.1 that, after participants solved an exercise about topic T, our explanation interfaces recommended three exercises to solve next. Participants could either accept one of these recommendations or ignore them and return to the exercise overview for topic T (Figure 1d) to select a next exercise themselves. Figure 8 shows that participants mostly decided to solve the first recommended exercise, followed by returning to the exercise overview. In addition, one-sided Mann-Whitney U tests revealed that the NONE group accepted significantly less recommendations than both REAL ($p = 0.007$, $U = 67$, CLES $= 0.827$) and PLACEBO ($p = 0.039$, $U = 72$, CLES $= 0.727$).

5 DISCUSSION

This section answers our research questions by discussing how adding real, placebo, or no explanations to our e-learning platform affected adolescents' initial trust in our platform. Then, based on the observations, it underlines the need for tailoring explanations, and reflects upon the broader scope of explanations and recommendations in e-learning.

5.1 Explanations Increase Multidimensional Initial Trust...

Previous work has shown that well-designed explanation interfaces can increase adults' trust in recommendations [22, 62, 77]. RQ1 asks whether the same holds for adolescents in an e-learning context. Two parts of our results suggest a confirmatory answer if trust is defined as an average of trusting beliefs, intention to return, and perceived transparency.

First, Table 1a shows that adding explanations significantly increased two out of three trust constructs: trusting beliefs and perceived transparency. The third construct, intention to return, was not significantly affected, which conflicts with the findings from Pu and Chen [62]: they reported that higher competence perception results in higher intention to return. One possible reason for this conflict might be that Pu and Chen's explanations assisted in buying expensive products, which seems more precarious than solving recommended exercises on an e-learning platform.

Second, participants with real explanations accepted significantly more recommended exercises than participants with placebo or no explanations. Building upon Cramer et al.'s [14] observation that acceptance of recommendations is correlated to trust, this further suggests that trust was higher for adolescents who saw real explanations.

5.2 ...But Not One-Dimensional Initial Trust

However, if trust is measured one-dimensionally with a single Likert-type question, there was *no* significant increase in trust compared to using placebo or no explanations. This shows that RQ1 cannot be answered in a univocal way, and puts our findings for increased trusting beliefs and multidimensional trust into perspective. First, our results seem to imply that multidimensional trust measurements are more nuanced than their one-dimensional counterpart, which matches with the well-known statement that trust is multi-faceted and cannot be fully captured by a single question [34, 59]. Second, as most participants across the three research groups reported relatively high one-dimensional trust (see Figure 4), the explanations may not have been the most important factor for trusting the e-learning platform. Instead, participants may have built initial trust mainly because of dynamically learned factors [34] such as the perceived accuracy of the recommender system, the exercises' overall quality, or the platform's appearance. This is further backed by the correlations in Figures 5 and 6: whereas one-dimensional trust is barely correlated to perceived transparency and need for explanations (Q16, Q17, Q19), it *is* correlated to integrity, competence, and being satisfied with the exercises' level (Q18). Thus, explanations for recommendations seem to increase competence, which in turn increases initial trust. This further justifies the presence of competence in many definitions of trust [30, 54, 75].

5.3 Placebo Explanations Are a Useful Baseline

RQ2 is concerned with how placebo explanations influence adolescents' initial trust in our e-learning platform. We found no significant differences in initial trust when using placebo explanations over no explanations. This differs from Eiband et al.'s results [22], who found that placebo explanations *do* increase trust compared to no explanations. Reasons for the differing results could be the low sample size in both their and our study, the different study context, or the different methods for measuring trust. On a methodological level, Eiband et al. [22] suggest using placebo explanations as a placeholder when insufficient information is available for real explanations. Based on our results, however, we would discourage this as it may undermine the platform's perceived transparency, competence, and integrity (see Figure 4 and Table 1b; the p-value for integrity is only slightly larger than 0.05).

However, when studying the impact of explanations, we do see several advantages for using placebo explanations as a baseline. For example, they allow to collect information about how critical participants stand towards explanations and how attentive they are. In our study, we find it rather encouraging that most adolescents noticed that our placebo explanations were meaningless. Furthermore, combining placebo explanations and qualitative responses allows to gain insights into how much transparency participants actually need. In our study, some adolescents required a more detailed explanation while others did not require much or any transparency. This underlines the importance of research on tailoring explanations based on transparency needs.

5.4 Tailoring Explanations Remains Important

Our qualitative data show that not all adolescents perceived the utility and transparency of our explanation interfaces in the same way. Some adolescents even had their own perception of what a good explanation is and sought explanations that go beyond our focus on exercises' difficulty level and estimated number of attempts. To accommodate different transparency needs, it seems essential to tailor explanations to the audience that sees them.

On the one hand, the think-aloud studies in our user-centered design process gave us some insights into *what* parts of our real explanation interface may be tailored. First, middle school students (7th and 8th grade) typically found it harder to understand the histogram in our explanation, which suggests that this particular age group might require additional clarification for the histogram or an entirely different (visual) explanation. Second, some participants valued explicit wordings in the interface as it allowed them to process the given information quicker and better, while others considered this as rather redundant.

On the other hand, we can only speculate on *how* to concretize the tailoring process. One possibility is to give adolescents *direct* control over the explanations' type or detail level, or over whether they see any explanations at all. In practice, this could be done by iteratively querying students who are exposed to explanations and then modifying those explanations based on their indicated needs. A potential drawback is that incomplete or no explanations can negatively impact adolescents' mental model of the recommender system, as illustrated by the participant in our NONE group who

believed that the exercise recommendation depended on their self-reported mastery in mathematics. Another possibility to tailor explanations is to *indirectly* customize them according to personal characteristics [8, 51]. There is, however, an ethical challenge here as underage adolescents cannot or should not always pass delicate personality information without parental consent.

5.5 Taking a Step Back: Recommendations and Explanations in E-Learning

To conclude, we briefly reflect upon the premise of recommending exercises and explaining the underlying algorithm in e-learning. Do recommendations always need explanations? Should e-learning platforms always recommend exercises? We distinguish between situations in which little or much is at stake.

In *low-stakes* situations, accepting unsuitable recommendations does not have severe repercussions, so quickly accepting whichever recommendation seems reasonable. In our short-term experiment, students understood that accepting recommendations involved little risk, which may explain why they most often selected the first recommended exercise (all participants were aware of *three* recommendations in our think-aloud studies, so we assume this holds for our final study). In addition, some teachers instructed students to drill a specific topic, so it is plausible that some students were more interested in solving as many exercises as possible rather than carefully choosing their next exercise. In such 'drilling' situations, recommending only one exercise (the best fit) at a time might be sufficient, and full-fledged explanations might be excessive. However, in our experiment, students who were left in the dark as to why an exercise was recommended were more eager to select one themselves in the exercises overview. Perhaps this was the case because they perceived the displayed difficulty levels (see Figure 1d) as a kind of explanation. Thus, even in low-stakes contexts, it seems desirable to provide some minimal information about the (recommended) exercises.

In *high-stakes* situations, it becomes more important to investigate the benefit of recommendations, and there, we hypothesize that explanations become more important too. When students have limited time to prepare for an exam, for example, it seems plausible that they seek a justification for why they should spend time solving a recommended exercise. Regarding recommendation, we have three remarks: (1) in a school context, teachers are in the perfect position to judge which topics are best suited for a particular student, so it is interesting to study how they can steer recommendations based on their domain knowledge; (2) we believe it remains important to give students the freedom to select exercises themselves, for example to follow teachers' instructions; (3) contrary to our basic recommender system with one overall Elo score for each student, more sophisticated algorithms [e.g., 1] could work with topic-specific Elo scores and process students' and teachers' feedback on the Elo scores to converge towards reasonable ratings more quickly.

5.6 Limitations and Future Work

Our research has limitations that affect the generalizability of our results. First, with only 37 participants divided over three research groups, our sample is relatively small. In addition, although we specifically focused on adolescents, the age range of 13–18 is still relatively large, especially given the turbulent stage of life that it spans. Thus, our results should be interpreted cautiously. Second, since Elo scores of students and exercises become more accurate as more students solve exercises, the accuracy of recommendations and explanations might have changed during the experiment. However, as participants were equally satisfied with the level of recommended exercises (Q18, see Figure 4), this should not have biased the results significantly. Third, some participants communicated that the exercises on our platform are rather basic. If solving an exercise takes an insignificant amount of time, the importance of picking a suitable recommendation becomes smaller. Future studies could thus be conducted with more challenging exercises to investigate whether our results hold. Fourth, although the post-study questions for trusting beliefs were based on those by Wang and Benbasat [7], we modified and translated them to match them to an e-learning context and adolescents. Future work can validate our questionnaire. Fifth, our short-term study could only assess initial trust, whereas trust evolves [36, 56, 59]. Long-term studies could measure trust implicitly through loyalty [49, 66]. Overall, our methods and our valuable data on how adolescents trust and interact with a recommender system can be used as starting points for future research.

6 CONCLUSION

This paper tackled the complex topic of trust in an e-learning platform that explains why it recommends certain exercises. Specifically, we investigated how real and placebo explanations affect initial trust. Contrary to the vast majority of other human-computer interaction research on this topic, we focused on adolescents as the target audience.

Our randomized controlled experiment with 37 high school students showed that our explanation interface increases adolescents' initial trust when trust is measured as a multidimensional construct of trusting beliefs, intention to return, and perceived transparency. However, this effect did not hold when we considered measurements of a single Likert-type question on trust. This two-sided result seems to imply that one question cannot capture the multifaceted nature of trust and that dynamically learned factors such as perceived accuracy of the recommendation algorithm and the website's appearance may be the leading cause for gaining initial trust in our e-learning platform. Furthermore, compared to using no explanations, we found that placebo explanations did not offer any significant trust differences quantitatively. However, the divisive qualitative responses revealed that tailoring explanations based on transparency needs remains essential. Finally, we reflected upon whether explanations and recommendations are always desirable in e-learning, distinguishing between low- and high-stakes situations.

In sum, while our study has some limitations, our results do seem to indicate that explaining recommendations on an e-learning platform is an asset for high school students. Therefore, accompanying recommendations with explanations should be considered when designing e-learning applications similar to ours for adolescents. We also advise researchers who study the impact of tailored explanations to include placebo baselines in their studies: they may

give more insights into how much transparency people actually need, compared to no-explanation baselines alone.

ACKNOWLEDGMENTS

We are very grateful to all involved adolescents for participating in our studies, their parents for giving parental consent, and their mathematics teachers for inviting us into their (virtual) classroom. This work was supported by the Research Foundation–Flanders (FWO, grant G0A3319N) and the imec.icon project AIDA financed by Flanders Innovation & Entrepreneurship (grant HB.2020.2373).

REFERENCES

[1] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2019. A Multivariate Elo-based Learner Model for Adaptive Educational Systems. *arXiv:1910.12581 [cs]* (Oct. 2019). arXiv:1910.12581 [cs]

[2] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2020. Complementing Educational Recommender Systems with Open Learner Models. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. Association for Computing Machinery, New York, NY, USA, 360–365.

[3] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156

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

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

[6] Jordan Barria-Pineda. 2020. Exploring the Need for Transparency in Educational Recommender Systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. Association for Computing Machinery, New York, NY, USA, 376–379.

[7] Izak Benbasat and Weiquan Wang. 2005. Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems* 6, 3 (March 2005), 72–101. https://doi.org/10.17705/1jais.00065

[8] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend? User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. Association for Computing Machinery, New York, NY, USA, 287–300. https://doi.org/10.1145/3025171.3025209

[9] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: A Visual Interactive Hybrid Recommender System. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. Association for Computing Machinery, New York, NY, USA, 35–42. https://doi.org/10.1145/2365952.2365964

[10] Susan Bull and Judy Kay. 2010. Open Learner Models. In *Advances in Intelligent Tutoring Systems*, Janusz Kacprzyk, Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi (Eds.). Vol. 308. Springer Berlin Heidelberg, Berlin, Heidelberg, 301–322. https://doi.org/10.1007/978-3-642-14363-2_15

[11] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. IEEE, Dallas, TX, USA, 160–169. https://doi.org/10.1109/ICHI.2015.26

[12] Li Chen (Ed.). 2008. *User Decision Improvement and Trust Building in Product Recommender Systems*. EPFL, Lausanne. https://doi.org/10.5075/epfl-thesis-4140

[13] K. Chopra and W.A. Wallace. 2003. Trust in Electronic Environments. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003.* IEEE, Big Island, HI, USA, 10 pp.–. https://doi.org/10.1109/HICSS.2003.1174902

[14] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The Effects of Transparency on Trust in and Acceptance of a Content-Based Art Recommender. *User Modeling and User-Adapted Interaction* 18, 5 (Nov. 2008), 455–496. https://doi.org/10.1007/s11257-008-9051-3

[15] Julie Bu Daher, Armelle Brun, and Anne Boyer. 2017. *A Review on Explanations in Recommender Systems*. Technical Report. LORIA - Université de Lorraine.

[16] Ole Halvor Dahl and Olav Fykse. 2018. *Combining Elo Rating and Collaborative Filtering to Improve Learner Ability Estimation in an E-Learning Context*. Master's thesis. NTNU.

[17] Brittany Davis, Maria Glenski, William Sealy, and Dustin Arendt. 2020. Measure Utility, Gain Trust: Practical Advice for XAI Researchers. In *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, Salt Lake City, UT, USA, 1–8. https://doi.org/10.1109/TREX51495.2020.00005

[18] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who Needs to Know What, When?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. Association for Computing Machinery, New York, NY, USA, 1591–1602.

[19] Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2020. Explaining Recommendations by Means of Aspect-Based Transparent Memories. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 166–176. https://doi.org/10.1145/3377325.3377520

[20] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]* (March 2017). arXiv:1702.08608 [cs, stat]

[21] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Vol. 12424. Springer International Publishing, Cham, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33

[22] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312787

[23] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo Japan, 211–223. https://doi.org/10.1145/3172944.3172961

[24] Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Pub, New York.

[25] Daniel Fitton, Janet C C. Read, and Matthew Horton. 2013. The Challenge of Working with Teens as Participants in Interaction Design. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*. ACM Press, Paris, France, 205. https://doi.org/10.1145/2468356.2468394

[26] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Transactions on Information Systems* 23, 2 (April 2005), 147–168. https://doi.org/10.1145/1059981.1059982

[27] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *International Journal of Human-Computer Studies* 72, 4 (April 2014), 367–382. https://doi.org/10.1016/j.ijhcs.2013.12.007

[28] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, Turin, Italy, 80–89. https://doi.org/10.1109/DSAA.2018.00018

[29] Rachel Glennerster and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, Princeton, New Jersey. https://doi.org/10.1515/9781400848447

[30] Tyrone Grandison and Morris Sloman. 2000. A Survey of Trust in Internet Applications. *IEEE Communications Surveys Tutorials* 3, 4 (2000), 2–16. https://doi.org/10.1109/COMST.2000.5340804

[31] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (Jan. 2019), 1–42. https://doi.org/10.1145/3236009

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

[33] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 241–250. https://doi.org/10.1145/358916.358995

[34] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434. https://doi.org/10.1177/0018720814547570

[35] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. *arXiv:1812.04608 [cs]* (Feb. 2019). arXiv:1812.04608 [cs]

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

[37] Y. Jin, N. Tintarev, and K. Verbert. 2018. Effects of Personal Characteristics on Music Recommender Systems with Different Levels of Controllability. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*. Association for Computing Machinery, Vancouver, British Columbia, Canada, 13–21. https://doi.org/10.1145/3240323.3240358

[38] Shotallo Kato. 2021. *Practicing the Right Math: Enhancing Trust in an E-Learning Platform Using an Explainable Recommender System.* Master's thesis. KU Leuven, Faculteit Ingenieurswetenschappen.

[39] S. Klinkenberg, M. Straatemeier, and H. L. J. van der Maas. 2011. Computer Adaptive Practice of Maths Ability Using a New Item Response Model for on the Fly Ability and Difficulty Estimation. *Computers & Education* 57, 2 (Sept. 2011), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

[40] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 379–390. https://doi.org/10.1145/3301275.3302306

[41] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, San Jose, CA, USA, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[42] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300717

[43] Ellen Langer, Arthur Blank, and Benzion Chanowitz. 1978. The Mindlessness of Ostensibly Thoughtful Action: The Role of "Placebic" Information in Interpersonal Interaction. *Journal of Personality and Social Psychology* 36, 6 (1978), 635–642. https://doi.org/10.1037/0022-3514.36.6.635

[44] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[45] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery. *Queue* 16, 3 (June 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[46] Maria Madsen and Shirley Gregor. 2000. Measuring Human-Computer Trust. In *Proceedings of the 11th Australasian Conference on Information Systems*, Vol. 53. Australasian Association for Information Systems, Brisbane, Australia, 6–8.

[47] Nikos Manouselis, Hendrik Drachsler, Katrien Verbert, and Olga C. Santos (Eds.). 2014. *Recommender Systems for Technology Enhanced Learning*. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4939-0530-0

[48] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for E-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (Sept. 2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81

[49] Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl. 2003. Interfaces for Eliciting New User Preferences in Recommender Systems. In *User Modeling 2003*, Peter Brusilovsky, Albert Corbett, and Fiorella de Rosis (Eds.). Vol. 2702. Springer Berlin Heidelberg, Berlin, Heidelberg, 178–187. https://doi.org/10.1007/3-540-44963-9_24

[50] Stephanie M. Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. 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 (June 2013), 520–534. https://doi.org/10.1177/0018720812465081

[51] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To Explain or Not to Explain: The Effects of Personal Characteristics When Explaining Music Recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 397–407. https://doi.org/10.1145/3301275.3302313

[52] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[53] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Aug. 2021), 24:1–24:45. https://doi.org/10.1145/3387166

[54] Bonnie M. Muir. 1987. Trust between Humans and Machines, and the Design of Decision Aids. *International Journal of Man-Machine Studies* 27, 5-6 (Nov. 1987), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

[55] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D. Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 97–105.

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

[57] Ingrid Nunes, this link will open in a new window Link to external site, and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User - Adapted Interaction* 27, 3-5 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[58] Jeroen Ooge. 2019. *Het personaliseren van motivationele strategieën en gamificationtechnieken m.b.v. recommendersystemen.* Master's thesis. KU Leuven, Faculteit Wetenschappen.

[59] Jeroen Ooge and Katrien Verbert. 2021. Trust in Prediction Models: A Mixed-Methods Pilot Study on the Impact of Domain Expertise. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, New Orleans, LA, USA, 8–13. https://doi.org/10.1109/TREX53765.2021.00007

[60] Umberto Panniello, Michele Gorgoglione, and Alexander Tuzhilin. 2016. In CARSs We Trust: How Context-Aware Recommendations Affect Customers' Trust and Other Business Performance Measures of Recommender Systems. *Information Systems Research* 27, 1 (2016), 182–196.

[61] Pearl Pu and Li Chen. 2006. Trust Building with Explanation Interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces - IUI '06*. ACM Press, Sydney, Australia, 93. https://doi.org/10.1145/1111449.1111475

[62] Pearl Pu and Li Chen. 2007. Trust-Inspiring Explanation Interfaces for Recommender Systems. *Knowledge-Based Systems* 20, 6 (Aug. 2007), 542–556. https://doi.org/10.1016/j.knosys.2007.04.004

[63] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 109–119. https://doi.org/10.1145/3397481.3450662

[64] Nava Tintarev and Judith Masthoff. 2007. Effective Explanations of Recommendations: User-Centered Design. In *Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys '07)*. Association for Computing Machinery, New York, NY, USA, 153–156. https://doi.org/10.1145/1297231.1297259

[65] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*. IEEE, Istanbul, Turkey, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[66] Nava Tintarev and Judith Masthoff. 2011. Designing and Evaluating Explanations for Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer US, Boston, MA, 479–510. https://doi.org/10.1007/978-0-387-85820-3_15

[67] Nava Tintarev and Judith Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4 (Oct. 2012), 399–439. https://doi.org/10.1007/s11257-011-9117-5

[68] Chun-Hua Tsai and Peter Brusilovsky. 2019. Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. Association for Computing Machinery, New York, NY, USA, 22–30.

[69] Chun-Hua Tsai and Peter Brusilovsky. 2019. Explaining Recommendations in an Interactive Hybrid Social Recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 391–396. https://doi.org/10.1145/3301275.3302318

[70] Raphael Vallat. 2018. Pingouin: Statistics in Python. *Journal of Open Source Software* 3, 31 (Nov. 2018), 1026. https://doi.org/10.21105/joss.01026

[71] Alfredo Vellido. 2020. The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care. *Neural Computing and Applications* 32, 24 (Dec. 2020), 18069–18083. https://doi.org/10.1007/s00521-019-04051-w

[72] Katrien Verbert, Nikos Manouselis, Xavier Ochoa, Martin Wolpers, Hendrik Drachsler, Ivana Bosnic, and Erik Duval. 2012. Context-Aware Recommender Systems for Learning: A Survey and Future Challenges. *IEEE Transactions on Learning Technologies* 5, 4 (Oct. 2012), 318–335. https://doi.org/10.1109/TLT.2012.11

[73] Giulio Vidotto, Davide Massidda, Stefano Noventa, and Marco Vicentini. 2012. Trusting Beliefs: A Functional Measurement Study. *Psicologica: International Journal of Methodology and Experimental Psychology* 33, 3 (2012), 575–590.

[74] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15.

[75] Y. Diana Wang. 2014. Building Trust in E-Learning. *ATHENS JOURNAL OF EDUCATION* 1, 1 (Jan. 2014), 9–18. https://doi.org/10.30958/aje.1-1-1

[76] Kelly Wauters, Piet Desmet, and Wim Van Den Noortgate. 2012. Item Difficulty Estimation: An Auspicious Collaboration between Data and Judgment. *Computers & Education* 58, 4 (May 2012), 1183–1193. https://doi.org/10.1016/j.compedu.2011.11.020

[77] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101. https://doi.org/10.1561/1500000066

A POST-STUDY QUESTIONNAIRE

Table 2: The questionnaire that participants answered at the end of the study. All questions were evaluated on a 7-point range. The group names in italics are for reference; participants did not see them. After each group, participants could motivate their answers and give additional comments in a text field.

No.	English original	Dutch translation
Competence		
Q1	Wiski is like an expert (for example, a teacher) for recommending math exercises.	Wiski is zoals een expert (bv. een leerkracht) in wiskunde-oefeningen aanraden.
Q2	Wiski has the expertise (knowledge) to estimate my math level.	Wiski heeft de expertise (kennis) om mijn wiskundeniveau te kunnen inschatten.
Q3	Wiski can estimate my math level.	Wiski kan mijn wiskundeniveau inschatten.
Q4	Wiski understands the difficulty level of math exercises well.	Wiski begrijpt de moeilijkheidsgraad van wiskunde-oefeningen goed.
Q5	Wiski takes my math level into account when recommending exercises.	Wiski houdt rekening met mijn wiskundeniveau om oefeningen aan te raden.
Benevolence		
Q6	Wiski prioritizes that I improve in math.	Wiski zet op de eerste plaats dat ik vorderingen maak in wiskunde.
Q7	Wiski recommends exercises so that I improve in math.	Wanneer Wiski oefeningen aanraadt, doet Wiski dat zodat ik vorderingen maak in wiskunde.
Q8	Wiski wants to estimate my math level well.	Wiski wil mijn wiskundeniveau goed inschatten.
Integrity		
Q9	Wiski recommends exercises as correctly as possible.	Wiski raadt oefeningen op een zo correct mogelijke manier aan.
Q10	Wiski is honest.	Wiski is eerlijk.
Q11	Wiski makes integrous recommendations.	Wiski maakt oprechte aanbevelingen.
Trust (one-dimensional)		
Q12	I trust Wiski to recommend me math exercises.	Ik vertrouw Wiski om mij wiskunde-oefeningen aan te raden.
Intention to return		
Q13	If I want to solve math exercises again, I will choose Wiski.	Als ik nog eens online wiskunde-oefeningen maak, dan kies ik voor Wiski.
Q14	If I want to be recommended math exercises again, I will choose Wiski.	Als ik nog eens wiskunde-oefeningen aangeraden wil krijgen, dan kies ik voor Wiski.
Perceived transparency		
Q15	I find that Wiski gives enough explanation as to why an exercise has been recommended.	Ik vind dat Wiski genoeg uitleg geeft over waarom een oefening aangeraden is.
General questions		
Q16	I do NOT want any explanations about why an exercise has been recommended when I use Wiski.	Wanneer ik Wiski gebruik, wil ik GEEN uitleg over waarom een oefening wordt aangeraden.
Q17	I find an explanation for why an exercise is recommended more important than for why a movie is recommended.	Ik vind uitleg krijgen over waarom een oefening wordt aangeraden belangrijker dan waarom een film wordt aangeraden.
Q18	I am NOT happy with the level of math exercises Wiski recommended.	Ik ben NIET blij met het niveau van de oefeningen die Wiski aanraadde.
Q19	I find it important to receive explanations when something (exercise/movie/product/...) has been recommended.	In het algemeen vind ik het belangrijk om uitleg te krijgen wanneer iets (oefening/film/product/...) wordt aangeraden.