Optimising Explanation Content of AI-Driven Hints for Transparency and User Satisfaction

JONATHAN MAXWELL GILMOUR*, University of British Columbia, Canada

This work details and evaluates a modified explanation interface for AI-driven Intelligent Tutoring System (ITS) hints. These modifications were designed to improve subjective user experience and encourage the user to explore more of the interface. To evaluate these modifications, an informal user study was conducted in which modified AI-driven hint explanations were presented to participants. The results were compared to a previous study using the same ITS. While participants spent more time exploring the explanation interface, qualitative feedback was generally more negative compared to previous explanation iterations. These results motivate and inspire further research into how the contextual element of XAI impacts user perception of explanations, contributing to a greater goal of AI transparency that is accessible to all.

1 INTRODUCTION

Explainable AI (XAI) is based around creating explanations for artificial intelligence systems that explain their inner workings, motivations, and outputs. Existing research has shown that, in some cases, XAI not only improves user performance for the given task [13], but also improves user acceptance of the AI assistance as a whole [6]. That said, explanations are not universally beneficial. Depending on the context [5] and content [2] of the explanations, they can sometimes be perceived as undesirable or intrusive. Based on these findings, it is crucial that XAI explanations are designed in a way that actually benefits end users.

To contribute to the goal of optimised XAI explanations, this work builds upon the 2021 research of Conati *et al.* [3], in which an Intelligent Tutoring System (ITS) provides AI-driven hints to assist users' learning of constraint satisfaction problems (CSPs). This ITS is called the ACSP (Adaptive CSP) applet. In this study, the experimental group's hints were bundled with explanations. Compared to the control group, users who were provided with hint explanations rated the hints as more helpful, more trustworthy, and reported a higher intention to use the ITS in the future. That said, certain variations between users were correlated with different perceptions of hint explanations. Notably, users with low Reading Proficiency (defined here as English vocabulary capabilities [11]), reported higher feelings of confusion when presented with explanations compared to being presented none at all. Additionally, users did not typically explore the entire explanation interface, typically closing the explanations after reading the first few pages. These findings motivate further modification and optimisation of the ACSP applet explanation interface.

In this paper, a version of the ACSP applet explanations, modified to assist user perception and attention, is presented and evaluated. This work contributes to the growing body of knowledge on XAI by showing how certain XAI content and contexts can impact user focus and satisfaction, informing future design choices and motivating further research.

2 RELATED WORK

Existing XAI research has produced promising results in both user performance and subjective user experience. Porayska-Pomsta and Chryssafidou found that adolescents performed better in real job interviews when the AI-based job interview tutoring program explained how it was evaluating the participants as they used the program [13]. Herlocker *et al.* evaluated the effects of different aspects of AI explanations on subjective user experience, highlighting specific design choices that make XAI more accepted, trusted, and understood by end users [6]. Larasati *et al.* evaluated different types of XAI for medical systems. They found that non-expert participants' trust

*This project was supervised by Dr. Cristina Conati of the University of British Columbia.

Author's address: Jonathan Maxwell Gilmour, maxgmr@student.ubc.ca, University of British Columbia, Vancouver, British Columbia, Canada.

levels can be moderated by explanations [9], supporting the idea that XAI can even help explain complicated concepts to non-professionals. Research by Kulesza *et al.* found that users were able to correct the mistakes of XAI more effectively, showing that user understanding of AI is genuinely impacted by explanations [8].

That said, XAI is not universally helpful. There have been many cases in which explanations have hindered user performance, understanding, and trust. The explanations must be considered necessary and informative by users— an evaluation of interface recommender explanations by Bunt *et al.* found that 40% of users perceived the explanations as either common sense or simply unneeded [2]. Wang *et al.* found that POMDP-generated explanations of robot reasoning to human teammates only improved trust when the actionable conclusions were explicitly expressed [15]. Work by Ehrlich *et al.* emphasises the need for accurate AI-driven recommendations— they found that incorrect recommendations with explanations harmed performance more than incorrect recommendations without explanations [5].

The use of XAI in pedagogical applications serves as a strong base for applications in other contexts due to the field's particular focus on high explainability, understanding, and trust [4]. In addition to the previously-described interview tutor study by Porayska-Pomsta and Chryssafidou [13], research by Long and Aleven found that transparent learning models that dynamically display students' live evaluations result in better learning outcomes in student self-assessment [10].

Naturally, the work most related to this paper is the 2021 study by Conati *et al.* that investigates the effectiveness of XAI within the context of the ACSP applet [3]. This study utilises a modified version of this ACSP applet to determine the effectiveness of the explanation changes by comparing the results with the 2021 study.

Finally, the changes to the explanation interface used by Conati *et al.* were motivated by existing research. For instance, a study by Pieters and Wedel found that increased text size captures user attention more effectively in advertisements [12]. An experimental trial by Selzer found that technical paragraph readability was improved by the following factors: use of repeated words instead of pronoun substitution, word repetition instead of synonyms, and presence of a topic sentence [14]. Finally, research by Bever *et al.* shows that text spacing also plays an element in readability— in particular, spacing the text to visually isolate key phrases improves readability, especially in low-proficiency readers [1].

3 THE ACSP APPLET

The adaptive constraint satisfaction problem (ACSP) applet is a pre-existing Intelligent Tutoring System (ITS) that tutors users in the AC-3 algorithm [3]. It does this by allowing users to step through algorithm operation and see its inner workings as the applet works through a variety of different example problems.

3.1 Hint Generation

The ACSP applet assists users with AI-driven hints personalised to the user's actions. The mechanism is fully delineated in the 2021 Conati *et al.* paper [3], but a short summary is provided here. Hint generation is based on the Framework for User Modeling and Adaptation (FUMA) [7]. FUMA builds a user model in two steps: Behaviour Discovery and User Classification.

Behaviour Discovery began with the data of users who utilised the applet without any adaptive hints. Their actions while using the applet and their learning performance were recorded. A clustering algorithm grouped the users together by similar applet usage behaviours while separating the groups by learning gains. After that, association rule mining was used to extract rules that distinguished between the cluster of users with low learning gains (the low-learning group) and the cluster of users with high learning gains (the high-learning group). Rules were assigned weights based on how well they distinguished between the two groups.

The User Classification phase occurs actively as a given user utilises the ACSP applet. The applet monitors their actions, and based on the weights of the rules extracted from the Behaviour Discovery phase satisfied by

that user, the applet will classify them into the low-learning or the high-learning group. This User Classification drives the hints— if a user is classified as low-learning, the applet will generate a hint that is personalised to the particular behavioural patterns and rules satisfied by the current user (Fig. 1).

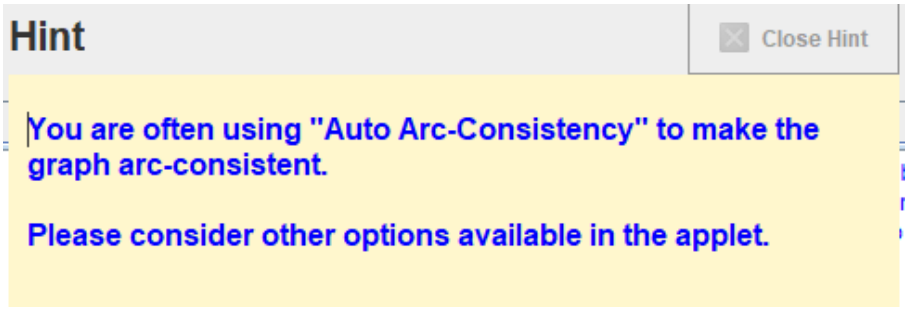

Fig. 1. An example hint. The content of the hint reflects the current assessment of the AI model— in this example, the user satisfies the "Frequently Auto-Solving the Problem" rule, causing them to be classified in the low-learning group. The hint suggests to the user how to modify their behaviour such that they will no longer satisfy that particular low-learning rule.

3.2 Explanation Access

Upon receiving a hint, users are given the option to press a button which opens up an explanation interface consisting of multiple pages. These explanation pages aim to explain the adaptive hints to the user. These pages are *WhyHint*, which details why the user was given the hint, *WhyLow*, which details why they were predicted to be lower learning, *HowScore*, which details how their score was computed, *HowHint*, which details why the rules were used for classification, *HowRank*, which details how their hint's rank was calculated, and *HowRules*, which details why these rules were used to classify them. (See Fig. 2 for an example explanation tab. All pages can be viewed in Appendix A.) They are partially personalised— some information is universal, such as the explanation of the Behaviour Discovery stage. Conversely, some information is specific to the user, such as a graph showing the user exactly which low-learning and high-learning rules their behaviour satisfies. Hence, the explanations themselves are inextricable from the AI system; they should be considered a part of the AI user model as a whole.

4 MODIFICATIONS TO THE ACSP EXPLANATIONS

The changes made to the ACSP explanations were designed to capture user attention and make said explanations feel less overwhelming, distracting, and more trustworthy. One major change was increase of text size. Previous research in advertising has found that increasing text size improves capture of attention [12], so this was a top priority.

Increase of text size necessitated reduction of word count. Most explanation text was shortened. This shortening had to be done whilst considering soundness and completeness— to remain a truly transparent AI system, no information could be lost and no information could be incorrect.

This word count reduction also allowed for the splitting of text into smaller blocks, sometimes separated by titles. Previous research found that visually isolating important text components through spacing helped with readability, especially for low-reading proficiency users [1]. The intention behind the use of titles was to provide not only a visual separation of text, but also the ability for readers to quickly see the content of the paragraph at a glance.

First-person pronouns were used in the 2021 study to emphasise the fact that the AI model was communicating details about itself to the user. However, motivated by research findings that substituting repeated phrases for

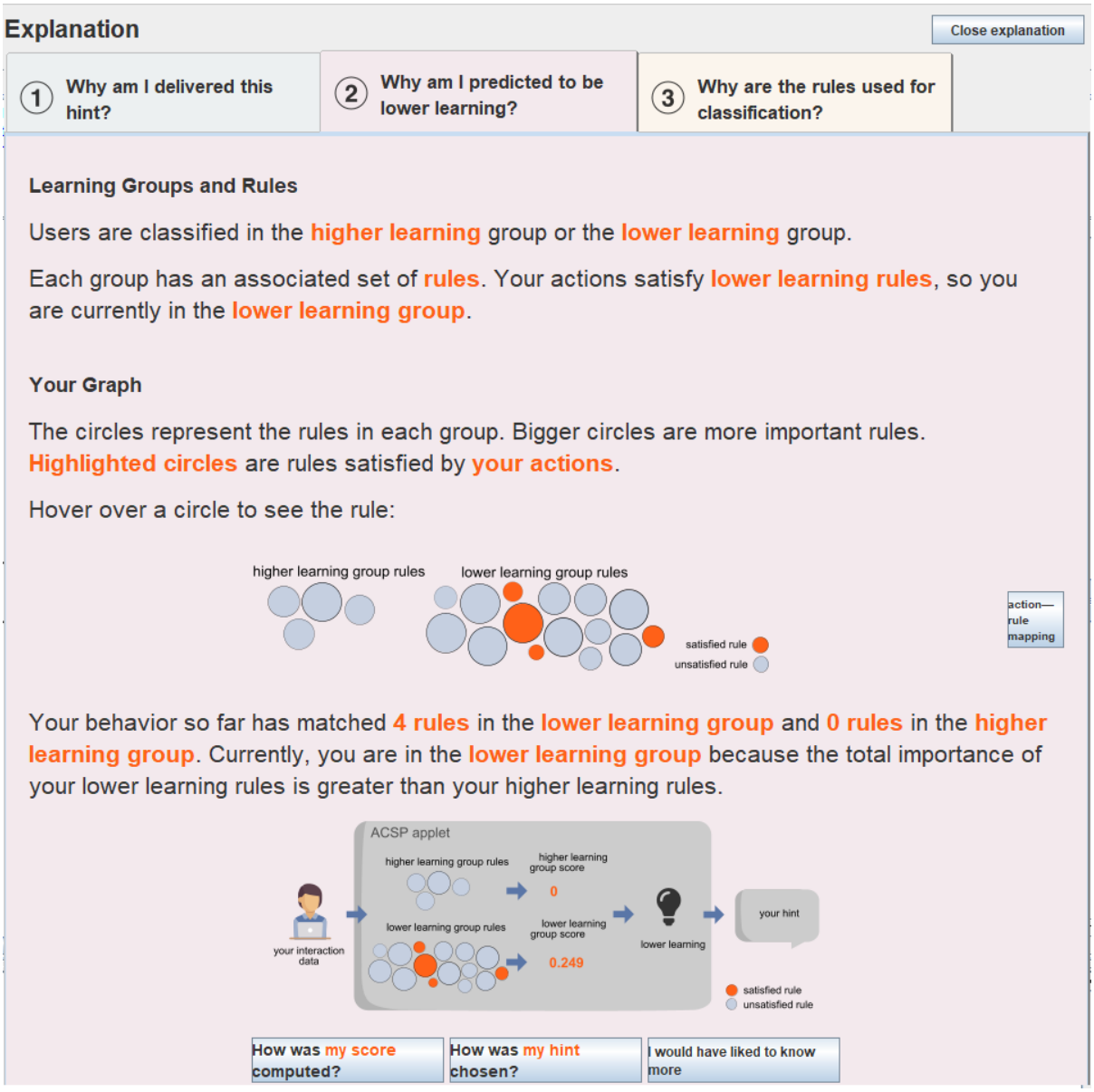

Fig. 2. An example *WhyLow* explanation page. While most of the text is global information always shown to the user, the graphs provide local information specific to this particular user in the form of highlighted satisfied rules and the calculated low-learning score. Some text is also dependent the AI user model— for example, the paragraph detailing how many rules in each group the user satisfies.

pronouns negatively affects technical paragraph readability [14], first-person pronouns were replaced with either passive voice or "this applet", "this model", "this hint", *etc.* (Fig. 3). Additionally, the use of passive voice in

particular was partially done to imitate the industry standard for explanations (*e.g.* Fig. 4), further signalling to the user that the content they are reading is an explanation designed to help them understand the product they are using.

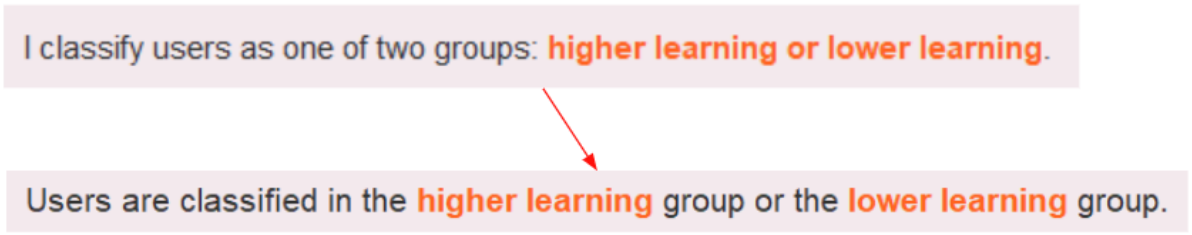

Fig. 3. An example replacement of first-person pronouns with passive voice.

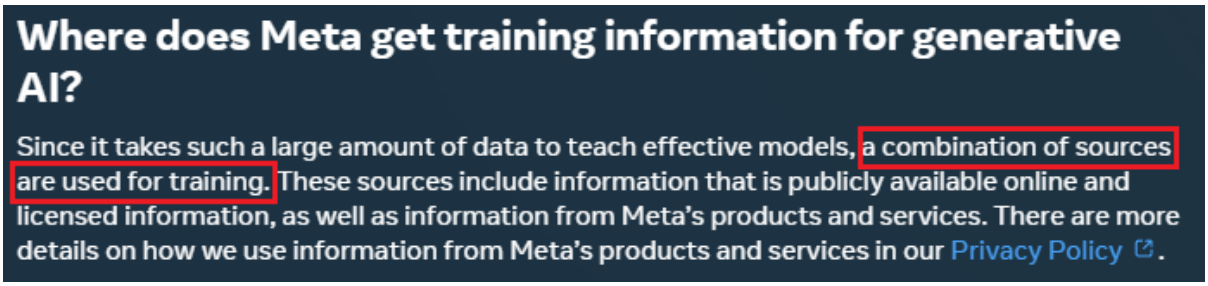

Fig. 4. An example of passive voice as an industry standard used in AI transparency pages. Screenshot taken April 2024 from the "Generative AI" page on Meta Inc.'s Generative AI Privacy Center page, highlight added by author (https://www.facebook.com/privacy/genai).

Finally, the technical language and jargon was reduced as much as possible without harming accuracy (Fig. 5). Technical words were substituted for equivalent words more well-known by the general population.

Fig. 6 provides a comparison of example pages from the old explanation interface and the new explanation interface. The complete modified ACSP explanation interface can be viewed in Appendix A.

5 USER STUDY

5.1 Participants

To evaluate the effectiveness of the explanation changes, a between-subject user study was conducted. The scale and scope of the project limited the available participant pool. 5 participants were recruited from the author's personal acquaintances. All participants already knew the AC-3 algorithm demonstrated by the ACSP applet but had no previous knowledge of the AI-driven hints or the explanations. The personal relationships to the author and their pre-existing algorithm knowledge were the only participant differences from the 2021 Conati *et al.* study against which this paper's results are compared.

5.2 Procedure

To compare the results against the 2021 Conati *et al.* study, a slightly modified version of that study's procedure was used. Their procedure, fully detailed in their paper [3], contained additional steps which measured for

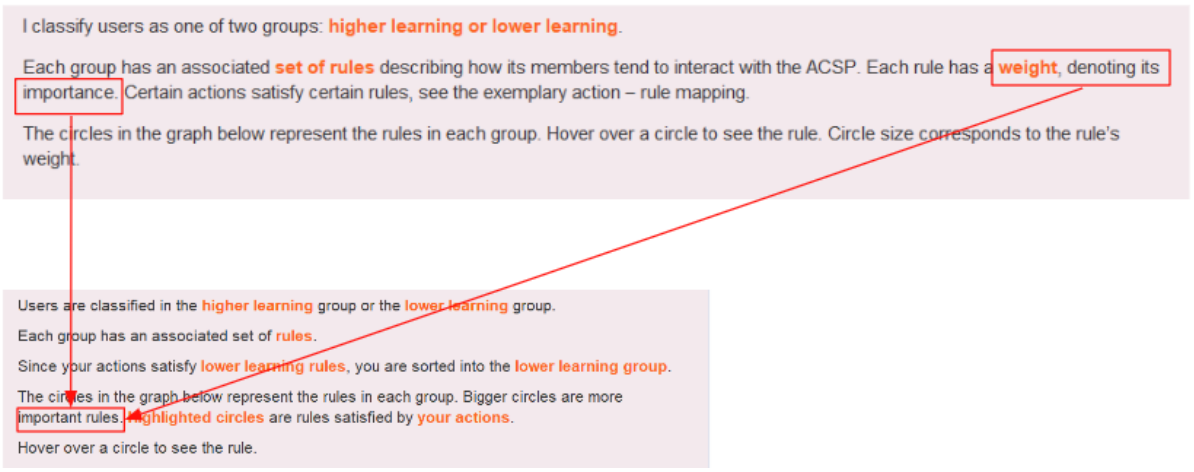

Fig. 5. An example of jargon substitution. While "weight" is the more commonly-used term amongst computer scientists, and may initially seem to be the clearest term here, the word "importance" has a contextually identical meaning and is more likely to be understood by the general population.

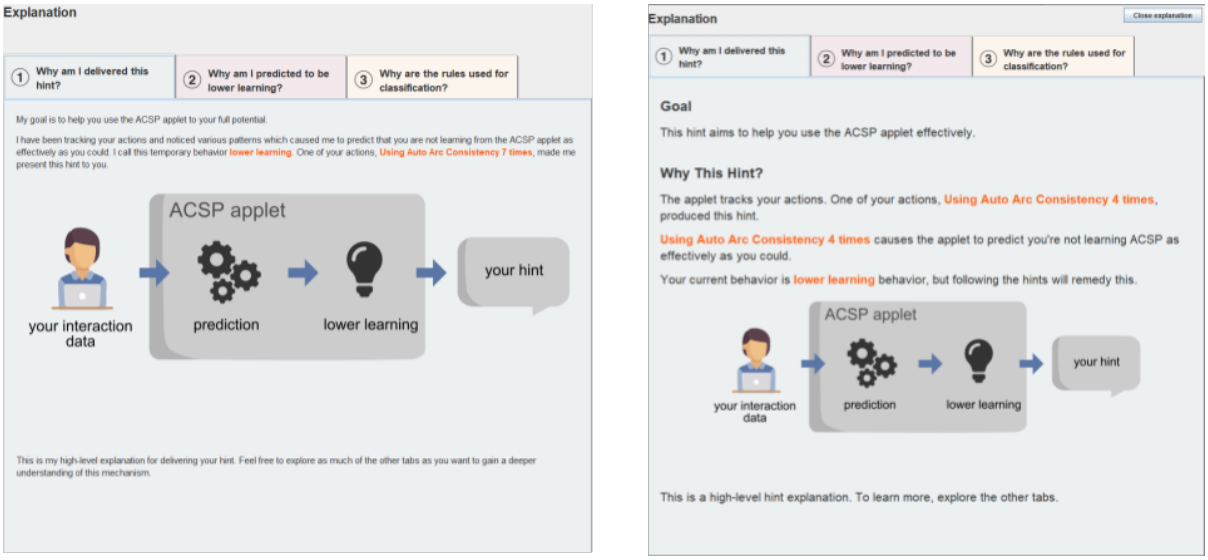

Fig. 6. A comparison of the old *WhyHint* page (left) with the new *WhyHint* page (right); the side-by-side comparison exemplifies the effect of increasing text size, reducing text, adding headers, and spacing out paragraphs.

learning gains and certain user characteristics that were not part of the scope of this project. The only further deviation from the Conati *et al.* experimental procedure was the addition of an open-ended oral interview. The procedure steps were as follows:

(1) To measure Reading Proficiency, participants take the X_Lex Vocabulary Test, which defines Reading Proficiency as English vocabulary size and reading comprehension ability [11].

(2) Participants read a textbook chapter on the AC-3 algorithm tutored by the ACSP applet. Although all participants had previous knowledge of AC-3, this step was preserved to ensure this knowledge was refreshed.
(3) Participants are shown a video detailing how to use the ACSP applet interface. This video does not mention the ACSP hints or explanations.
(4) Participants use the applet to solve three example problems. The ACSP applet monitors and logs all their actions. Focus is recorded by an eye tracker. Participants had to use the applet for a minimum of 10 minutes— if they finished before that time, they were instructed to simply explore the applet interface until the 10 minutes had passed.
(5) Participants fill out a questionnaire, identical to the 2021 study [3], which was designed to measure qualitative user perception of the ACSP applet hints and explanations. Participants rate, from 1 to 5, how strongly they agree with each of the statements. The questions related to the hints are shown in Table 1. This data was unused, but was collected regardless to mitigate the influence of participants' perception of the hints on the questions evaluating the explanations themselves. Table 2 shows the list of questions related to the explanations. These questions measure a comprehensive range of sentiments users could feel from the hints. Participants who do not receive any hints do not receive any questions.
(6) After debriefing the participant and informing them that the study is evaluating the explanation interface, a recorded oral interview is conducted. Due to the small sample size, this study placed more focus on each individual's subjective experience. The oral interview enabled collection of sentiments and feedback that participants were unable to express in the questionnaire due to the lack of text-based questions. Participants who did not receive any hints were debriefed but did not answer any interview questions. The oral interview questions are detailed in Table 3.

Table 1. Questionnaire Items - Perception of Hints

Question Label	Category	Question Text
H1	Usefulness	I would choose to have the hints again in the future.
H2	Usefulness	I am satisfied with the hints.
H3	Usefulness	The hints were helpful for me.
H4	Intrusiveness	The hints distracted me from my learning task.
H5	Intrusiveness	The hints were confusing.
H6	Understanding and Trust	I understand why hints were delivered to me in general.
H7	Understanding and Trust	I understand why specific hints were delivered to me.
H8	Understanding and Trust	I trust the system to deliver appropriate hints.
H9	Understanding and Trust	Given my behaviour, I agree with the hints that were delivered to me.

6 QUANTITATIVE RESULTS

The following data was collected from the 4 participants who received hints (out of 5 total). Any interpretations or comparisons must be done acknowledging that the scale of this study prevents any results from being

Table 2. Questionnaire Items - Perception of Explanations

Question Label	Category	Measured Sentiment	Question Text
E1	Usefulness	Intention to Use	I would choose to have the explanations again in the future.
E2	Usefulness	Satisfaction	I am satisfied with the explanations.
E3	Usefulness	Helpful	The explanations were helpful to me.
E4	Intrusiveness	Distracting	The explanations distracted me from my learning task.
E5	Intrusiveness	Confusion	The explanations were confusing.
E6	Intrusiveness	Overwhelming	I found the explanations overwhelming.
E7	Usability	Access	It was clear to me how to access the explanations.
E8	Usability	Navigation	The explanation navigation was clear to me.
E9	Usability	Content	The explanation content (i.e., wording, text, figures) was clear to me.

Table 3. Oral Interview Questions

Question
I. When you opened the explanations, what was your first impression? How did you feel?
II. Is there anything you'd personally change about how the explanations were delivered?
III. Did the explanations make you trust the hints more? Why or why not?
IV. Would you feel more comfortable using artificial intelligence if its functionality was explained in such a manner? Why or why not?
V. Given the option to enable similar hints and explanations in a real-life tutoring program you were using, would you choose to enable them? Why or why not?
VI. Was there anything in particular you liked about the explanations?
VII. Was there anything in particular you disliked about the explanations?
VIII. Are there any other comments you have about the hints, the explanations, or the applet as a whole?

statistically significant. The numerical results and conclusions (detailed in later sections) are presented in the context of a small-scale, informal user study; they are used to inform further design choices and research rather than make definitive claims.

The following metrics were collected and compared with the identical metrics from the 2021 Conati *et al.* study:

- *Hints before first explanation initiation*: The amount of hints participants received before they opened an explanation.

- *Explanation initiations per hint received*: The proportion of hint pop-ups in which participants opened the explanations.
- *Number of page accesses per initiation*: The number of different explanation pages accessed by a participant after opening an explanation.
- *Attention to explanations per hint received*: The time participants spent focusing on the explanation content after receiving a hint. Not opening the explanation at all would equal a result of 0 seconds. Unfortunately, the eye tracker lost the gaze of 2 out of the 4 participants while they were reading the explanations, meaning that this metric could not be collected.
- *Explanation types accessed*: There are 6 different pages in the explanation interface. Each participant received a score of 0-6 based on how many different explanation pages they viewed at least once.

The results, shown next to the results of the 2021 study [3], can be seen in Table 4.

Table 4. Comparison of Quantitative Results to Conati *et al.* (2021)

Metric	Original Explanations (2021) Mean [3]	Modified Explanations (2024) Mean
Hints before first explanation initiation	1.08	1.25
Explanation initiations per hint received	0.75	0.75
Number of page accesses per initiation	3.11	3.38
Attention to explanations per hint received	38.5s	N/A
Explanation types accessed	2.79	3.75

The average 1.25 hints before first explanation initiation was caused by 1 out of the 4 participants opening the explanation only on their second hint.

The 0.75 explanations per hint received was identical to the 2021 study.

On average, participants accessed more pages when reading the explanations compared to the 2021 study.

As stated above, no results could be collected for explanation attention, but it is worth noting that, with the modified explanations, participants spent an average of 81.5s with the explanation window open. Assuming even half of that time was dedicated to focusing on the explanation contents (a reasonable assumption, considering the explanations are the only accessible part of the applet when the explanation window is open), participants likely spent more time reading the explanations. However, this cannot be known based off the data successfully collected in this study.

On average, users opened a more diverse range of pages than users in the 2021 study. One user viewed all 6 explanation pages— something not seen across the 24 participants who received explanations in the 2021 study. Matching the 2021 study, the *Why* pages (*WhyHint, WhyLow, WhyRules*) were more accessed than the *How* pages (*HowScore, HowHint, HowRank*). A visual comparison is provided in Fig. 7.

Finally, no analysis was done on the interaction between Reading Proficiency and confusion. This was because all subjects had very high Reading Proficiency scores with very little variation (mean=94.07, SD=6.12 compared to the 2021 study's mean of 80.32 [3]), meaning that there were no low-Reading Proficiency users to analyse.

7 QUALITATIVE RESULTS

7.1 Questionnaire Results and Comparison

Questionnaire results (questions E1-E9; refer back to Table 2 to review the measured categories and sentiments), designed to express qualitative user perception of the ACSP explanations, were mostly contrary to expectations. A visual comparison of user questionnaire responses can be seen in Fig. 8.

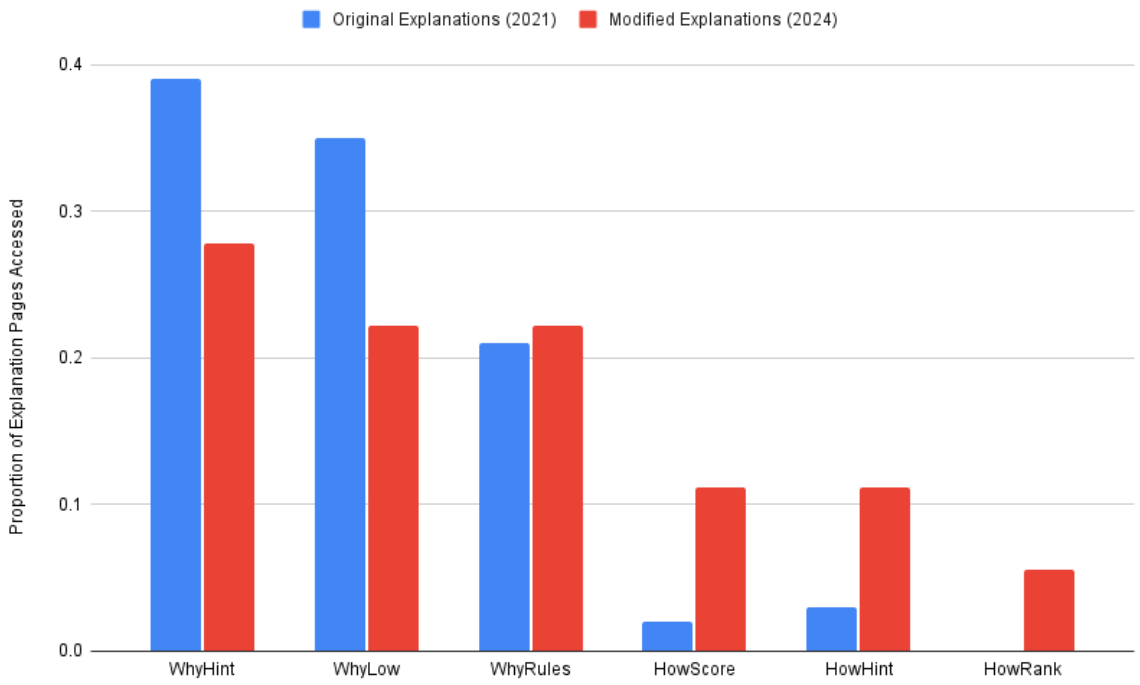

Fig. 7. A visual comparison of the proportion of pages accessed of each type. The larger bars for *WhyHint* and *WhyLow* for the original explanations, and smaller bars for the others, show that participants accessed a greater diversity of pages under the modified explanations.

The comparatively negatively-expressed sentiments were as follows: lower Intention to Use, lower Satisfaction, higher Distracting, higher Overwhelming, lower Navigation, and lower Content (compared to the 2021 explanations).

The comparatively positively-expressed sentiments were as follows: higher Helpful and lower Confusion (compared to the 2021 explanations).

In general, the modified explanations produced more negative sentiments. Particularly, sentiments in the Usefulness category were more negative overall. The largest differences were much higher ratings of Distraction and much lower ratings of Confusion.

As stated in the Quantitative Results section, this study's sample size is too small for formal statistical analysis, but the questionnaire results provide interesting insight into how the different participant groups perceived the hint explanations.

7.2 Oral Interview Results

The following results represent sentiments that were common among multiple participants or otherwise notable.

I. When you opened the explanations, what was your first impression? How did you feel?

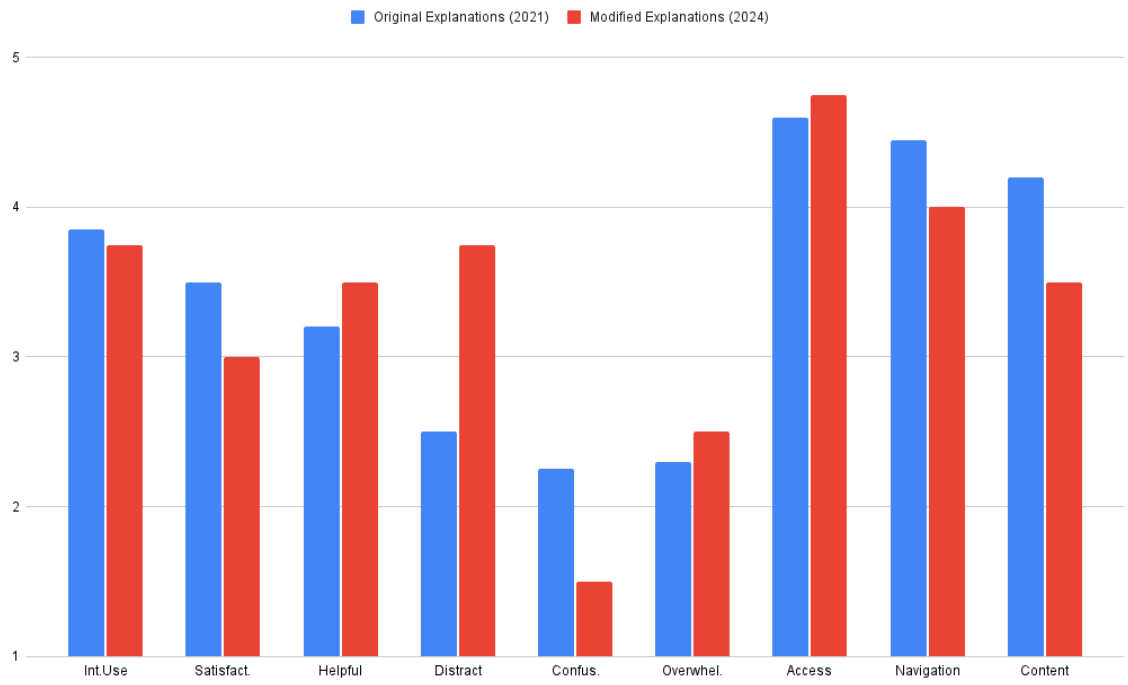

Fig. 8. A visual comparison of user questionnaire responses. Of particular note are the results contrary to the hypothesised user response— lower Intention to Use, lower Satisfaction, higher Distraction, higher Overwhelming, and lower Content. However, the lower Confusion aligns with expectations.

"[The explanations] were well worded, I understood it, but once I had done [the first example problem] once, I understood how it worked and just used Auto-Step [low learning behaviour]."

"I was annoyed with the message I got, that I wasn't learning enough because I was doing this basic action too many times, even though I was only doing it at the end after I already figured everything out."

Exemplified by the previous two quotes, the sentiment shared by every participant was that they felt their hints were unwarranted; for example, one participant reported feeling frustrated that they got a hint because they used the Reset button once.

II. Is there anything you'd personally change about how the explanations were delivered?

"The data was hard to digest and didn't offer much insight[...] I didn't read it too much"

"Compared to the rest of the applet, where information is pretty short and sparse, going to the explanation where as soon as you clicked it, it was like a few paragraphs of explanation immediately, my eyes started glazing over."

Most participants still felt overwhelmed by the amount of information even after the text was reduced. The volume of text and information remains a reason why people are not exploring the interface more.

III. Did the explanations make you trust the hints more? Why or why not?

Participants generally stated that the explanations make sense, but they felt that the things they were doing were not bad for their learning. "I still didn't understand why resetting was a bad thing." Interestingly, the explanations reduced trust in these participants because they felt like the hints were unwarranted. After reading why they got the hint in the explanation, they thought this rationale was unreasonable, making them question the hint delivery system even though they were not questioning it before.

Additionally, two participants pointed out that the applet was only transparent upon receiving a hint: "When seeing the hints for the first time, you're not even aware you're being evaluated."

IV. Would you feel more comfortable using artificial intelligence if its functionality was explained in such a manner? Why or why not?

Responses to this question were lukewarm. No participants were opposed to the addition, but they either had problems with the implementation itself or simply stated they would not use it.

V. Given the option to enable similar hints and explanations in a real-life tutoring program you were using, would you choose to enable them? Why or why not?

Responses to this question were too varied to draw out any meaningful shared sentiments. Some participants would enable hints while others would not.

VI. Was there anything in particular you liked about the explanations?

Interestingly, several participants reported liking that explanations were "short and to-the-point", contrasting with previous sentiments of feeling overwhelmed. Additionally, several participants liked how in-depth the explanations were— they reported feelings of agency and control over how much they could see.

VII. Was there anything in particular you disliked about the explanations?

All sentiments expressed by responses to this question were covered in responses to the previous questions.

VIII. Are there any other comments you have about the hints, the explanations, or the applet as a whole?

Once again, most of the sentiments expressed here were previously expressed in other responses. However, one participant was interested in the nature of the AI driving the hints. They were surprised to learn it was a trained user model— a fact expressed in the explanations that, as shown by their surprise, did not reach the intended audience. This means that the explanations' description of how the AI model got its rules (*WhyRules*, see Fig. 14) were unclear even to users who were personally interested in seeking out those facts.

8 DISCUSSION

As shown by the data, this new experimental iteration produced mixed results. While participants generally explored the interface more, their reported subjective experiences were less positive compared to the original 2021 study's interface [3]. These unexpected results, which run contrary to the original hypothesis that the explanation modifications would produce a better subjective experience, inspire reflection and theorising on how such results could have occurred.

The unexpected results could have arisen because of differences between this study and the 2021 study. There were three deviations from the 2021 Conati *et al.* study against which the data was compared. All three must be considered as potential confounds. They are as follows: first, several steps were removed, and one added, to the 2021 procedure; second, participants were recruited as personal acquaintances, while the 2021 participants were recruited from a general subject pool and compensated for their time; third, all participants had previous

experience with the AC-3 algorithm tutored by the ACSP applet, while the 2021 participants were required to have no AC-3 experience.

8.1 Changes to Procedure

It is unlikely the modifications to the procedure had any impact on the results. Regarding the user characteristic tests dropped from the procedure, the 2021 study analysed the effects of various user characteristics on participant responses to the explanations. Therefore, participants took not only the test for Reading Comprehension, but also tests on 10 other user characteristics [3]. This study did not measure any data related to these tests. It was entirely focused on the response to the explanations themselves, so there is no known way the absence of these tests could have affected results.

Additionally, the steps in which participants take a pre-test and a post-test on the tutored algorithm were removed because learning gains were not a relevant metric to the goals of this work. It is possible that taking these tests primed participants' mental states for the algorithm, but reading the textbook chapter would have done the same thing.

Finally, the addition of the oral interview could not have impacted questionnaire responses or interface exploration because it took place after the applet usage and questionnaire steps. Hence, it is unlikely any procedural changes impacted results.

8.2 Personal Relationships With Participants

It is unlikely that participants interacted with the interface more extensively because of a previously-existing personal acquaintanceship with the author. While the desire to spend more time fully exploring the explanation interface may have been motivated by this acquaintanceship, three pieces of evidence taken together make this unlikely. First, participants took about as much time on the other experimental procedure steps as previous studies, meaning they did not spend significantly longer amounts of time on other actions. Second, participants did not know that the explanations were the evaluated portion of the study, so they would not have known that exploring more of the interface would have produced positive results. Third, considering that the questionnaire responses were generally more negative than previous studies, it seems unlikely that participants would artificially inflate other statistics.

Additionally, there is the possibility that the participants' personal relationship with the author made them feel more comfortable rating the interface more harshly in the questionnaire, thus bringing the results down compared to the 2021 study. This theory is hampered by the fact that all participants were clearly informed at multiple points during the study, including right before the questionnaire, that all their results are anonymous, even to the author. This was done to simulate the relative anonymity the 2021 participants had. Furthermore, not all questionnaire metrics were rated more negatively. It would be unreasonable to claim without good reason that a personal relationship influenced participants to rate some metrics higher than they should whilst rating other metrics lower than they should.

That said, even taking these points into consideration, the participants' personal acquaintanceship with the author remains a deviation from the compared study's design, so any confounding effects cannot be ruled out with absolute certainty.

8.3 Previous Algorithm Experience

The potential confound of the subjects' previous knowledge of the AC-3 algorithm was known before the experiment was conducted. This concession was made out of necessity due to the fact that this project was a smaller scale than the 2021 study. At the time, this deviation was justified by the fact that the focus of the study was entirely on the explanations rather than any learning gains or hint feedback. Previous AC-3 knowledge was

thought to be irrelevant with respect to explanation perception because no participants in this study had ever been previously exposed to the explanation or hint interfaces. None were aware of their existence in the first place.

However, the theory that previous AC-3 experience affected results holds some weight. While further research would be needed to make any concrete claim, this is supported by the unexpectedly lower questionnaire ratings for certain metrics. As shown above, compared to participants in the 2021 study, users in this study rated the explanations with lower Satisfaction, higher Distraction, lower Intention to Use, higher Overwhelming, and lower Content. The most likely interpretation of these results is that, due to their pre-existing knowledge, this study's participants received the explanations in an entirely different context. The ACSP applet's user model was trained on users learning AC-3 for the first time. Users with previous knowledge in AC-3 would likely feel more confident and be more comfortable auto-solving the problem, stepping through the problem quickly, or performing other behaviours that would classify them as low learning. Previously-discussed research also supports this, such as the Bunt *et al.* paper that found a great deal of participants who reacted in similar ways to their explanations [2].

This theory is further supported by the oral interview responses. As previously stated, the response that was universally expressed was a feeling that the hints were unwarranted. One user reported seeing the hint and assuming it was trustworthy. Then, upon opening the hint and seeing that hint was delivered because they reset the problem a single time, they trusted the hint less. Two users reported feeling annoyed when they got a hint because they already understood the algorithm and were just auto-solving the problem for convenience. The fact that, compared to the 2021 study, Distraction was the largest increase, supports this theory: it would be only natural for users to feel interrupted by explanation pop-ups if they were just trying to finish a simulation of a problem they already understood. It is reasonable to imagine that these participants had already mentally considered the task completed, leading them to feel frustrated and dissatisfied by any further interruptions, especially because these hints were so unexpected and not telegraphed to the user beforehand.

This theory also aligns with the more positive questionnaire results. The most significant change from the 2021 study was the rating of Confusion, which was far lower than the 2021 study's mean rating. It is possible that the lower rating of Confusion was caused by the modified explanation interface, and it is very likely these modifications played some role in the rating— the ratings were just for the explanations, not the rest of the applet. The explanations were completely new for users both in this study and the 2021 study. That said, it is possible that since the users in the study already knew AC-3, they were in a less confused state in general, leading them to feel to rate the explanations as less confusing. This would align with previous observations; there appears to be a general pattern of the context in which the users receive the hints potentially having a large effect on their responses. While the questionnaire tried to mitigate any sentiments about the hint delivery bleeding over into the ratings of the explanations, it is reasonable to assume that the context colours all further experiences with that interface.

The theory that users' previous AC-3 knowledge caused the unexpectedly negative subjective response to the modified explanations has a great deal of support from the above observations. However, it must be emphatically restated that this remains a theory— no formal conclusions can be drawn from these points.

8.4 Supportive Quantitative Results

The quantitative results, while still limited by a small sample size, provide modest support for the original hypothesis. The greater average diversity of page accesses, i.e., users exploring more pages of the interface than the 2021 participants, supports the hypothesis that the modified explanations reduce attentional fatigue and make users feel more comfortable exploring the full interface. The interface navigation buttons were entirely unchanged, so both groups of participants navigated the interface in identical ways. It was not easier nor harder to access the different interface pages. One of the common positive sentiments received in the oral interviews

was a general appreciation of the page-based delivery— while this was already in the explanation interface before this study's modifications, it is possible that the modified content made individual pages less intimidating and inspired a level of further exploration of other pages that was not seen in the previous study. The increased exploration aligns with previous research on the effects of these changes on attentional capture specifically [12].

9 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

While user perception did not align with expectations, user behaviour certainly did— the explanation modifications did indeed encourage further interface exploration compared to the 2021 study by Conati *et al.*. However, the main conclusion that can be drawn from this work is not a conclusive statement on the effectiveness of the modified explanations due to the small sample size and confounding factors detailed above. Rather, the primary contribution of this work is support for a theory: existing user knowledge, and context in general, significantly affects user perception of XAI enough to counteract any changes to the content of the explanations themselves.

Should this theory be true, the applicability to real-life XAI applications would be extensive. More emphasis would have to be placed on the accuracy of the user model, and the user model itself would have to take into account further metrics that potentially would be otherwise under-valued— for instance, AI-driven software tutorials could take into account what the user is actually trying to do at the given moment, in addition to gauging their knowledge of the program, when deciding whether to present them a hint. Interruptions by explanations would result in more frustration when the user is trying to do certain tasks or for highly-experienced users.

Naturally, the largest limitation of this study was the scope. A larger participant group more similar to the 2021 study would produce more directly comparable results and more convincing conclusions.

A limitation shared with the 2021 study is the intended audience of the explanations. Most software users are not well-educated university students with computer science backgrounds. The effectiveness of different XAI explanations must be analysed in a wide variety of applications and audiences. Certain design choices that work well for computer science-related Intelligent Tutoring Systems may not work well, for instance, within an XAI-assisted social media app for the general public, or even an Intelligent Tutoring System for high school students.

One final limitation was the lack of range in Reading Proficiency across participants. Findings from the 2021 study suggest that low-Reading Proficiency users do not receive the same benefits of XAI as the general population [3]. The interface changes in this study, such as reduced text and larger text size, should be tested on low-Reading Proficiency users to see if they can mitigate or eliminate the negative effects.

This work motivates research into further XAI interface improvements. The oral interview responses prove people still find the information intimidating and difficult to access. Users' ability to access different levels of detail through the different explanation pages was a highly-praised feature. A tiered information delivery system that allows users to access exactly the amount and complexity of explanatory content they want appears to be helpful. More work into further implementing this could further improve XAI readability. In particular, tooltips, i.e. the ability to hover over a technical word or phrase to see its definition, could reduce the amount of text the user is immediately presented with and give users further control over how much information they are exposed to. The extent of their efficacy should be measured.

Anecdotally, responses outside of the experimental context to the side-by-side comparisons of the old and new explanation interface (as seen in Appendix A) have been universally positive. The specific differences between the old and the new interfaces could be directly measured with a within-subject user study in which users are shown both explanation interfaces and asked to compare the two. However, caution must be exercised, because as this study's results suggest, the context in which a user sees a given interface can dramatically affect that user's perception of that interface. If the explanations are taken out of their actual context, it is unlikely the results would have any practical, real-world applications.

While the efficacy of the modified XAI explanations could not be measured to the intended extent, the findings of this work support and inspire future research into both the context and content of effective XAI.

ACKNOWLEDGMENTS

This work was supervised by Dr. Christina Conati at the University of British Columbia. Special thanks to Vedant Bahel from the University of British Columbia.

REFERENCES

[1] Thomas Bever, R. Burwell, S. Jandreau, Ronald Kaplan, and Annie Zaenen. 1991. Spacing printed text to isolate major phrases improves readability. *Visible Language* 25 (01 1991).

[2] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (Lisbon, Portugal) *(IUI '12)*. Association for Computing Machinery, New York, NY, USA, 169–178. https://doi.org/10.1145/2166966.2166996

[3] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (2021), 103503. https://doi.org/10.1016/j.artint.2021.103503

[4] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. arXiv:1807.00154 [cs.AI]

[5] Kate Ehrlich, Susanna E. Kirk, John Patterson, Jamie C. Rasmussen, Steven I. Ross, and Daniel M. Gruen. 2011. Taking advice from intelligent systems: the double-edged sword of explanations. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (Palo Alto, CA, USA) *(IUI '11)*. Association for Computing Machinery, New York, NY, USA, 125–134. https://doi.org/10.1145/1943403.1943424

[6] 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* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 241–250. https://doi.org/10.1145/358916.358995

[7] Samad Kardan. 2017. *A data mining approach for adding adaptive interventions to exploratory learning environments.* Ph. D. Dissertation. University of British Columbia. https://doi.org/10.14288/1.0348694

[8] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[9] Retno Larasati, Anna De Liddo, and Enrico Motta. 2023. Meaningful Explanation Effect on User's Trust in an AI Medical System: Designing Explanations for Non-Expert Users. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 30 (dec 2023), 39 pages. https://doi.org/10.1145/3631614

[10] Yanjin Long and Vincent Aleven. 2017. Enhancing learning outcomes through self-regulated learning support with an Open Learner Model. *User Modeling and User-Adapted Interaction* 27, 1 (01 Mar 2017), 55–88. https://doi.org/10.1007/s11257-016-9186-6

[11] Paul Meara. 1991. EFL vocabulary tests. https://eric.ed.gov/?id=ED362046

[12] Rik Pieters and Michel Wedel. 2004. Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects. *Journal of Marketing* 68, 2 (2004), 36–50. https://doi.org/10.1509/jmkg.68.2.36.27794 arXiv:https://doi.org/10.1509/jmkg.68.2.36.27794

[13] Kaśka Porayska-Pomsta and Evi Chryssafidou. 2018. Adolescents' Self-regulation During Job Interviews Through an AI Coaching Environment. In *Artificial Intelligence in Education*, Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay (Eds.). Springer International Publishing, Cham, 281–285.

[14] Jack Selzer. 1982. Certain Cohesion Elements and the Readability of Technical Paragraphs. *Journal of Technical Writing and Communication* 12, 4 (1982), 285–300. https://doi.org/10.1177/004728168201200403 arXiv:https://doi.org/10.1177/004728168201200403

[15] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 997–1005.

A THE COMPLETE MODIFIED ACSP EXPLANATION INTERFACE

Appendix A contains all the explanation interface pages from the 2021 Conati *et al.* user study [3] compared with the equivalent updated explanation interface pages used in this study.

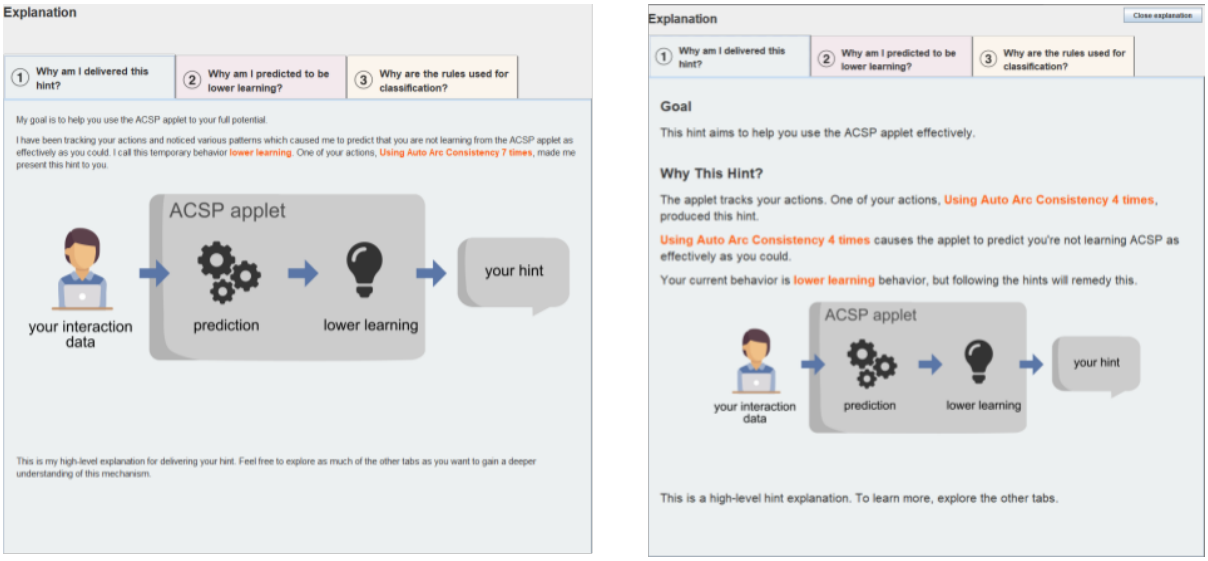

Fig. 9. The old *WhyHint* window on the left, the new *WhyHint* window on the right.

Fig. 10. The old *WhyLow* window on the left, the new *WhyLow* window on the right.

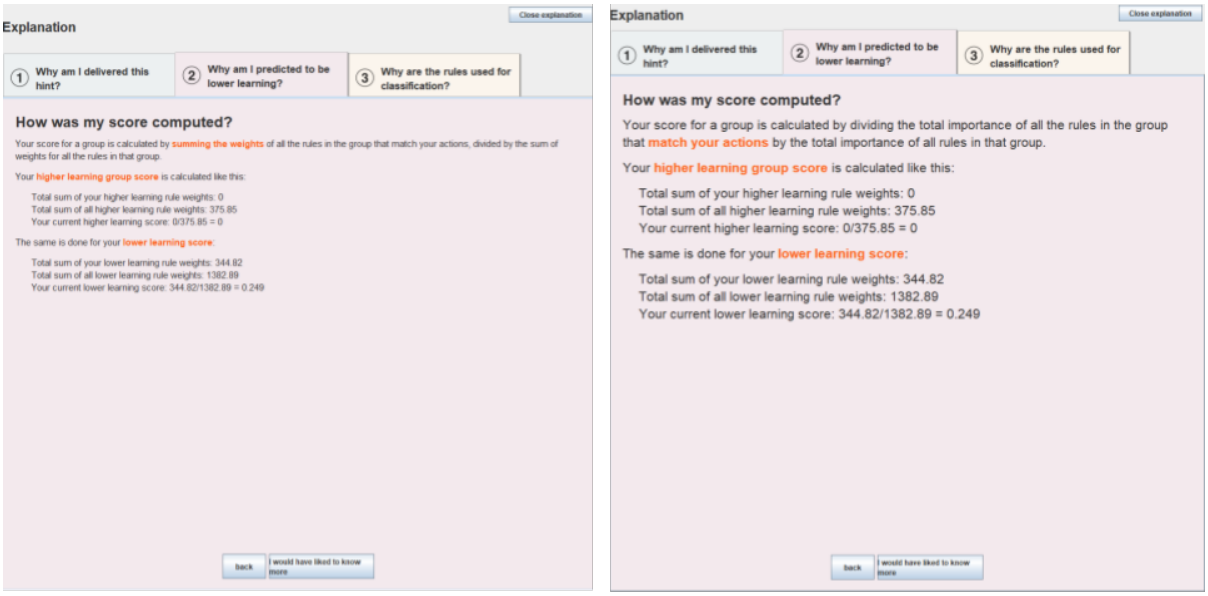

Fig. 11. The old *HowScore* window on the left, the new *HowScore* window on the right.

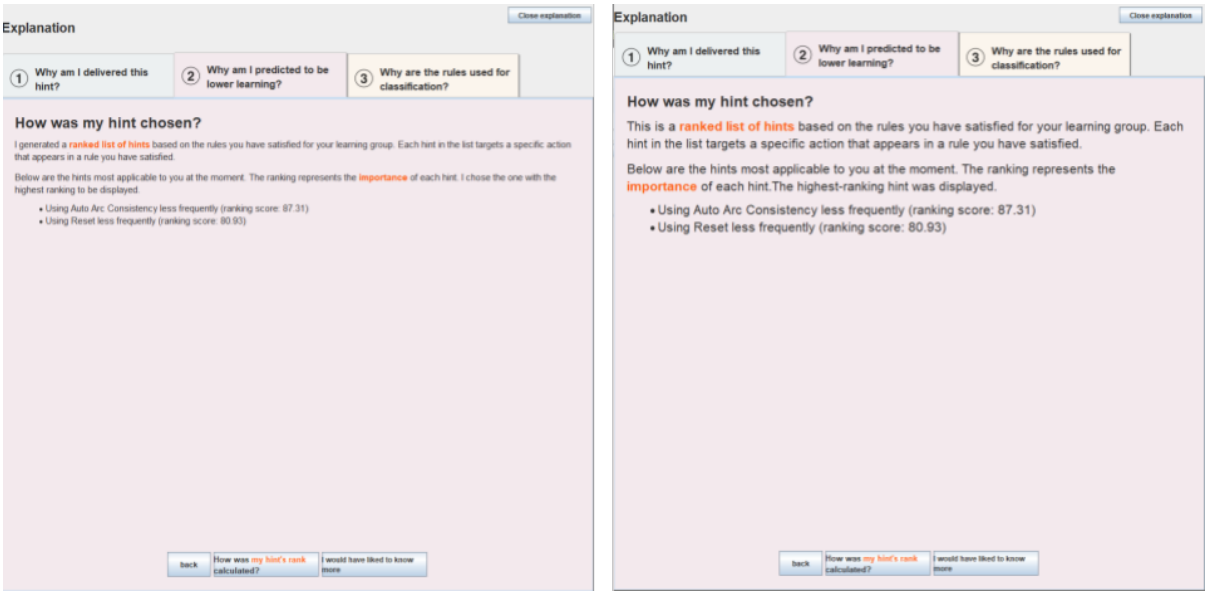

Fig. 12. The old *HowHint* window on the left, the new *HowHint* window on the right.

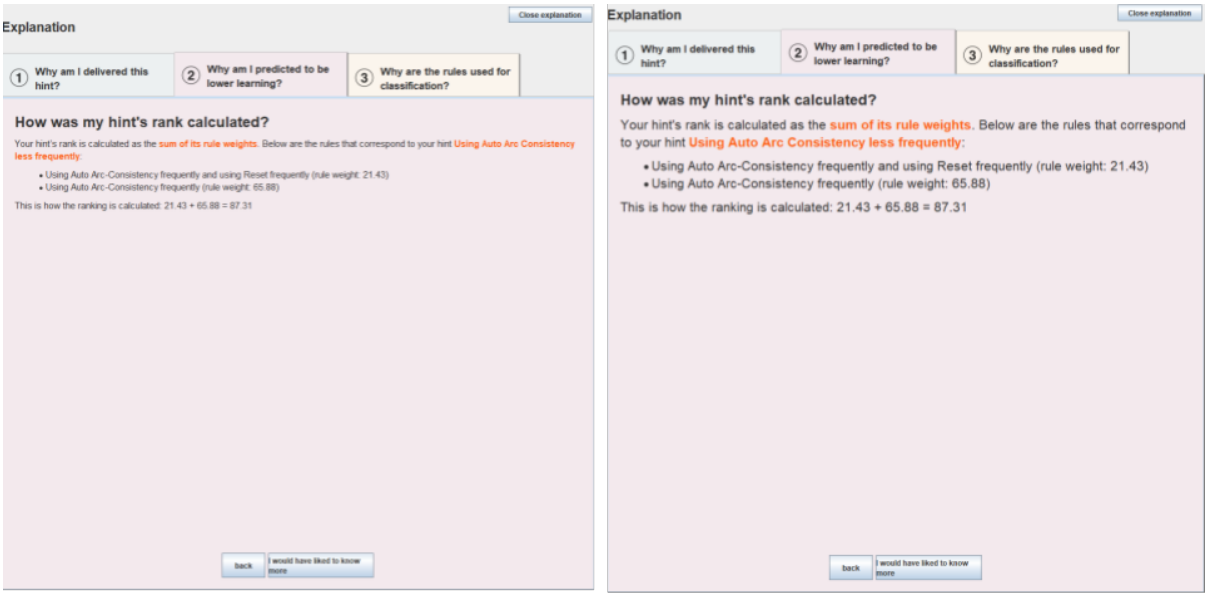

Fig. 13. The old *HowRank* window on the left, the new *HowRank* window on the right.

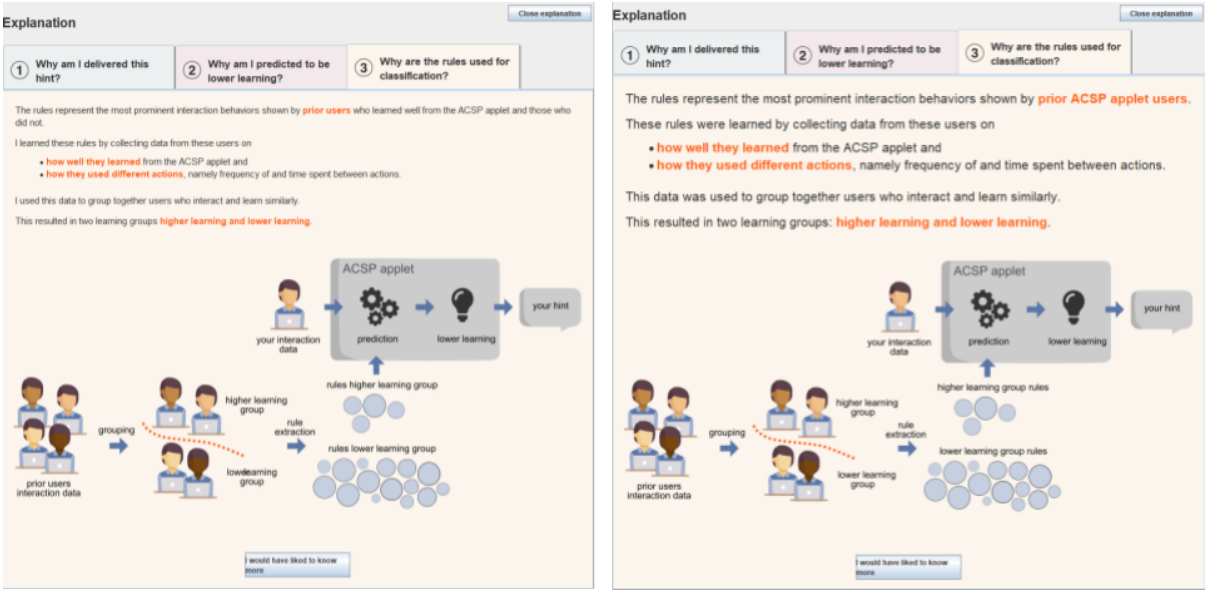

Fig. 14. The old *WhyRules* window on the left, the new *WhyRules* window on the right.