

Toward XAI for Intelligent Tutoring Systems: A Case Study

Vanessa Putnam

Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

Lea Riegel

Department of Computer Science
Augsburg University
Augsburg, Germany

Cristina Conati

Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

ABSTRACT

Our research is a step toward understanding when explanations of AI-driven hints and feedback are useful in Intelligent Tutoring Systems (ITS). We added an explanation functionality for the adaptive hints provided by the Adaptive CSP (ACSP) applet, an intelligent interactive simulation that helps students learn an algorithm for constraint satisfaction problems. We present the design of the explanation functionality and the results of an exploratory study to evaluate how students use it, including an analysis of how students' experience with the explanation functionality is affected by several personality traits and abilities. Our results show a significant impact of a measure of curiosity and the Agreeableness personality trait and provide insight toward designing personalized Explainable AI (XAI) for ITS.

CCS CONCEPTS

Human Centered Computing → User Studies; Laboratory experiment

KEYWORDS

Explainable Artificial Intelligence (XAI); Intelligent Tutoring Systems (ITS); User Modeling

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2020. Toward XAI for Intelligent Tutoring Systems: A Case Study. In *Proceedings of the ACM IUI 2020 Conference*. ACM, Cagliari, Italy, 10 pages. <https://doi.org/10.1145/1234567890>

1 Introduction

Existing research on Explainable AI (XAI) suggests that having AI systems explain their inner workings to their users can help foster transparency, interpretability, and trust (e.g., [8][14][22]). However, there are also results suggesting that such explanations are not always wanted by or beneficial for all users (e.g., [4][5][11]). Our long-term goal is understanding when having AI systems provide explanations to justify their behavior is useful, and

how this may depend on user differences such as expertise, personality, cognitive abilities, and transient states like confusion or cognitive load. Our vision is that of a personalized XAI, endowing AI agents with the ability to understand to whom, when, and how to provide explanations.

As a step toward this vision, in this paper, we present and evaluate an explanation functionality for the hints provided in the Adaptive CSP (ACSP) applet, an Intelligent Tutoring System (ITS) that helps students learn an algorithm to solve constraint satisfaction problems. ITS research investigates how to create educational systems that can model students' relevant needs, states, and abilities (e.g., domain knowledge, meta-cognitive abilities, affective states) and how to provide personalized instruction accordingly [38]. We chose to focus on an ITS in this paper because—despite increasing interest in XAI research encompassing applications such as recommender systems [14][23][29][31][37], office assistants [5], and intelligent everyday interactive systems (i.e. Google Suggest, iTunes Genius, etc.) [4]—thus far there has been limited work on XAI for ITS. Yet, an ITS's aim of delivering highly individualized pedagogical interventions makes the educational context a high-stake one for AI, because such interventions may have a potentially long-lasting impact on people's learning and development. If explanations can increase ITS's transparency and interpretability, this might improve both their pedagogical effectiveness as well as the acceptance from both students and educators [8].

Related research has looked at the effects of having an ITS show its assessment of students' relevant abilities via an *Open Learner Model* (OLM, [3]), with initial results showing that this can help improve student learning (e.g., [26]) and learning abilities (e.g., ability to self-assess [32]). There is also anecdotal evidence that an OLM can impact students' trust [27].

In this paper, we go beyond OLM and investigate the effect of having an ITS generate more explicit explanations of both its assessment of the students as well as the pedagogical actions that the ITS puts forward based on this assessment. We also evaluate whether a set of student traits and abilities affect student usage and perception of the explanations. The goal here is to ascertain if these user differences can account for parts of the variance we detected in users' reactions to the explanation, and eventually inform guidelines on how to address this variance via explanations personalized to the relevant differences. Despite the fact that varied reactions to explanations have been observed with several AI-

driven interactive systems (e.g., [4][11][14][22]), thus far, there has been little work looking at linking these reactions to individual differences in XAI. Existing results have shown an impact of *Need for Cognition* (a personality trait) [29] and of user *decision-making style* (rational vs. intuitive) [30] on explanations in recommender systems, and of *perceived user expertise* for explanations of an intelligent assistant [34]. Our results contribute to this line of research by looking at explanations for a different type of intelligent system (an ITS), and showing the impact of a measure of *Curiosity* and of the *Agreeableness* personality trait, thus broadening the understanding of which user differences should be further investigated when designing personalized XAI in a variety of application domains.

In the following sections, we first describe related work. Next, we introduce the ACSP and the AI mechanisms that drive its adaptive hints. Then, we illustrate the explanation functionality we added to the ACSP and the study to evaluate it, followed by the results of the study, conclusions, and future work.

2 Related Work

There are encouraging results on the helpfulness of explanations in intelligent user interfaces. For example, Kulesza et al. [22] investigated explaining the predictions of an agent that helps its users organize their emails. They showed explanations helped participants understand the system’s underlying mechanism, enabling them to provide feedback to improve the agent’s predictions. Coppers et al. [9] added explanations to an intelligent translation system, to describe how a suggested translation was assembled from different sources, and showed that these explanations helped translators identify better quality translations. Other substantial positive results on explanations were found in the field of recommender systems (RS, e.g., [23][31]). In particular, Kulesza et al. [23] investigated the soundness (“nothing but the truth”) and completeness (“the whole truth”) of explanations in a music RS and found that explanations with these attributes helped users to build a better mental model of the music recommender.

There is, however, also research showing that explanations might not always be useful or wanted. Herlocker et al. [14] evaluated an explanation interface for an RS for movies. Although 86% of the users liked having the explanations the remaining 14% did not. Similarly, Bunt et al. [5] added explanations to a mixed-initiative system suggesting personalized interface customization, and showed 60% of users appreciated the explanations whereas others considered the explanation as common sense or unnecessary. In [4], the authors conducted a survey study asking participants if they would like to receive explanations on the workings of everyday AI-driven applications (e.g., Google Suggest, iTunes Genius), qualified as low-cost in terms of their impact on the users’ stakes. Users were also asked their intuition on how the underlying AI worked. Most users had reasonable mental models of this, without the help of explanations. Only a few wanted additional information.

Some research looking at the role of individual differences in XAI has focused on user preferences, mainly in the context of RSs. For instance, Cotter et al. [10] showed that users prefer explanations for why a recommender works the way it does to explanations that

describe how it works when receiving recommendations in the Facebook news feed. Kouki et al. [21] report a crowd-sourced study showing that users prefer item-centric to user-centric or socio-centric explanations, although preference for the latter type is modulated by levels on the Neuroticism personality trait. Furthermore, users preferred textual explanations. Tsai and Brusilovsky [35] evaluated twelve visual explanations and three text-based explanations in an RS for conference attendees. Participants reported a preference for visual explanation over text-based explanation, although it was shown that the preferred explanation type was not always the most effective.

Going beyond user preferences, Millicamp et al. found moderating effects of *Need for Cognition* (a personality trait) [6] on user confidence in the recommendations with and without explanations, delivered by a music RS, as well as on user preference for different types of explanations [29]. Naveed et al. [30] found an impact of user decision-making style (rational vs. intuitive) on user perception of different types of explanations when looking at mocked-up recommendations for buying a camera. Schaffer et al. [34] found that explanations of the suggestions generated by an intelligent assistant that helped play a binary decision game were only useful for users who declared low ability at the game, whereas they had no impact on users who were overconfident of their ability.

Within ITS, there has been research on increasing transparency via Open Learner Modeling, namely tools that allow learners to access the ITS’s current assessment [3]. Although there is no clear understanding of how OLM can be beneficial for interpretability and explainability of ITS, there is evidence of an effect on learning. For instance, Porayska-Pomsta and Chrysafidou [32] did a preliminary evaluation of the OLM for a job interview coaching environment, with results suggesting that the OLM helped users to improve their self-perception and interview skills. Long and Aleven [26] report on the positive effect of an OLM for an ITS designed to foster student self-assessment abilities in algebra skills. There is also anecdotal evidence that an OLM can impact students’ trust [27], where interestingly students trusted an ITS with an OLM more when they could not change assessment in the student model. Barria-Pineda et al. [2], add explanations to an OLM, but the explanations are essentially textual rephrasing of the OLM assessment. Our work goes beyond OLMs by investigating more explicit explanations of an ITS underlying AI mechanism.

3 The ACSP Applet

3.1 Interactive Simulation for AC-3

The ACSP applet is an interactive simulation that provides tools and personalized support for students to explore the workings of the Arc Consistency 3 (AC-3) algorithm for solving constraint satisfaction problems [25]. AC-3 represents a constraint satisfaction problem as a network of variable nodes and constraint arcs. The algorithm iteratively makes individual arcs consistent by removing variable domain values inconsistent with a given constraint, until it has considered all arcs and the network is consistent. Then, if there remains a variable with more than one domain value, a procedure called domain splitting is applied to that variable in order to split the CSP into disjoint cases so that AC-3 can recursively solve each case. The ACSP applet demonstrates the

AC-3 algorithm dynamics through interactive visualizations on graphs using color and highlighting (see Figure 1). The applet provides several mechanisms (accessible via buttons in the toolbar at the top of the ACSP interface) for the interactive execution of the AC-3 algorithm on available problems [1], including: *Fine Step*: goes through AC-3 three basic steps of selecting an arc, testing it for consistency, removing domain values to make the arc consistent; *Direct Arc Click*: allows the user to select an arc to apply all these steps at once. *Auto AC*: automatically fine step on all arcs one by one. *Domain Split*: select a variable to split on and specify a subset of its values for further application of AC-3 (see the pop-up box on the left side of Figure 1). *Backtrack*: recover alternative networks during domain splitting. *Reset*: return the graph to its initial status.

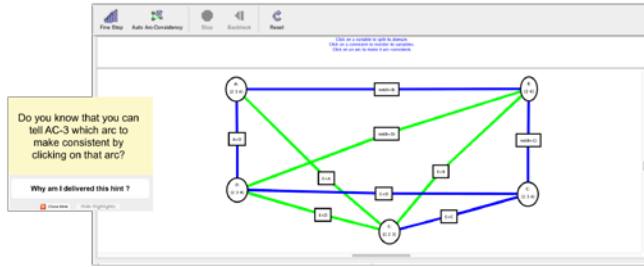


Figure 1: The ACSP applet with an example CSP and hint.

The ACSP also includes a user model that monitors how a student uses the available tools and recognizes interaction patterns that are not conducive to learning. It then leverages the predictions of the user model to generate hints guiding the student towards a more effective usage of the available tools. This user model and hint delivery mechanisms are derived based on a general framework for modeling and supporting exploratory, open-ended interactions (FUMA, Framework for User Modeling and Adaptation [15][18]). The next two sections summarize these mechanisms since they are the targets of the explanations that we added to the ACSP.

3.2 Modeling User Behaviors in the ACSP

Figure 2 illustrates how the FUMA framework is integrated into the ACSP. In FUMA, the process of building a user model consists of two phases: Behavior Discovery (Figure 2 – top) and User Classification (Figure 2– bottom right). In the following, numbers in curly braces correspond to the graph’s elements in Figure 2. The Behavior Discovery phase leverages existing datasets of students working with the CSP applet without adaptive hints. Data from existing interaction logs {1} is preprocessed into feature vectors consisting of statistical measures that summarize users’ actions (i.e., action frequencies, time interval between actions) {2, 2.1}. Each vector summarizes the behaviors of one user. These vectors, along with data on each student’s learning gains with the system {3}, are fed into a clustering algorithm. The algorithm groups the feature vectors according to their similarities while also ensuring groups have significantly different learning performance. Therefore, the algorithm identifies clusters of users who interact and learn similarly with the interface {4, 4.1}. Next, association rule mining is applied to each cluster to extract its identifying interaction behaviors {5, 5.1}.

The rules are weighted based on how well they discriminate between the two clusters, namely based on a combination of their confidence (i.e., the relative frequency of a rule in this cluster compared to others) and their support (i.e., how frequently a rule appears in a cluster) {6, 6.1}. Based on these rules, a human designer then defines a set of hints {14, 14.1} aimed at discouraging behaviors associated with lower learning and promote behaviors associated with higher learning.

This behavior discovery mechanism was applied to a data set of 110 users working with the CSP applet without adaptive support [18][15]. Learning gains for these users were derived from tests on the AC-3 algorithm taken before and after using the system. From this data set, Behavior Discovery generated two clusters of users that achieved significantly different levels of learning, labeled as Higher Learning Gain (HLG) and Lower Learning Gain (LLG). A total of four and fifteen rules were found for the HLG and LLG, respectively, a selection of which is presented in Table 1. The hints that were derived from these rules are listed in Table 2.

Rules for HLG cluster

- Rule 1: Infrequently auto solving the CSP
- Rule 2: Infrequently auto solving the CSP and infrequently stepping through the problem
- Rule 3: Pausing for reflection after clicking CSP arcs

Rules for LLG cluster

- Rule 4: Frequently backtracking through the CSP and not pausing for reflection after clicking CSP arcs
- Rule 8: Frequently auto solving the CSP and infrequently clicking on CSP arcs
- Rule 10: Frequently resetting the CSP

Table 1: A subset of representative rules for HLG and LLG clusters.

- Use Direct Arc Click more often;
- Spend more time after performing Direct Arc Clicks;
- Use Reset less frequently;
- Use Auto Arc-consistency less frequently;
- Use Domain Splitting less frequently;
- Spend more time after performing Fine Steps;
- Use Back Track less frequently;
- Use Fine Step less frequently;
- Spend more time after performing reset for planning;

Table 2: Hint descriptions

User Classification is the second phase involved in building the ACSP applet’s user model (Figure 2– bottom right). In this phase, the clusters, association rules, and corresponding rule weights extracted in the Behavior Discovery phase are used to build an online classifier {9}. As a new user interacts with the ACSP applet, the classifier predicts the user’s learning after every action. This is done by (i) incrementally building a feature vector based on the interface actions seen so far {7, 8} and (ii) classifying this vector in one of the available clusters {11, 11.1, 11.2, 12}. Note that the classification can change over time, depending on the evolution of the user’s interaction behaviors.

3.3 Adaptive Hints

In addition to classifying a user in one of the available clusters, the ACSP’s user model also returns the satisfied association rules causing that classification {10}. These rules represent the characteristic interaction behaviors of a specific user so far. If the

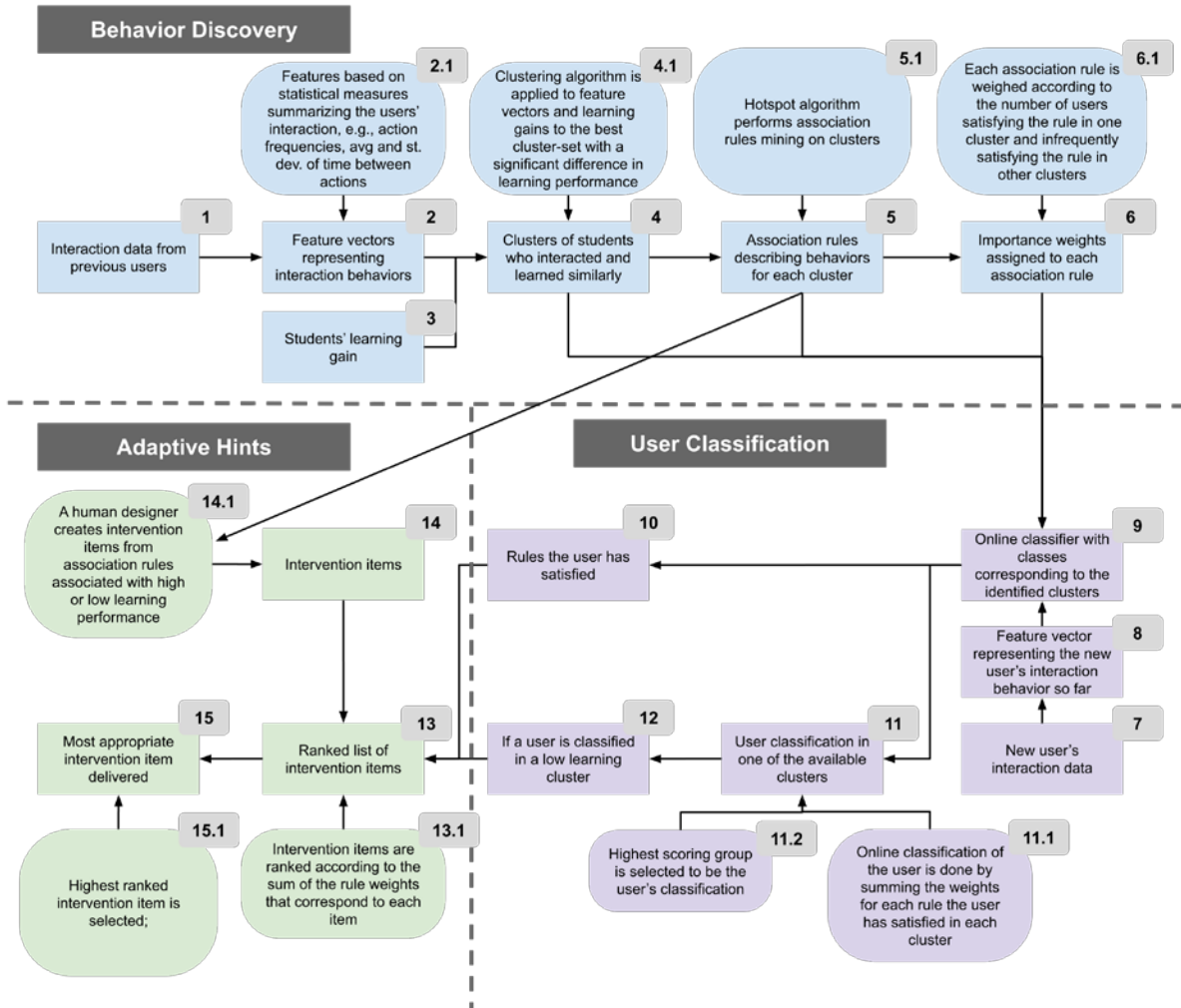


Figure 2: ACSP User Modeling Framework broken down into three phases: Behavior Discovery, User Classification, and Adaptive Hints; rectangular nodes represent inputs and states, oval nodes represent processes

user is classified as belonging to a cluster associated with lower learning, the process of providing adaptive hints triggers (Figure 2–bottom left). This process starts by identifying which of the hints in Table 2 should be provided when a student is classified as a lower learner at a given point of their interaction with the ACSP. More specifically, when a user is classified as a lower learner, the ACSP identifies which detrimental behaviors this user should stop performing or which beneficial behaviors they should adopt, based on the association rules that caused the classification.

Generally, a combination of rules causes the user to be classified as a lower learner, and thus, several hints might be relevant. However, to avoid confusing or overwhelming the user, the applet only delivers one hint at a time, chosen based on a ranking that reflects how predominant each of the behaviors associated with the possible hint is. After each prediction of lower learning, every item in Table 2 is assigned a score proportional to the sum of the weights of the association rules that triggered that item and the lower learning classification {13, 13.1}. The hint with the highest score is chosen to be presented to the student {15, 15.1}.

The ACSP delivers its adaptive hints incrementally. Each hint is first delivered via a textual message that prompts or discourages a target behavior. For instance, a hint for the *Use Direct Arc Click more often* item in Table 2 is “Do you know that you can tell AC-3 which arc to make consistent by clicking on that arc?” (see Figure 1). After receiving the hint, the student is given some time to change their behavior accordingly (a reaction window equal to 40 actions). During this time, the user model will keep updating its user classification. At the end of this time window, the user model determines whether the user has followed the hint for the target item or not, and if not, the target item is selected for delivery again, this time accompanied by stronger guidance, e.g., highlighting of relevant interface items.

The ACSP was evaluated against a non-adaptive version with a formal study where two groups of 19 students studied three CSP problems with the adaptive and control version, respectively [18]. The study showed that students working with the ACSP learned the AC-3 algorithm better than students in the control conditions and followed on average about 73 % of the adaptive hints they received.

Although these results are very positive, it is worth investigating if and how explanations of the ACSP adaptive hints might increase students' uptake and learning.

4 Explanation Interface

4.1 Pilot User Study

To gain an initial understanding of the type of explanations that students would like to have about the ACSP hints, we instrumented it with a tool to collect this information. Namely, we added to each hint's dialogue box a button "explain hint" that enables a panel allowing students to choose one or more of the following options for explanations they would have liked for these hints: (i) *why* the system gave this hint; (ii) *how* the system chose this hint; (iii) some other explanation about this hint (including a text field for user input); (iv) no explanation.

We ran a pilot with nine university students with adequate prerequisites to use the ACSP applet. We told participants that we were looking for feedback on how to enrich the ACSP applet with explanations for its hints. During their interaction with the ACSP, the participants accessed the "explain hint" functionality for 51% of the hints delivered. Of these responses, 47% asked for a *why* explanation, 30% for *how*, 14% for none, and 9% for other. These results confirm that participants are generally interested in explanations, although to different extents, which is consistent with findings from Castelli et al. [7] and Cotter et al. [10] in terms of users preferring *why* explanations, followed by *how* explanations.

Based on the results from this pilot study, we designed and implemented an explanation interface that conveys to the ACSP users the motivations (*why*) and processes used (*how*) for each of the hints they receive. Essentially, these explanations should provide the ACSP users with insights on the user modeling and hint provision mechanisms described in Section 3.

4.2 Design Criteria

As guidance for the explanation design, we rely on some of the criteria articulated by Kulesza et al. [22]. Specifically, in principle we want our explanations to be

- *Iterative*, namely accessible at different levels of detail based on the user's interest
- *Sound*, namely conveying an accurate, not simplified nor distorted description of the relevant mechanisms
- *Complete*, namely exposing all aspects of the relevant mechanisms
- *Not overwhelming*, namely comprehensible and not conducive to excessive cognitive load or other negative states such as confusion and frustration

There is a trade-off that needs to be made between complying to the requirements of soundness, completeness, and avoiding that the explanations become overwhelming. The iterative criterion is an important means to achieve this tradeoff, and it has a predominant role in the explanation functionalities we are designing. However, the AI driving the ACSP hints is a complex combination of three different algorithmic components (behavior discovery, user classification, and hint selection, see Section 3). To determine the explanation's content, the authors discussed these components at length and decided to start designing and evaluating a version of

the explanation that sacrifices completeness when it is needed to avoid excessive complexity. We do so by prioritizing *why* over *how* explanations, following the results from the pilot study described in the previous section. The rationale for this choice is to start evaluating a meaningful, albeit incomplete, set of explanations and get feedback from the users regarding how much more information they would like to see.

Based on this strategy, we identified three self-contained *why* explanations, as well as three *how* explanations, described in Section 4.3. We derived these explanations from the graph in Figure 2, which represents all the inputs and states (rectangular nodes) involved in the hint computation as well as the specific processes (oval nodes) that generate each state from preceding ones. We use the states in Figure 2 to justify specific aspects of the rationale for hint computation (*why* explanations) and the processes to explain *how* some of the relevant algorithm components work.

We then came up with several designs to structure and navigate through these explanations, which we prototyped using a tool called Marvel¹ to create fast wire-frame interfaces for the different designs. Test piloting the different designs revealed that the most intuitive and easy to use navigation is the tab-based design illustrated in the next section.

4.3 Navigation and Content

We structured the explanation interface around three tabs, each providing a self-contained, incremental part of the explanation for a given hint, as shown in Figure 3. Each tab displays a *why* explanation; for one of these *why* explanations (tab in 3(B)), we allow users to ask for more details on *how* three specific aspects where computed (Figure 3 (D)–(E)).

We refer to the different parts of the explanation as pages (WhyHint, WhyLow, WhyRules, HowScore, HowHint, and HowRank page for future reference). The content of each page, not shown in Figure 3, will be illustrated later in this section. Since we did not explain the ACSP's User Modeling and Behavior Discovery to full extent, users can provide feedback on the explanation's content by using a button labeled "I would have liked to know more" that is accessible on every page.

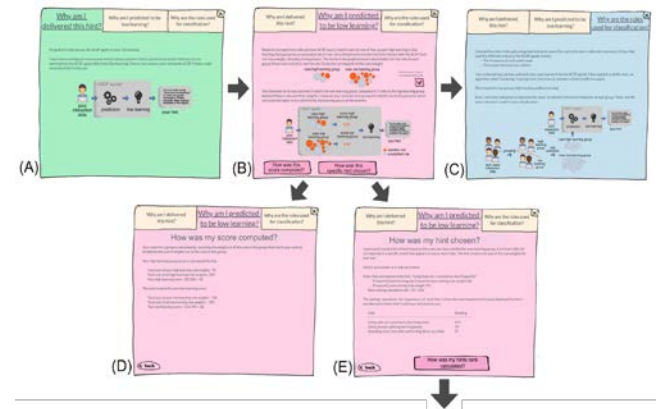


Figure 3: Flow Chart of Explanation Navigation (A) Why am I delivered this hint? (B) Why am I predicted to be lower learning? (C) Why are the rules used for classification? (D)

¹ <https://marvelapp.com>

How was this score computed? (E) How was this specific hint chosen? Page: “How was my hint’s rank calculated?” not shown (see arrow from (E)).

As mentioned above, we built these six pages of explanations from the graph in Figure 2. We selected and assembled various elements of the graph to create sound and coherent incremental explanations that the user can access at will. Note that we were originally hoping to have a one-to-one mapping between elements in the graph and explanation pages, but quickly realized that this would result in explanations that were too fragmented. The rest of this section provides the full content of each explanation page, including text and accompanying visualizations. Added numbers correspond to the graph’s elements in Figure 2 that are discussed by that text. These numbers have been added here for illustration. They are not present in the explanation seen by the users. It is important to note that our explanation is personalized and dynamically updates according to the user’s real-time interaction. The following explanation is exemplary for a hint stating, “You have used the Reset button excessively. I recommend that you limit your usage of this action.”

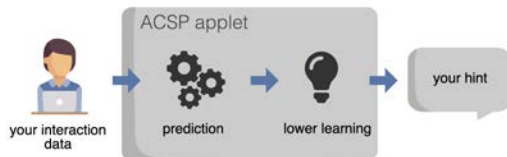
The user can activate the explanation functionality once the hint has been delivered, by clicking the button labeled “Why am I delivered this hint?” (see Figure 1). In response to this request, the explanation window appears, with the first tab to the left active, as shown in Figure 3(A). The next three subsections describe the *why* explanations provided in the three tabs, as well any *how* explanation that can be requested from there.

4.3.1 Why am I delivered this hint?

The explanation in this tab provides a high-level explanation of the user classification component and how it is linked to the hint received.

My goal is to help you use the ACSP applet to your full potential. I have been tracking your actions {7} and noticed various patterns {10} which caused me to predict that you are not learning from the ACSP applet as effectively as you could.

*I call this temporary behavior **lower learning** {12}. One of your actions, **Using Reset 4 times**, made me present this hint to you.*



Note that, although the first two sentences of the explanation illustrate general aspects of the rationale for hint provision, the last one provides information that is specific to this user.

4.3.2. Why am I predicted to be lower learning?

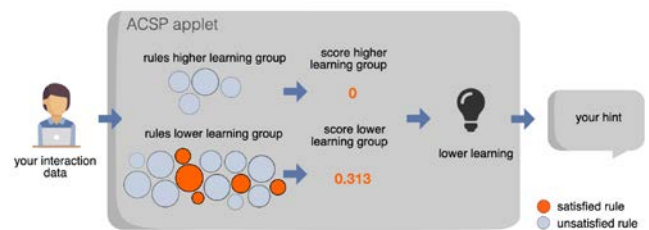
Selecting the second tab in the interface “Why am I predicted to be lower learning?”, will give access to a more specific explanation on why the ACSP user model came up with this classification.

*I classify users as one of two groups: **higher learning or lower learning** {4,9}. Each group has an associated **set of rules** describing how its members tend to interact with the ACSP {5}. Each rule has a **weight**, denoting its importance {6}. Certain actions satisfy certain rules [examples can be accessed here]. The circles in the graph below represent the rules in each group. Hover over a circle to see the rule. Circle size corresponds to the rule’s*



weight.

*Your behavior so far has matched **5 rules** in the **lower learning group**, compared to **0 rules** in the **higher learning group** {10}. Based off these rules’ weights, I computed your **score for each group** and classified you in the group for which you have the*



*higher score at the moment, namely the **lower learning group**.*

Within this tab, the user can access an additional visualization linking their actions to the satisfied rules and their weights (not shown for lack of space). Users can also choose to ask more details on (i) *how* their scores for each group were computed and (ii) *how* the specific hint delivered was selected, see buttons at the bottom of Figure 3(B), and resulting pages 3(D) and 3(E) respectively. Their content is presented below.

How was my score for each group computed?

*Your score for a group is calculated by **summing the weights** of all the rules in the group that match your actions, divided by the sum of weights for all the rules in that group {11.1}.*

***Your higher learning group score** is calculated like this:*

Total sum of your higher learning rule weights: 0
Total sum of all higher learning rule weights: 376
Your current higher learning score: $0/376 = 0$

*The same is done for your **lower learning score**:*

Total sum of your lower learning rule weights: 432
Total sum of all lower learning rule weights: 1383
Your current lower learning score: $432/1383 = .313$

How was my hint chosen?

*I generated a **ranked list of hints** {13} based on the rules you have satisfied for your learning group {10}. Each hint in the list targets a specific action that appears in a rule you have satisfied. Below are the hints most applicable to you at the moment. The ranking represents the **importance** of each hint. I chose the one with the highest ranking to be displayed {15, 15.1}.*

- *Using Reset less frequently (ranking : 98)*
- *Using Auto Arc Consistency less frequently (ranking:87)*
- *Spending more time after performing Fine Steps (ranking: 18)*

Within page 3(E), the user can navigate further to read more about how their hint’s rank was computed (button Figure 3(E) bottom).

How was my hint's rank calculated?

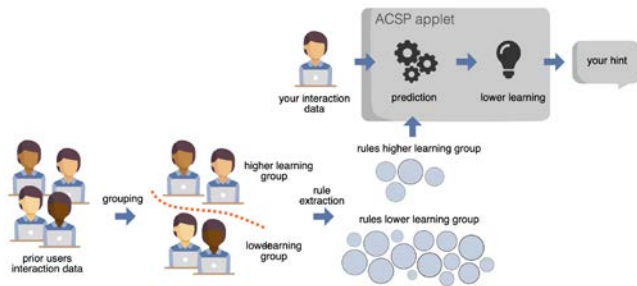
Your hint's rank is calculated as the **sum of its rule weights** [13.1]. Below are the rules that correspond to your hint **Using Reset less frequently**:

- Using Reset less frequently and short pausing after performing Fine Step (rule weight: 18)
- Using Auto Arc-Consistency frequently and using Reset frequently (rule weight: 21)
- Using Reset frequently and regularly pausing after performing Domain Splitting (rule weight: 19)
- Using Reset frequently (rule weight: 40)

4.3.3 Why are the rules used for classification?

Selecting the third tab in the interface (see Figure 3(C)), will provide a high-level description of the Behavior Discovery phase, including background information on the data used to create the classifier and how this relates to what has already been explained.

The rules represent the most prominent interaction behaviors {5} shown by **prior users** who learned well from the ACSP applet and those who did not {1}. I learned these rules by collecting data from these users on **how well they learned** from the ACSP {3} and **how they used different actions** {2}, namely frequency of and time spent between actions. I used this data to group together users who interact and learn similarly. This resulted in two learning groups **higher learning and lower learning** {4}.



Note that in this tab, we could have enabled explanations on *how* different parts of the Behavior Discovery process work, e.g., clustering and rules extraction. However, because explaining these algorithms can be quite complicated, here is where we chose to give up explanation completeness and see how users react to this choice in the formal study described in Section 5.

5 User Study

This section illustrates the exploratory study we conducted to evaluate the ACSP's explanation functionality for usability and user attitude (i.e., whether participants use the explanation functionality and how they perceived it). Given the complexity of the explanation described in the previous sections, we argue that before engaging in a formal controlled study to compare the ACSP with and without the explanation, it is crucial to have a clear sense of whether such an explanation is wanted and accessed in the first place. With this study, we also took the opportunity to start

investigating the impact of individual differences on users' attitudes toward the ACSP explanation.

5.1 Participants and Procedure

43 participants (21 female, 22 male) were recruited through advertising at our campus. They were required to have enough computer science knowledge to learn the concept of CSPs, e.g., basic graph theory and algebra, and to not have colorblindness.

The procedure for our study followed the one used in [18] to evaluate the ACSP applet hints, without a control condition and with minor modifications to cover the evaluation of the explanation and individual differences. The study task was to use the ACSP applet to understand how the AC-3 algorithm solves three CSP problems [18]. Participants were told that the ACSP would provide adaptive hints during their interaction and that they could access the explanation on *why* and *how* the hints were provided. Participants were shown how to access the explanation functionality but were told that it was up to them to decide whether to use it or not. The experimental procedure was as follows: participants (1) took tests on individual differences (see next section); (2) studied a textbook chapter on the AC-3 algorithm; (3) wrote a pre-test on the concepts covered in the chapter; (4) watched an introductory video on how to use the main functionalities of the ACSP applet; (5) used the ACSP applet to solve three CSPs; (6) took a post-test analogous to the pre-test; and (7) answered a post-questionnaire (see section 6.3). The study took between 2.5 and 3 hours in total. Participants were compensated with \$30.

5.2 Individual Differences

The individual differences considered in this study include *cognitive abilities* that can affect how easy it is for a user to process the explanation's content, as well as *traits* that can impact a user's perception of the explanations. All the individual differences were measured using state-of-the-art tests from Psychology.

For *cognitive abilities*, we measured *Perceptual Speed* (i.e., speed in comparing figures or symbols [11]), *Visual Working Memory* (i.e., the quantity of visual information that can be temporarily maintained and manipulated in working memory [36]), and *Reading Proficiency* (i.e., vocabulary and reading comprehension ability in English [28]) to uncover differences in users' abilities to process the diagrams and text in the explanation. We also measure users' *Locus of Control* [33] or the degree to which they attribute outcomes to their own behavior or outside forces. For *user traits*, we included

- *Need for Cognition* (extent to which one is inclined towards effortful cognitive activities), because it was found to have an impact on explanation effectiveness in [29]
- The five personality dimensions *Agreeableness*, *Conscientiousness*, *Extraversion*, *Neuroticism*, and *Openness* [12] since at least one of them was found to have an impact on explanation preference in [21]
- Two dimensions of *Curiosity* [19]²: *Joyous Exploration* (i.e., the extent to which one derives positive emotions from learning new information and experiences) and *Deprivation Sensitivity* (i.e., the desire to reduce gaps in knowledge because they generate feelings of anxiety and tension). We added these traits because

² The test here measures three other dimensions of curiosity not relevant to our context

some users in the pilot study in Section 4.1 mentioned curiosity when asked reasons for wanting explanations.

5.3 Measurements

To ascertain how participants accessed the explanation, we tracked all their interaction events and extracted a variety of explanation-related actions, upon which we computed the summative statistics described in Section 6.1. These actions include:

- *Explanation initiation*: starting the explanation for a given hint;
- *Page accessed*: viewing any one of the explanation’s pages; during each initiation, there can be multiple pages accessed for each available page;
- *Explanation type accessed*: accessing one of the six types of explanations available (see Figure 3); thus, the number of explanation type accessed ranges from 1 to 6.

We also collected subjective feedback on the explanation functionality (see Table 3).

Items on Usefulness
I would choose to have the explanations again in the future.
I am satisfied with the explanations.
The explanations were helpful for me.
Items on Negative Impressions
The explanations distracted me from my learning task.
The explanations were confusing.
I found the explanations overwhelming.
Items on Usability
It was clear to me how to access the explanations.
The explanation navigation was clear to me.
The explanation content (i.e., wording, text, figures) was clear to me.

Table 3: Explanation Questionnaire Items

The items in Table 3 were rated on a 5-point scale ranging from *strongly disagree* (1) to *strongly agree* (5) and were selected from a variety of sources including the Usefulness, Satisfaction, and Ease of use (USE) questionnaire, as well as established XAI literature (e.g., [5], [29], [20]). The first three items target the general usefulness of the explanation, gauging *users’ intention* to use again, *satisfaction*, and perceived *helpfulness*. We evaluate users’ negative impressions the same way as Kardan et al. previously evaluated the ACSP in terms of both *confusion* and *distraction* [18]. We added an item for *overwhelming* because it is one of the specific design criteria for explanations in [22]. We evaluate usability in terms of clarity of accessibility, navigation, and content to ensure none of these factors inhibited users from using the explanations. Instead of this questionnaire, participants who did not view the explanation answered the open-ended question “Please describe why you did not access the explanation, using the button ‘Why was I delivered this hint?’”.

Participants filled out a second questionnaire for the ACSP hints, that included the items for usefulness and negative impressions, but replaced the items on usability with items related to trust and understanding why the hints were delivered.

6 Results and Analysis

Of the 43 study participants, 17 did not receive any hints during their interaction with the ACSP because the system assessed that they did not need help to learn effectively. This group, in fact, obtained an average percentage learning gain (PLG)³ of 56% (SD = 21%), which is higher than the average PLG of the group who received hints (46%, SD = 29%) and in line with the PLGs of higher learners reported in previous studies on the CSP applet without hints [17][16]. Since these 17 participants did not have the opportunity to access the explanation, the analyses and results in the following sections focus on the 26 participants (14 female, 12 male) who did receive hints.

6.1 Interaction with Explanation Interface

Out of the 26 participants who received hints, 20 of them (77%) accessed explanations, showing that there is substantial interest for this functionality, but also confirming previous findings that not all users want explanations. The six participants who did not access explanations stated the following reasons in their free text post-questionnaire answers: three said that they were not interested in the hints, they just wanted to complete their task on their own; the other three reported that the hints did not need further explanation. In the following, if not stated differently, the statistics presented entail the 20 participants who initiated the explanation. Any formal comparison between these participants and the 6 who did not access explanations is not feasible due to the small number of the latter group.

Figure 4 breaks down, for each participant, the number of hints received, compared to their *why* page vs. *how* page accesses, and gives a general sense of the variability with which these 20 participants engaged with the explanation functionality.

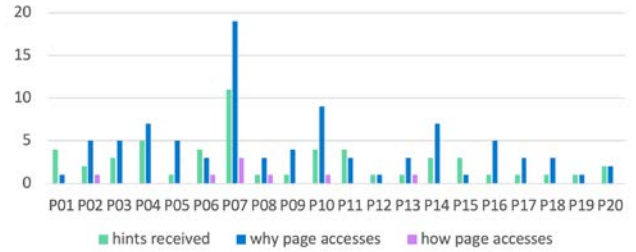


Figure 4: Number of hints, number of why page accesses, and number of how page accesses per participant

Participants received on average 2.7 hints, with large standard deviation (2.4) and range (minimum of one hint and maximum of 11). Table 4 provides detailed summative statistics on how participants approached the explanation interface.

The first row concerns how participants initiated the explanation in response to hints. Participants tended to initiate the explanation on the first hint, or the second at the latest. The ratio of explanation initiations over the number of hints (second row) received is 0.76 on average, i.e., participants initiated the explanation for 3/4 of the hints received, indicating that some participants were eager to view explanations and went back to the explanations for subsequent hints.

³ Difference between post-test score and pre-test score over the difference between the tests’ maximum score and pre-test score

The last three rows in Table 4 give a sense of how much participants actually explored the explanation interface. An average of almost 3 pages were accessed per each initiation, with a minimum of 1 and maximum of 5. Participants spent an average of 66.2s in the explanation interface, with a notable standard deviation of 55.5s and a range between 5.4s and 191s. Note that, although total time spent could depend on number of hints received, the two measures were not significantly correlated (Pearson $r = 0.38$, $p = 0.1$), thus the large variance in total time spent is likely due to reasons other than hints received. Finally, of the 6 different types of explanation pages available, close to 3 were seen on average, with a range between 1 and 5.

	Mean	SD	Min	Max
Hints before first explanation initiation	1.10	0.31	1	2
Explanation initiations over number of hints received	0.76	0.30	0.25	1.00
Number of pages accessed per initiation	2.95	1.36	1	5
Total time spent in explanation	66.2s	55.5s	5.4s	191.6s
Distinct types accessed	2.80	1.24	1	5

Table 4: Summative statistics on usage of the explanation

Going into more detail, Figure 5 visualizes the proportion of each explanation type accessed, as well as the proportion of time spent on each type. This gives the picture of users mainly being interested in the first two *why* pages, as taking together the proportions for WhyHint and WhyLow (Figure 3A and 3B), makes up about two thirds of total accesses or duration. The third type of *why* explanation (WhyRules, Figure 3C) takes up most of remaining third. As far as the *how* explanations are concerned, the proportions of accesses and time spent decrease from HowScore to HowHint (Figure 3D) and 3E), the two types that can be directly accessed from WhyLow, and reach zero for the *how* page on how a hint's rank was computed (HowRank, Figure 3F). Only one participant made use of the "I would have liked to know more" button, wishing for more details on the rules.

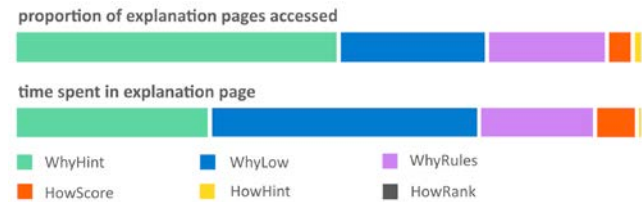


Figure 5: Proportion of time spent in and number of accesses for each type of explanation page

6.2 Subjective Ratings

Analyzing the questionnaire items on the explanation functionality (Table 3) reveals that users were in general positive about it. This can be seen in (Figure 6 (A)), with the high ratings for the items related to intention to use (int) and satisfaction (sat), whereas helpfulness (help), has more room for improvement. The low ratings for distraction (dist), confusion (conf), and overwhelming (over) (Figure 6(B)) also speak in favor of the explanation functionality, although distracting is the one with the most negative (for mean) rating of the three. Most users strongly agreed that the explanation is clear in access, navigation, and content (Figure 6(C)), suggesting a strong usability.

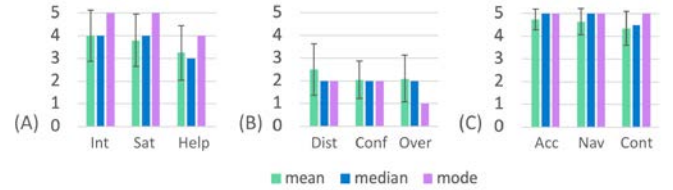


Figure 6: Subjective ratings of the explanation

6.3 Impact of Individual Differences on Explanation Access and Ratings

To ascertain whether the user characteristics tested in the study (see Section 5.2) modulated explanation access, for each of them we ran a MANCOVA with that individual difference as a co-variate, and total time in the explanation interface and number of accesses per initiation (Table 4, rows 3 and 4) as dependent variables. We chose these two dependent measures as representative of the amount of effort a participant was willing to put into exploring the explanation interface. We ran separate MANCOVAs to avoid overfitting our models by including all co-variables at once. Since there was no strong correlation among the tested individual differences, each MANCOVA can be considered as an independent analysis on the impact of the target individual difference on explanation usage. We also run a MANCOVA with pretest score as co-variate, to ascertain the possible effect of existing knowledge on explanation access.

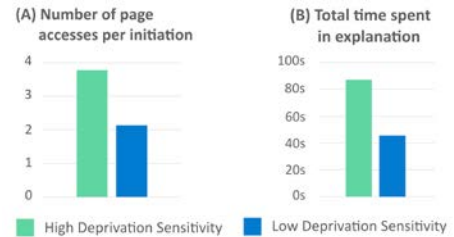


Figure 7: Distribution of total time spent in explanation and number of page accesses per initiation for participants with high and low deprivation sensitivity, split by the median of their scores for this dimension of curiosity.

We found a significant effect (with a large effect size) of the curiosity dimension *Deprivation Sensitivity* (DS) on the number of pages per initiations ($p = .011^*$, $\eta^2 = .310$, $F(1,18) = 8.085$). Users high in DS accessed more pages per initiation than users who are low in DS (Figure 7 (A)). High DS users tend to seek further information because they experience anxiety when they have knowledge gaps. Thus, participants with high levels of this trait may be more inclined to access explanations to better understand why they received a hint. We found a consistent marginally

significant effect (with medium effect size) of DS on the total time spent in the explanation interface ($p = .081$, $\eta^2 = .159$, $F(1,18) = 3.441$), with users high in DS showing a trend of higher time than users low in DS (see Figure 7(B)).

We also checked for possible impacts of individual differences on user ratings in the explanation questionnaire in Table 3. To do so, we ran independent samples Kruskal-Wallis tests on two dependent measures: one derived by taking the average of the three ratings on usefulness, i.e., intention to use, satisfaction, and helpfulness; the other derived by averaging the three ratings on negative impressions, i.e., distraction, confusion, and overwhelming. As we did for the analysis above, we ran separate Kruskal-Wallis tests with each of our individual differences as independent measures. We found a significant effect of the personality trait *Agreeableness* on the combined measure for negative user impressions ($\eta^2 = .381$, $p = .021$, $df = 11$), where lower levels of *Agreeableness* result in more negative impressions. Looking at the specific ratings generated by users with high and low *Agreeableness* (computed via median split over the test values for this personality trait), we see that most of the difference comes from the ratings for *distracting* and *overwhelming* (see Figure 8).

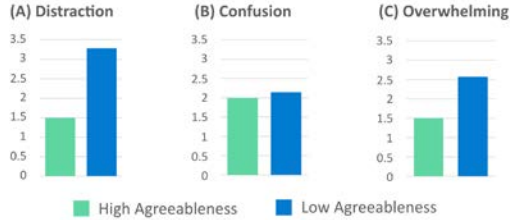


Figure 8: Distribution of combined ratings for distraction, confusion, and overwhelming, for high and low agreeableness split by the median agreeableness value.

6.4 Discussion

Designing an explanation functionality that conveys at least some of the AI mechanisms driving the ACSP adaptive hints has proven to be challenging, because of the complexity of such mechanisms. The study presented in this paper was mainly geared to ascertain that there were no major usability and acceptance issues with the explanation functionality we designed. Our results indicated that, overall, the functionality was rather extensively used, with only six out of 26 users not accessing it, two out of three hints triggering an explanation initiation on average, and almost 3 explanation pages accessed per initiation on average. The functionality also received overall positive subjective ratings, suggesting that it makes sense to move to the next step of evaluating it more formally for impact on students' experience with ACSP, by conducting a user study that compares versions of ACSP with and without explanations.

Our results found a significant impact of two individual differences on explanation access and subjective evaluation. Specifically: (1) users with higher values of the curiosity dimension Deprivation Sensitivity (DS) accessed more explanation pages than their low DS counterparts; (2) users with lower values of the *Agreeableness* personality trait perceived the explanations as more distracting and overwhelming than those with high *Agreeableness*. These results suggest that it is important to continue investigating these two individual differences as factors that could drive personalized explanations in the ACSP, and possibly in other ITS and Intelligent Interfaces. For the ACSP, for instance, we could

- modify the ACSP explanation functionality so that it more proactively encourages users who are known to be low on DS to access explanations.
- investigate what makes low agreeableness users perceive the current ACSP explanations as more distracting and overwhelming, and design a version of the explanations for these users that is modified accordingly.

This personalization, geared toward increasing explanation access and acceptance, will of course be most relevant if further studies to confirm that leveraging explanations is beneficial to improve students' experience with the ACSP applet. Note that information on the relevant individual differences can be collected upfront using the standard tests we used in the study, after which personalization can be enable by setting a related parameter in the ACSP. However, we can also explore the option of predicting these values in real-time from interaction data as students work with the ACSP, as it has been done, for instance, [24].

Due to the complexity of the AI mechanisms underlying the ACSP adaptive hints, we chose to start evaluating explanations that sacrificed completeness to focus on usability and clarity. We found that no participant accessed all the available types of explanations, and none but one participant mentioned wanting more information. This suggests further investigation on the value of having complete explanations, as advocated by [22] when the mechanisms to be explained are exceedingly complex.

The current study cannot provide reliable results on what difference explanations can make, because of too few users not seeing explanations. However, there are some promising trends. As mentioned in Section 5.3, users rated the ACPS hints for usefulness, confusion, distraction and trust. Looking at the percentage difference between the ratings of the 20 users who viewed the explanation and those of the six who did not, most differences are below 10%, except for *confusion*: here users who accessed explanations gave ratings 38% lower than the others. This trend suggests a potential impact of explanations on making the hints more clear. Furthermore, participants who accessed the explanations show a trend of higher learning gains than users who did not (48% vs. 35% average).

7 Conclusions and Future Work

This paper represents a step toward understanding the value of XAI in Intelligent Tutoring Systems. Although there has been research on how to increase ITS transparency via Open Learner Models, thus far work on enabling ITS to provide explicit explanations on the AI underlying their user modeling and decision making has been preliminary at best. The contributions of this paper include:

- An interface enabling incremental access to *why* and *how* explanations for the adaptive hints generated by the ACSP, an ITS that supports learning via an interactive simulation.
- An evaluation for usability and acceptance of the explanations, showing both encouraging results along these dimensions as well as the importance of investigating student individual differences to further their experience with the explanations.

Our results also confirm that some users do not access explanations. Although we uncovered some general reasons for this behavior (not wanting hints in the first place, or feeling that the hints do not need explanations), we plan to collect additional data to perform a formal analysis of which individual differences might cause these reactions, and possibly how to overcome them. We also plan to

conduct a formal user study to compare the effectiveness of the ACSP with and without explanations, in terms of hints perception and follow rate, as well as impact on student learning.

Finally, it is important to remember that the ACSP is designed to be used by learners that have some computer science background, and thus might be more interested in understanding the underlying AI via explanations. It is crucial to investigate explanations in ITS designed to work with less technology-savvy students, as they might generate very different reactions than the ones we observed.

ACKNOWLEDGMENTS

Left blank for submission.

REFERENCES

- [1] Saleema Amershi, Giuseppe Carenini, Cristina Conati, Alan K. Mackworth, and David Poole. 2008. Pedagogy and Usability in Interactive Algorithm Visualizations: Designing and Evaluating CIspace. *Interact. Comput.* 20, 1: 64–96. <https://doi.org/10.1016/j.intcom.2007.08.003>
- [2] Jordan Barria-Pineda, Kamil Akhuseyinoglu, and Peter Brusilovsky. 2019. Explaining Need-based Educational Recommendations Using Interactive Open Learner Models. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP'19 Adjunct)*, 273–277. <https://doi.org/10.1145/3314183.3323463>
- [3] Susan Bull and Judy Kay. 2016. SMILI © : a Framework for Interfaces to Learning Data in Open Learner Models, Learning Analytics and Related Fields. *International Journal of Artificial Intelligence in Education* 26, 1: 293–331. <https://doi.org/10.1007/s40593-015-0090-8>
- [4] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. *International Conference on Intelligent User Interfaces, Proceedings IUI*. <https://doi.org/10.1145/2166966.2166996>
- [5] Andrea Bunt, Joanna McGrenere, and Cristina Conati. 2007. Understanding the Utility of Rationale in a Mixed-Initiative System for GUI Customization. In *User Modeling 2007 (Lecture Notes in Computer Science)*, 147–156.
- [6] John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment* 48, 3: 306–307. https://doi.org/10.1207/s15327752jpa4803_13
- [7] Nico Castelli, Corinna Ogonowski, Timo Jakobi, Martin Stein, Gunnar Stevens, and Volker Wulf. 2017. What Happened in my Home?: An End-User Development Approach for Smart Home Data Visualization. 853–866. <https://doi.org/10.1145/3025453.3025485>
- [8] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. In *presented at 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*. Retrieved September 18, 2019 from <http://arxiv.org/abs/1807.00154>
- [9] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An Intelligible Translation Environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, 524:1–524:13. <https://doi.org/10.1145/3173574.3174098>
- [10] Kelley Cotter, Janghee Cho, and Emilee Rader. 2017. Explaining the News Feed Algorithm: An Analysis of the “News Feed FYI” Blog. 1553–1560. <https://doi.org/10.1145/3027063.3053114>
- [11] Kate Ehrlich, Susanna Kirk, John Patterson, Jamie Rasmussen, Steven Ross, and Daniel Gruen. 2011. Taking advice from intelligent systems: the double-edged sword of explanations. 125–134. <https://doi.org/10.1145/1943403.1943424>
- [12] Ruth B. Ekstrom, Diran Dermen, and Harry Horace Harman. 1976. *Manual for kit of factor-referenced cognitive tests*. Educational testing service Princeton, NJ.
- [13] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6: 504–528. [https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1)
- [14] 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)*, 241–250. <https://doi.org/10.1145/358916.358995>
- [15] Samad Kardan. 2017. A data mining approach for adding adaptive interventions to exploratory learning environments. University of British Columbia. <https://doi.org/10.14288/1.0348694>
- [16] Samad Kardan and Cristina Conati. 2011. A Framework for Capturing Distinguishing User Interaction Behaviors in Novel Interfaces. 159–168.
- [17] Samad Kardan and Cristina Conati. 2013. Comparing and Combining Eye Gaze and Interface Actions for Determining User Learning with an Interactive Simulation. In *User Modeling, Adaptation, and Personalization (Lecture Notes in Computer Science)*, 215–227.
- [18] Samad Kardan and Cristina Conati. 2015. Providing Adaptive Support in an Interactive Simulation for Learning: An Experimental Evaluation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, 3671–3680. <https://doi.org/10.1145/2702123.2702424>
- [19] Todd Kashdan, Melissa Stikma, David Disabato, Patrick Mcknight, John Bekier, Joel Kaji, and Rachel Lazarus. 2017. The Five-Dimensional Curiosity Scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality* 73. <https://doi.org/10.1016/j.jrp.2017.11.011>
- [20] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 411:1–411:14. <https://doi.org/10.1145/3290605.3300641>
- [21] 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 (IUI '19)*, 379–390. <https://doi.org/10.1145/3301275.3302306>
- [22] 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 (IUI '15)*, 126–137.
<https://doi.org/10.1145/2678025.2701399>
- [23] 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.
<https://doi.org/10.1109/VLHCC.2013.6645235>
- [24] L. Küster, C. Trahms, and J. Voigt-Antons. 2018. Predicting personality traits from touchscreen based interactions. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6.
<https://doi.org/10.1109/QoMEX.2018.8463375>
- [25] Poole David L. and Mackworth Alan K. 2010. *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press, New York, NY, USA.
- [26] 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: 55–88. <https://doi.org/10.1007/s11257-016-9186-6>
- [27] Andrew Mabbott and Susan Bull. 2006. Student Preferences for Editing, Persuading, and Negotiating the Open Learner Model. In *Intelligent Tutoring Systems (Lecture Notes in Computer Science)*, 481–490.
- [28] Paul Meara. 1992. EFL Vocabulary Tests.
- [29] 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. 397–407.
<https://doi.org/10.1145/3301275.3302313>
- [30] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. 2018. Argumentation-Based Explanations in Recommender Systems: Conceptual Framework and Empirical Results. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18)*, 293–298.
<https://doi.org/10.1145/3213586.3225240>
- [31] Ingrid Nunes 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: 393–444.
<https://doi.org/10.1007/s11257-017-9195-0>
- [32] Kaśka Porayska-Pomsta and Evi Chryssafidou. 2018. Adolescents' Self-regulation During Job Interviews Through an AI Coaching Environment. In *Artificial Intelligence in Education (Lecture Notes in Computer Science)*, 281–285.
- [33] Julian B. Rotter. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs* 80, 1: 1–28. <https://doi.org/10.1037/h0092976>
- [34] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better Than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*, 240–251.
<https://doi.org/10.1145/3301275.3302308>
- [35] 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 (UMAP '19)*, 22–30.
<https://doi.org/10.1145/3320435.3320465>
- [36] E. K. Vogel, G. F. Woodman, and S. J. Luck. 2001. Storage of features, conjunctions and objects in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance* 27, 1: 92–114.
<https://doi.org/10.1037//0096-1523.27.1.92>
- [37] Michelle Wiebe, Denise Y. Geiskkovitch, and Andrea Bunt. 2016. Exploring User Attitudes Towards Different Approaches to Command Recommendation in Feature-Rich Software. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*, 43–47.
<https://doi.org/10.1145/2856767.2856814>
- [38] Beverly Park Woolf. 2007. *Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing e-Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.