Toward XAI for Intelligent Tutoring Systems: A Case Study

by

Vanessa Putnam

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2020

© Vanessa Putnam, 2020
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis/dissertation entitled:

Toward XAI for Intelligent Tutoring Systems: A Case Study

submitted by Vanessa Putnam in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Examiner committee:

Cristina Conati, Professor, Department of Computer Science, UBC
Supervisor

David Poole, Professor, Department of Computer Science, UBC
Supervisory Committee Member
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.
Lay Summary

Artificial intelligence is the study of and design computational agents that act intelligently [29]. Lack of transparency in many artificial intelligence (AI) techniques has created a growing interest in incorporating explanation in artificial intelligence systems, in order to express an intelligent system’s modeling technique in a way that is interpretable and understandable. This field of research, known as explainable artificial intelligence (XAI for short), aims to make AI techniques more transparent in hope of increasing user trust, and providing users with information that can develop their understanding of an intelligent system’s learning mechanism.

This work is a step towards understanding when and if it is necessary for an Intelligent Tutoring System (ITS), (i.e. a computer system that aims to provide customized instruction or feedback to students), to explain its underlying user modeling techniques to students.
Preface

This master’s thesis is an outcome of a research project done at the University of British Columbia in collaboration with M.Sc. student Lea Reiger, advised by Prof. Cristina Conati. Vanessa Putnam and Lea Rieger contributed equally to all parts of the work presented in this thesis, except for the parts described in the following sections:

4.1, Pilot Study: Vanessa Putnam
4.2, First Design Iteration: Vanessa Putnam
4.4, Navigation: Lea Rieger¹
5.1, Participants and Procedure: Vanessa Putnam
6.3, Analyses on Different Study Elements: Vanessa Putnam
6.4, Impact of Individual Differences on Explanation Access and Ratings: Vanessa Putnam

The following papers were written collaboratively during the course of the research project and can be found in Chapter 4:


¹ Lea Rieger’s thesis contains more work on the gaze analysis that was not covered in this thesis. This analysis can be found in “Explainable AI in Intelligent Tutoring Systems: An Eye-Tracking Assisted Investigation”.

v
Table of Contents

Abstract........................................................................................................................................ iii
Lay Summary ................................................................................................................................... iv
Preface .............................................................................................................................................. v
Table of Contents .............................................................................................................................. vi
List of Tables ...................................................................................................................................... ix
List of Figures .................................................................................................................................... x
List of Abbreviations ........................................................................................................................ xii
Acknowledgements ........................................................................................................................... xiii
Dedication ........................................................................................................................................ xiv

Chapter 1: Introduction ..................................................................................................................... 1

Chapter 2: Related Work ................................................................................................................... 4
  2.1 Mixed Results on the Effectiveness of Explanation ................................................................. 4
  2.2 The Role of Individual Differences in Explanation .............................................................. 5
  2.3 For Whom Explanations are Designed .................................................................................... 7
  2.4 Explanations in Intelligent Tutoring Systems ....................................................................... 7

Chapter 3: The ACSP Applet ........................................................................................................... 9
  3.1 Interactive Simulation for AC-3 ............................................................................................. 9
  3.2 Modeling User Behaviors in the ACSP .................................................................................. 10
  3.3 Adaptive Hints ....................................................................................................................... 13

Chapter 4: Explanation Interface .................................................................................................... 16
  4.1 Pilot User Study ...................................................................................................................... 16
4.2 First Iteration of Explanation Design .................................................. 21
4.3 Design Criteria .................................................................................. 24
4.4 Navigation and Content ..................................................................... 26
  4.4.1 Why am I delivered this hint? ......................................................... 30
  4.4.2 Why am I predicted to be lower learning? .................................. 31
  4.4.2.1 How is my score for each group computed? ............................. 33
  4.4.2.2 How was my hint chosen? ......................................................... 33
  4.4.2.3 How was my hints rank calculated? ......................................... 34
  4.4.3 Why are the rules used for classification? .................................... 34

Chapter 5: User Study ............................................................................... 36
  5.1 Participants and Procedure ............................................................... 36
  5.2 Individual Differences .................................................................... 37
  5.3 Measurements .................................................................................. 38

Chapter 6: Results and Analysis ............................................................. 41
  6.1 Interaction with Explanation Interface ............................................ 41
  6.2 Subjective Ratings ........................................................................... 44
  6.3 Analyses on Different Study Elements ........................................... 49
  6.4 Impact of Individual Differences on Explanation Access and Ratings 50
  6.5 Discussion ....................................................................................... 53

Chapter 7: Conclusions and Future Work .............................................. 56

Bibliography ........................................................................................... 57

Appendices .............................................................................................. 60
  Appendix A Materials Used in Experiments ....................................... 60
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1</td>
<td>Pre and Post CSP tests (Pilot and Full Study)</td>
<td>60</td>
</tr>
<tr>
<td>A.2</td>
<td>Post Study Questionnaires (Pilot Study)</td>
<td>64</td>
</tr>
<tr>
<td>A.3</td>
<td>Post Study Interview (Pilot Study)</td>
<td>67</td>
</tr>
<tr>
<td>A.4</td>
<td>Post Study Questionnaires (Full Study)</td>
<td>67</td>
</tr>
</tbody>
</table>
List of Tables

Table 1: A subset of representative rules for HLG and LLG clusters. ............................................. 11
Table 2: Hint descriptions .................................................................................................................. 12
Table 3: Explanation Questionnaire Items. ......................................................................................... 38
Table 4: Hint Questionnaire Items. ..................................................................................................... 39
Table 5: Summative statistics on usage of the explanation ................................................................. 43
Table 6: Results of the Kendall Rank Test on explanation questionnaire items. Numbers in each box represent the correlation coefficient, * indicates a significance level at .05 and ** indicates a significance level at .01. The table is symmetrical along the diagonal, this is why half of the graph is greyed out. ......................................................................................................................... 48
Table 7: Results of the Kendall Rank Test on hinting questionnaire items. Numbers in each box represent the correlation coefficient, * indicates a significance level at .05 and ** indicates a significance level at .01. The table is symmetrical along the diagonal, this is why half of the graph is greyed out. ......................................................................................................................... 48
List of Figures

Figure 1: The ACSP applet with an example CSP and hint. ................................................................. 10

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. ........................................................................................................ 13

Figure 3: Explanation responses over time. X-axis normalizes session time for all students, broken up into quartiles. Y-axis is a ratio of reports for each type of explanation response over the total number of responses for each session quartile................................................................. 18

Figure 4: First iteration of explanation design. Explanation beginning with the ACSP behavior discovery phase and ending with the hint that was delivered. ......................................................... 21

Figure 5: The explanation within each of the circles in Figure 4. ......................................................... 23

Figure 6: 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) How was this score computed? (E) How was this specific hint chosen? Page: “How was my hints rank calculated” not shown (See arrow from (E)). ................................................................. 27

Figure 7: Breadcrumb Navigation ........................................................................................................... 29

Figure 8: Graph Navigation .................................................................................................................. 29

Figure 9: Number of hints, number of why page accesses, and number of how page accesses per participant. The x axis denotes each participant (i.e. Pij) and y axis denotes the number of number of hints, number of why page accesses, and number of how page accesses. ................. 42

Figure 10: Proportion of time spent in and number of accesses for each type of explanation page ............................................................................................................................................. 44

Figure 11: Subjective ratings of the explanation. .................................................................................. 45
Figure 12: Subjective ratings on hints for all 26 users who received hints. ................................. 47

Figure 13: Average hint ratings for users of both groups (i.e., users who chose to view the explanation and users who did not chose to view the explanation). .......................................................... 47

Figure 14: Blue boxes represent different elements of our study. Solid lines with numbers indicate analyses we tried on different elements, and dotted lines indicate analyses for future work. The direction of each arrow illustrates the potential effect(s) of one element on another. 49

Figure 15: 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. ........................................................................................................ 51

Figure 16: Distribution of combined ratings for distraction, confusion, and overwhelming, for high and low agreeableness split by the median Agreeableness value. .................................................. 52
List of Abbreviations

- ACSP - Adaptive Constraint Satisfaction Problems
- ML - Machine Learning
- PLG - Performance Learning Gain
- RS - Recommender System
- XAI - Explainable Artificial Intelligence
Acknowledgements

I would like to thank the University of British Columbia’s Computer Science Department for allowing me to research my master’s thesis and providing me with a state of the art education in Computer Science. Particularly, I would like to thank my supervisor Dr. Cristina Conati for providing me with the help and guidance I needed to conduct this research. I would also like to thank my friends Sarah Espinosa and Zoe Sherman for their support along the way. Sarah, thank you for answering my late night phone calls and providing me with pep talks on request. Zoe, thank you for always reminding me of my worth and teaching me the importance of sticking up for myself. This work would not have been possible without the both of you.

Lastly, but definitely not least, I would like to thank Lea Reiger for being the other half of this project, and the greatest partner I could have asked for during this research. I am grateful for you hard work, insight, attention to detail, and most of all our friendship. I could not have done it without you.
Dedication

I dedicate this work to my family for their love and support through all of the ups and downs of my studies. I would like to thank my grandmother for encouraging me to continue higher education after my undergraduate studies and teaching me the importance of hard work to accomplish my goals. I would like to thank my grandfather for sparking my interest in math at a young age, and teaching me how to get back up after I fall. Lastly, I would like to dedicate this work to my fiancé, Alexander, thank you for always believing in me and supporting me no matter what. Also, thank you for cleaning and cooking for me when I was too busy to do it myself. You are the strongest pillar that held me up during this process, and this work would not have been possible without you. I love you.
Chapter 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 [7][13][20]. However, there are also results suggesting that such explanations are not always wanted by or beneficial for all users [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 [13][21][26][28][37], office assistants [5], and intelligent everyday interactive systems (e.g., 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 [7].
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., [23]) and learning abilities (e.g., ability to self-assess [30]). There is also anecdotal evidence that an OLM can impact students’ trust [24].

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][13][20]), 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) [26] and of user decision-making style (e.g. rational vs. intuitive) [27] on explanations in recommender systems, and of perceived user expertise for explanations of an intelligent assistant [33]. 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, 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.
Chapter 2: Related Work

2.1 Mixed Results on the Effectiveness of Explanation

There are encouraging results on the helpfulness of explanations in intelligent user interfaces. For example, Kulesza et al. [20] investigated explaining the predictions of an agent that helps its users organize their emails. They showed that explanations helped participants understand the system’s underlying mechanism, enabling them to provide feedback to improve the agent’s predictions. Coppers et al. [8] 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., [6][21][28][32]). For instance, Kulesza et al. [21] 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. Chang et al. [6] incorporated natural language explanations into a movie recommender by mining quotes from movie reviews and explaining recommendations in a way that mimics word-of-mouth recommendations while still modeling a users’ preferences. An example recommendation of this type might be, “From your profile you prefer movies tagged as intense, the movie a pretty intense ninety minutes, with Bullock's character constantly battling one catastrophe after another, and all of it is amazing to see.” They compared these explanations based on natural language quotes against state-of-the-art content-based explanation [35] (e.g. “We recommend the movie because you like the following features: [tag1, ..., tag5]”), and found that users perceived the natural language explanations as more trustworthy, contain a more appropriate amount of information and offer a better user experience. This suggests that even
though baseline explanations have already proven to be beneficial, there is still room to enhance explanations further.

There is, however, also research showing that explanations might not always be useful or wanted. Herlocker et al. [13] evaluated an explanation interface for a 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.

2.2 The Role of Individual Differences in Explanation

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. [9] 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. [19] 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 [34] 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. Arnold et al. [1] found that novice and expert users express preferences for different types of explanations in a knowledge based expert insolvency system. Results indicate that novices will have stronger preferences for feedforward (i.e. declarative explanations or explanations about how inputs in the KB system are used in terms of relevant information). Experts, on the other hand, will have a greater interest in feedback explanations (i.e. procedural explanations or explanations about how a decision was made and what information lead to the current decision) that, as the authors claim, are the type that have generally been available in prior experimental studies.

Going beyond user preferences, Millecamp et al. found moderating effects of Need for Cognition (a personality trait) 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 [26]. Dodge et al. [10] conducted an empirical study to understand how different explanations impact people’s fairness judgments of machine learning (ML) systems. Authors show that users prior positions of algorithmic fairness impact how people react to different styles of explanation. More specifically, people who perceive machine learning systems as fair gain an even higher confidence in the fairness of a prediction given a global explanation (i.e., explanations that describe how the system works overall) compared too a local explanation (i.e. explanations that describe the systems decision for a specific input or outcome). Naveed et al. [27] 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. Shaffer et al. [33] 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.

2.3 For Whom Explanations are Designed

Hohman et al. [14] surveys different interrogative questions (i.e., Why, Who, What, How, When, and Where) for visualizing deep learning models (i.e. a form of explaining a deep learning model to users). This work highlights an important question for designers of explanations to ask, “Who are the types of people and users that would use and stand to benefit (from an explanation)?”. The survey identifies three types of users: model developers (individuals whose job is primarily focused on developing, experimenting with, and deploying the model), model users (users who may have some technical background about the system but are novices to the underlying system’s model), and non-experts (individuals who typically have no prior knowledge about the system’s model). For the purpose of this thesis’s explanations, we design our explanation for the third category of users: non-experts. This is because students will be the primary users of our ITS and explanations, and do not have any prior knowledge on the mechanisms that drive the ACSPs intelligent behavior.

2.4 Explanations in Intelligent Tutoring Systems

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 Chryssafidou [30] 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 [23] 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 [24], 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 mechanisms
Chapter 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 [29]. 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, 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.

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 [16][15]). The next two sections summarize these mechanisms since they are the targets of the explanations that we added to the ACSP.

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

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 clustering algorithm then groups these 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 \[16][15\]. 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.

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

<table>
<thead>
<tr>
<th>Rules for HLG cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1: Infrequently auto solving the CSP</td>
</tr>
<tr>
<td>Rule 2: Infrequently auto solving the CSP and infrequently stepping through the problem</td>
</tr>
<tr>
<td>Rule 3: Pausing for reflection after clicking CSP arcs</td>
</tr>
</tbody>
</table>

| 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 2: Hint descriptions

<table>
<thead>
<tr>
<th>Hint Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Direct Art Click more often;</td>
</tr>
<tr>
<td>Spend more time after performing Direct Arc Clicks;</td>
</tr>
<tr>
<td>Use Reset less frequently;</td>
</tr>
<tr>
<td>Use Auto Arc-consistency less frequently;</td>
</tr>
<tr>
<td>Use Domain Splitting less frequently</td>
</tr>
<tr>
<td>Spend more time after performing Fine Steps;</td>
</tr>
<tr>
<td>Use Back Track less frequently</td>
</tr>
<tr>
<td>Use Fine Step less frequently</td>
</tr>
<tr>
<td>Spend more time after performing reset for planning;</td>
</tr>
</tbody>
</table>

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.

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.

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 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. It is important to note that a users classification can change at any time depending on their behaviors during interaction.

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.
Chapter 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. In order to ensure participants understood the aforementioned explanation types we asked participants to give a verbal explanation, to the best of their ability, on how the system provides hints and why the system provides hints in our interview. From this feedback we determined that 5 of the 8 participants understood the distinction between a how type and why type explanation.

We ran a pilot with nine university students with adequate prerequisites (i.e. currently enrolled in the university undergraduate introduction to AI course) 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. The procedure for our study was as follows: (1) students studied a textbook chapter on the AC-3 algorithm; (2) wrote a pre-test on the concepts covered in the chapter; (3) watched an introductory video on how to use the main functionalities of the ACSP applet; (4) used the ACSP applet to solve two CSPs; (5) took a post-test analogous to the pre-test (See Appendix Chapter 7:A.1); and (6) answered a post-questionnaire (See Appendix Chapter 7:A.2) and a follow up interview (See Appendix Chapter 7:A.3) that solicited feedback on the explanations they selected for “explain hint” in the dialogue box.
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 Cotter et al. [9] in terms of users preferring why explanations, followed by how explanations.

We also analyzed the temporal pattern of students’ responses regarding explanations, to uncover when students are more likely to need or not need explanation during interaction. Specifically, we normalized the session time for all students and broke the session up into quartiles (i.e. 1/4, 1/2, 3/4, 4/4). Then we took a ratio of the responses for why, how, other explanation and no explanation over the total number of responses at that time. The results of this analysis are summarized in Figure 3, indicating that students want to know why more during the first half of their interaction (i.e., first and second quartile). A possible design direction for explanation would mean explaining why earlier in the student’s interaction with the applet. Figure 3 also shows a tendency increasing with time for participants to explicitly declare that they do not want explanation in the “explain hint” dialogue box. We do not attribute the decrease in wanting explanation to be because participants know they will not be getting any explanation. If this were the reason, participants would not access the “explain hint” functionality a second time after already discovering the ACSP had no explanation after the first hint. This indicates that participants are more interested in explanations earlier in their interaction and may find them less necessary closer to the end of their interaction.
Figure 3: Explanation responses over time. X-axis normalizes session time for all students, broken up into quartiles. Y-axis is a ratio of reports for each type of explanation response over the total number of responses for each session quartile.

Finally, we analyzed the open-ended feedback from our follow-up questionnaire and interview to understand the reasons participants want explanations for system hints. From this source, multiple participants expressed that they need explanations and consider them valuable when they are curious, and when they disagree with the system’s decision making. More specifically, one student stated they wanted explanations because they were curious to know how the hint was created. We use this finding later in our full study to evaluate curiosity as an individual difference (see Section 5.2). Other students expressed that they wanted explanation when they disagreed with the system. One student expressed value in an explanation that would justify a student’s decision to ignore system hints, claiming:
Some students learn at different paces. So for a student that learns quickly, a reasoning behind a slow down hint may allow the student to see the reasoning and know ‘oh this does not apply to me’ and they can take note of that.

This is an interesting finding since it suggests that explanations may be needed when students feel the systems hints are not useful, justified, or do not apply to them.

Participants also expressed other reasons why explanations would be valuable to them. One participant expressed that wanting to know the system’s decision making was important to them. This implies that explanations for system hints would be necessary for participants who want to know the systems reasoning. Additionally, the same participant that suggested that a type of explanation that specifies which options the applet would like him to try would be valuable because the explanation could guide him to trying different ways of solving the problem. This response expresses why explanations may be specifically valuable to ITSs, and reinforces the idea that incorporating explanation to the ACSP may be a feature students would like to see.

Table 3: Pilot participants reasons for wanting explanation.

<table>
<thead>
<tr>
<th>Themes</th>
<th>Participants</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curiosity</td>
<td>P4; P7</td>
<td>Participants were curious why the system told them to adopt a new behavior or told them to slow down.</td>
</tr>
<tr>
<td>Disagree with the systems decision making</td>
<td>P1; P2; P10</td>
<td>Participants wanted to utilize the explanation when they did not agree with the systems decision making. They expressed situations when they felt a given hint was not appropriate for them or when they were told to stop performing an action.</td>
</tr>
<tr>
<td>Want to know the system’s decision making</td>
<td>P2</td>
<td>Participant wanted to know how the system works in</td>
</tr>
</tbody>
</table>
Other participants addressed situations when they needed an explanation during their interaction. One participant claimed explanation was needed for a hint when it was first delivered. This is useful information because it may be a reason for the large number of reports wanting to know why in the first quartile of Figure 3. Another participant expressed that they wanted an explanation for why a hint was delivered at that moment in time and not earlier, indicating that explaining the timing of a hint is also important to students.

Apart from participants who expressed positive feelings toward explanation, two out of the eight participants responded that they did not experience a situation when explanation was necessary. This indicates that not all participants needed or wanted explanation during their interaction and supports our future work investigating if and how individual student differences influence the effect of explanations. The remaining participants not accounted for in this section either did not answer the question or gave an answer that was not interpretable. For these participants we did not add their responses to our analysis. Table 3 accounts for the different theme’s participants expressed in their responses as well as the individual participants that expressed those themes.
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 Chapter 3.

4.2 First Iteration of Explanation Design

**Explanation for System Hints.**

![Diagram of explanation design](image)

**Figure 4: First iteration of explanation design. Explanation beginning with the ACSP behavior discovery phase and ending with the hint that was delivered.**

The first iteration of our explanation design was based on the assumption that we could just generate an explanation for “why” or “how” a user has been delivered a specific hint by simply explaining in sequence each phase of the user modeling framework, as in Figure 2, from beginning (i.e., behavior discovery) to end (i.e., adaptive hints). In this design, once a hint is delivered, all of Figure 4 would be available to the user. Each circle represents an element of the user modeling framework in Figure 2. Users can click on each circle to view more details, this is shown in Figure 5 where each numbered box corresponds to the explanation for each of the higher level circles (1-6) in Figure 4. Each level of the explanation contains a progress bar to indicate how far along the user is in the explanation.
For each user who worked with the ACSP applet, it stores a summary of how they used all the available actions.

Here are the two groups that were generated.

High Learning Group
Low Learning Group

Here are some sample rules.

High Learning Group
Low Learning Group

Infrequently solving the CSP
Frequently reating the CSP

The size of each circle denotes the weight (importance) of the rule in that group.
Learning groups, rules, and rule weights are used to create a classifier that will predict at any given point in the interaction if you are a high or low learner based on your interaction behaviors so far.

Your classification score for each learning group:
- Low Learning Score: 60
- High Learning Score: 40
You are a member of the **Low Learning Group**

A list of hints are generated based on the rules you have satisfied for your learning group. Each hint corresponds to a specific behavior or action that could appear in one or more rules.

- **Using Auto Arc Consistency Least Frequently**
  - Frequently Backtracking and Auto solving
  - Frequently Auto Solving

Figure 5: The explanation within each of the circles in Figure 4.
As noted previously, this design explains a hint starting from the very beginning of the process, namely from the behavior discovery phase. Essentially, this design relies on having a one-to-one mapping between elements in the graph and explanation pages. Although this is a faithfully rendition of the process with which the ACSP hints are generated, and the diagram in Figure 4 can be a useful high level explanation of the process, we quickly realized that it would not be suitable for a user who would like to know more details of why or how a hint was delivered, because it would result in explanations that are too fragmented. With this explanation approach, the user needs to view the details of each step in order to understand the subsequent steps in Figure 4 until the question of “why was I delivered this hint” is answered. From a usability perspective this design can be confusing if the user does not finish the whole explanation, or unnecessarily time consuming if the user does not wish to view all of the steps. Thus, we decided to investigate an approach that no longer relies on having a one-to-one mapping between explanation pages and the individual steps in Figure 4, but collapses some of them to facilitate starting the explanation with a direct answer to the question of why a given hint was delivered, and enable users to incrementally access more details if they so wish. The following sections describe the design criteria we used to determine our final explanation design (4.3), and the final explanation functionality that we generated.

4.3 Design Criteria

As guidance for the explanation design, we rely on some of the criteria articulated by Kulesza et al. [20]. 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 Marvel\(^2\) 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.

4.4 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 6. Each tab displays a why explanation; for one of these why explanations (tab in Figure 6(B)), we allow users to ask for more details on how three specific aspects were computed (Figure 6 (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 6, 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.

\(^2\) https://marvelapp.com
Figure 6: 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) How was this score computed? (E) How was this specific hint chosen? Page: “How was my hints 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. 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.

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 6(A). The following three subsections describe the why explanations provided in the three tabs, as well as any how explanation that can be requested from there.

We designed our explanations to be personalized and dynamically update according to the user’s real-time interaction. This is done by logging the users interaction data (i.e. creating a feature vector) and using this data to deliver an appropriate hint, according to the user modeling mechanism described in Section 3.2. As the user interacts, more actions are added to the users feature vector and the system’s user model and classification change overtime. As described in section 3.3 this change will cause the type of hint delivered to change as well. To ensure a dynamic explanation that is personalized for the user, this change is also made within the explanation by querying the user model in real-time. More specifically, we extract personalized data on the user’s current classification, the group scores, the satisfied rules, the hint ranking, and the actions a user has performed.

It is important to note that we settled on this tab-based navigation for our explanation after a few different design iterations. Figure 7 (i.e., the breadcrumb navigation) and Figure 8 (i.e., the graph navigation) are previous navigation designs that we evaluated with three pilot users.
Figure 7: Breadcrumb Navigation

Figure 8: Graph Navigation
In the bread crumb navigation users navigated through the explanation while tracking where they are in the navigation using bread crumbs (i.e., like a trail of bread crumbs left by Hansel and Gretel in the German fairy tale). The user also has the ability to visit different parts of the explanation by selecting any of the titles in the breadcrumb heading. In the graph navigation the graph at the top of the page indicates where in the explanation the users are, and as with the breadcrumbs users are able to visit different parts of the explanation by clicking in different titles of the graph. We piloted these three navigation designs (tab, bread-crumb and graph) with three pilot users and in the end the tab navigation was chosen. Two of the pilot users expressed the breadcrumb navigation was not easy to notice and they lost track of where they were in the explanation. All three of the pilots expressed that the graph navigation was not preferred because it was not clear what the graph at the top of the explanation was communicating. Overall all three participants expressed that they preferred the tab navigation because this is the type of navigation that they were all most familiar with.

The following sections include an exemplary explanation for a hint stating, “You have used the Reset button excessively. I recommend that you limit your usage of this action.”

4.4.1 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.

### 4.4.2 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 (See below).

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 6(B), and resulting pages Figure 6(D) and Figure 6(E) respectively. Their content is presented below.
4.4.2.1 How is 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 = 0.313

4.4.2.2 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 Figure 6(E), the user can navigate further to read more about how their hint’s rank was computed (button Figure 6(E) bottom).
4.4.2.3 How was my hints 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)

4.4.3 Why are the rules used for classification?

Selecting the third tab in the interface Figure 6(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 Chapter 5.
Chapter 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 [16]. Participants were told that the ACPS 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 [25]) to uncover differences in users abilities to process the diagrams and text in the explanation. We also measure users’ Locus of Control [31] 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 [26]
• 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 [19].
Two dimensions of Curiosity [17]: 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 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 6); thus, the number of explanation type accessed ranges from 1 to 6.

We also collected subjective feedback on the explanation functionality (see Table 4).

Table 4: Explanation Questionnaire Items.

<table>
<thead>
<tr>
<th>Items on Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would choose to have the explanations again in the future.</td>
</tr>
<tr>
<td>I am satisfied with the explanations.</td>
</tr>
<tr>
<td>The explanations were helpful for me.</td>
</tr>
</tbody>
</table>

3 The test here measures three other dimensions of curiosity not relevant to our context
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 5: Hint Questionnaire Items.

I would choose to have the hints again in the future.
I am satisfied with the hints.
The hints were helpful for me.
I understand why hints were delivered to me in general.
I understand why specific hints were delivered to me.
I trust the system to deliver appropriate hints.
Given my behavior, I agree with the hints that were delivered to me.
The hints distracted me from my learning task.
The hints were confusing.
The hints appeared at the right time.

The items in Table 4 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], [26], [18]). 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 [16]. We added an item for overwhelming because it is one of the specific design criteria for explanations in [20]. 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.

To see both full questionnaires please refer to Appendix Chapter 7:A.3.
Chapter 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) 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.

4 Difference between post-test score and pre-test score over the difference between the tests’ maximum score and pre-test score.:
Figure 9 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 9: Number of hints, number of why page accesses, and number of how page accesses per participant. The x axis denotes each participant (i.e. Pij) and y axis denotes the number of number of hints, number of why page accesses, and number of how page accesses.](image)

Participants received on average 2.7 hints, with large standard deviation (2.4) and range (minimum of one hint and maximum of 11). Table 6 provides detailed summative statistics on how participants approached the explanation interface. The first two rows concern how participants initiated explanation in response to hints. Participants tended to initiate the explanation on the first hint, or the second at the latest (second row). The ratio of explanation initiations over the number of hints (third 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.

The last three rows in Table 6 give a sense of how much participants actually explored the explanation interface. An average of almost 3 distinct 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.

Table 6: Summative statistics on usage of the explanation

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hints before first explanation initiation</td>
<td>1.10</td>
<td>0.31</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Explanation initiations over number of hints received</td>
<td>0.76</td>
<td>0.30</td>
<td>0.25</td>
<td>1.00</td>
</tr>
<tr>
<td>Number of pages accessed per initiation</td>
<td>2.95</td>
<td>1.36</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Total time spent in explanation</td>
<td>66.2s</td>
<td>55.5s</td>
<td>5.4s</td>
<td>191.6s</td>
</tr>
<tr>
<td>Distinct types accessed</td>
<td>2.80</td>
<td>1.24</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Going into more detail, Figure 10: Proportion of time spent in and number of accesses for each type of explanation page) 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 6A and Figure 6B), makes up about two thirds of total accesses or duration. The third type of Why explanation (WhyRules, Figure 6C) 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 6D and Figure 6E), 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 6F). Only one participant made use of the “I would have liked to know more” button, wishing for more details on the rules.

![Proportion of explanation pages accessed](image)

![Time spent in explanation page](image)

**Figure 10:** 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 4) reveals that users were in general positive about it. This can be seen in (Figure 11(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 11 (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 11 (C)), suggesting a strong usability.
Figure 11: Subjective ratings of the explanation.

Next, we analyze the questionnaire items on the systems hints (Figure 12) for all 26 participants (i.e., 20 users who chose to view the explanation and 6 users who did not choose to view the explanation). Participants reported a neutral rating for intention to use (int) [mean: 3.38, median: 3, mode: 3], helpfulness (help) [mean: 2.92, median: 3, mode: 3], and satisfaction (sat) [mean: 3.19, median: 3, mode: 3] with the hints. This is shown not only by participants average rating of each item but also in the most frequently reported rating for each (i.e. each item having a mode of 3). In comparison, the ratings for general understanding of why hints are delivered (gen und) [mean: 4.15, median: 4, mode: 5] and why specific hints were delivered to them (spec und) [mean: 4.07, median: 5, mode: 5], are quite high (numbers). Ratings of trust (trust) [mean: 3.65, median: 4, mode: 4] as well as agreement with the system hints (agree) [mean: 3.5, median: 4, mode: 5] are reported to be neutral but unlike ratings for int, help, and sat the medians and modes for both items are both 4 or higher. Also, average ratings on agreement have a high standard deviation, suggesting although many participants agree with the hints that were delivered to them (as shown by the mode of 5 for this item), some highly disagree. This suggests that many users trusted and agreed with the systems hints, but there is still room for improvement for others who did not feel the same. As with the explanation, the low ratings for distraction (dist) [mean: 2.5, median: 2, mode: 2] and confusion (conf) [mean: 2.23, median: 2,
mode: 1] also speak in favor of the hints. These ratings have a high standard deviation indicating users have split opinions on the confusion and distraction of the system hints. However, taking a closer look at the mode’s for these items indicate that the majority felt that they disagree that the hints are distracting, and strongly disagree that the hints are confusing. Lastly, a neutral average rating for hint timing (time) [mean: 3, median: 3, mode: 2] but a mode of 2 (disagree) indicates that most users are unsatisfied with timing of hints.

We are also interested in comparing the differences between the two groups. In Figure 13 we analyze the difference between these average ratings for people who have and have not seen the explanation. It is important to note that these findings are preliminary and are subject to change since our group of users who did not view the explanation is small in comparison to the group of users who chose to view the explanation. If we look at the percentage difference of the ratings provided by the 20 participants who viewed the explanation against the ratings of the six users who did not, most differences are small (at or below 10%) with the exception of the difference in the ratings on an item asking whether the hints are confusing: users who accessed explanations gave ratings 38% lower than the others for this item. This is a positive observation because it indicates that the presence of explanation could be an important factor for making sure users are not confused by the system’s hints. Further analyses should be conducted to compare the two groups to ensure this is a plausible conclusion for the difference in ratings. For analyzing the remaining ratings between the two groups we do not report any other observations because the difference in ratings are too small. Future studies should be conducted with larger and more balanced groups to evaluate if viewing explanations influence users rating of the system hints.
Figure 12: Subjective ratings on hints for all 26 users who received hints.

Figure 13: Average hint ratings for users of both groups (i.e., users who chose to view the explanation and users who did not choose to view the explanation).

We ran a Kendall Rank correlation analysis over the questionnaire responses (see Table 7 and Table 8) as a sanity check to identify any sets of questions whose answers were inconsistent or contradictory. For example, responses to a negatively phrased measure rating explanation distraction (“The hints distracted me from my learning task.”) and a positively phrased measure rating explanation helpfulness (“The hints were helpful for me.”) should correlate negatively; otherwise, this would mean that whenever participants rated the explanations as very distracting they also rated the explanation as very helpful, indicating that the questions we asked our participants were unclear. Our sanity check was successful in that all measures from Tables 6
and 7 show correlations with the expected directionality (expect for 5 indicated in red in Table 7.
However, for these the correlation values are extremely low and not significant (blue boxes
indicating a .05 significance level and red boxes indicating a .01 significance level).

Table 7: Results of the Kendall Rank Test on explanation questionnaire items. Numbers in
each box represent the correlation coefficient, * indicates a significance level at .05. Green
cells indicate a positive correlation and orange cells indicate a negative correlation.

<table>
<thead>
<tr>
<th></th>
<th>Usage</th>
<th>Satisfaction</th>
<th>Help</th>
<th>Distract</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction</td>
<td>.34</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Help</td>
<td>.323</td>
<td>.620*</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distract</td>
<td>-.264</td>
<td>-.014</td>
<td>-.106</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Conf</td>
<td>-.180</td>
<td>-.558*</td>
<td>-.242</td>
<td>.381</td>
<td>X</td>
</tr>
<tr>
<td>Overwhelm</td>
<td>-.154*</td>
<td>-.195</td>
<td>-.203</td>
<td>.403*</td>
<td>.221</td>
</tr>
</tbody>
</table>

Table 8: Results of the Kendall Rank Test on hinting questionnaire items. Numbers in each
box represent the correlation coefficient, * indicates a significance level at .05. Green cells
indicate a positive correlation and orange cells indicate a negative correlation. Red cells
indicate that the directionality of the correlation is not in the expected direction.

<table>
<thead>
<tr>
<th></th>
<th>Satisfy</th>
<th>Help</th>
<th>Gen Under</th>
<th>Spec Under</th>
<th>Trust</th>
<th>Agree</th>
<th>Distract</th>
<th>Conf</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfy</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Help</td>
<td>.065</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gen</td>
<td>-.031</td>
<td>-.199</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Under Spec Under</td>
<td>Trust</td>
<td>Agree</td>
<td>Distract</td>
<td>Conf</td>
<td>Timing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>-------</td>
<td>-------</td>
<td>----------</td>
<td>------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.063</td>
<td>.171</td>
<td>.463*</td>
<td>.007</td>
<td>.210</td>
<td>.449*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.071</td>
<td>-.083</td>
<td>.242</td>
<td>.210</td>
<td>.449*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.049</td>
<td>-.453*</td>
<td>.242</td>
<td>.210</td>
<td>.449*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.217</td>
<td>-.072</td>
<td>-.419*</td>
<td>-.372</td>
<td>-.255</td>
<td>-.371</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.124</td>
<td>-.259</td>
<td>.130</td>
<td>.184</td>
<td>.387*</td>
<td>-.079</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.081</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[X\]

### 6.3 Analyses on Different Study Elements

![Diagram](image.png)

Figure 14: Blue boxes represent different elements of our study. Solid lines with numbers indicate analyses we tried on different elements, and dotted lines indicate analyses for future work. The direction of each arrow illustrates the potential effect(s) of one element on another.
One of the long-term goals of our research is to understand how individual differences impact how a user interacts, perceives and benefits from explanations. A complete analysis to achieve this understanding would include analyzing a possible direct relation between user differences and 1) usage of the explanation functionality and 2) explanation ratings, as well as the structural relations of these variables with hint usage (follow rate), hint perception, and learning outcomes. These variables along with possible structural relations among them are shown in Figure 14. Unfortunately we do not have enough data to run a complete structural equation analysis on the model in Figure 14, thus for this thesis we chose to focus on testing a subset of the relations in Figure 14. In solid numbered arrows between variables indicate the analyses we performed. For example, in analysis (1) and (2) we investigate the effect of individual differences on users’ usage of the explanations and on explanation ratings. We will discuss the findings from analyses (1) and (2) in the next sections. For analyses (3) and (4) we investigated the effect of individual differences and explanation usage on perceived user trust of hints. We chose trust because explanations have shown to influence user trust in AI systems [21] [24]. We performed analysis 3 by running separate Kruskal-Wallis tests with each of our individual differences (i.e., connection (3)) as independent measures and user hint ratings of trust as the dependent measure. Similarly, we performed analysis 4 by running an additional Kruskal-Wallis with explanation usage (i.e., connection (4)) as the independent measures and user hint ratings of trust as the dependent measure. We found no significant effects in either analysis.

6.4 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 6, rows 4 and 5) 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-variates 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 15: 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*, η² = .310, F(1,18) = 8.085). Users high in DS accessed more pages per initiation than users who are low in deprivation sensitivity (Figure 15(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 ($p = .081$, $\eta^2 = .159$, $F = 3.441$), with users high in DS showing a trend of higher time than users low in DS (see Figure 15(B)).

We also checked for possible impacts of individual differences on user ratings in the explanation questionnaire in Table 4. To do so, we ran independent samples Kruskal-Wallis tests on 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 tests 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.

![Figure 16: Distribution of combined ratings for distraction, confusion, and overwhelming, for high and low agreeableness split by the median Agreeableness value.](image)

Figure 16: Distribution of combined ratings for distraction, confusion, and overwhelming, for high and low agreeableness split by the median Agreeableness value.
6.5 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 student’s 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 traits 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, [22].

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 [20] 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).
Chapter 7: Conclusions and Future Work

This thesis 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.
Bibliography


Interaction. 27, 3–5 (Dec. 2017), 393–444. DOI:https://doi.org/10.1007/s11257-017-9195-0.


Appendices

Appendix A  Materials Used in Experiments

A.1   Pre and Post CSP tests (Pilot and Full Study)

CSP Post-Test

*If you could not finish the question, please state why (i.e. not enough time, didn’t understand question, etc)*

* If you need extra room write on the back of the test*

1. Consider the following constraint network.
a) Is this constraint network arc consistent?

b) If it is, explain why the constraint network is arc consistent. If it isn't, make the network arc consistent and give all solutions.

2. Consider the following constraint network: Note that \((X+Z) \mod 2 = 1\) means that \(X+Z\) is odd.

a) Is this constraint network arc consistent?

b) If it is, explain why the constraint network is arc consistent. If it isn't, explain why the constraint network is not arc consistent AND state which arcs are not arc consistent.

3. Consider the problem of scheduling four activities labeled \(\{W, X, Y, Z\}\). The domain of each activity is \(\{1, 2, 3\}\). Suppose that the constraints on scheduling are as follows: \(W > X, W < Z, W = Y, Y = X\) and \(Z \neq 2\).
a) Draw the initial constraint network, including domains of all nodes and arcs representing all binary relations.

b) Make the network arc consistent and give all solutions.

/4. Consider the following constraint network. Note that \((X+Y) \mod 2=1\) means that \(X+Y\) is odd.

a) Make the constraint network arc consistent. For each domain element removed from a node, identify the constraint responsible for the removal, i.e., you have to write a sequence of steps of the form “val1,…,valK removed from VarX because of constraintC”.

b) Why is domain splitting necessary for this CSP?

c) Show how domain splitting can solve this problem. Choose to split on X, and then find all solutions to the network. Write out the steps as you did for part a above, and include steps for domain splitting and backtracking where necessary.
CSP Pre-Test

*If you could not finish the question, please state why (i.e. not enough time, didn’t understand question, etc)*

* If you need extra room write on the back of the test*

1. Consider the following constraint network: Note that \((X+Y) \mod 2 = 1\) means that \(X+Y\) is odd.

   a) Is this constraint network arc consistent?

   b) If it is, explain why the constraint network is arc consistent. If it isn't, explain why the constraint network is not arc consistent AND state which arcs are not arc consistent.

2. Consider the following constraint network.

   /

   a) Is this constraint network arc consistent?
b) If it is, explain why the constraint network is arc consistent. If it isn’t, make the network arc consistent and give all solutions.

3. Consider a scheduling problem with four activities labeled \{A, B, C, D\}. The domain of each activity is \{2, 3, 4\}. Suppose that the constraints on scheduling are as follows: \(A \neq B\), \(C < A\), \(A < D\), \(B = D\) and \(C < D\).

   a) Draw the initial constraint network, including domains of all nodes and arcs representing all binary relations.
   b) Make the network arc consistent and give all solutions.

4. Consider the following constraint network. Note that \((X+Y) \ mod \ 2 = 1\) means that \(X+Y\) is odd.

   /

   a) Make the constraint network arc consistent. For each domain element removed from a node, identify the constraint responsible for the removal, i.e., you have to write a sequence of steps of the form “\(val_1, …, val_K\) removed from VarX because of constraintC”.
   b) Why is domain splitting necessary for this CSP?
   c) Show how domain splitting can solve this problem. Choose to split on X, and then find all solutions to the network. Write out the steps as you did for part a above, and include steps for domain splitting and backtracking where necessary.

A.2 Post Study Questionnaires (Pilot Study)

1. For the following statements, rate your agreement or disagreement & try to explain your answer: (check a box for each row)
The applet helped me learn the material from the book.

Please Explain:

The book alone would have been enough to learn the material.

Please Explain:

I liked studying with the applet.

Please Explain:

2. For each pair of adjectives, check one box that reflects the extent to which you believe the adjectives describe the applet *(please read adjectives carefully)*.

<table>
<thead>
<tr>
<th>Confusing</th>
<th>Clear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boring</td>
<td>Exciting</td>
</tr>
<tr>
<td>Pleasing</td>
<td>Annoying</td>
</tr>
</tbody>
</table>
3. How would you rate your level of confidence after the study on each of the topics below: (circle a number for each topic)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Poor</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Variable domains</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Constraints</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Constraint Satisfaction Problem 1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Definition of arc consistency 1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Arc consistency AC-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Domain splitting</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Backtracking</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

4. For the following statements, rate your agreement or disagreement & try to explain your answer: (check a box for each row)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Agree</th>
<th>Somewhat Agree</th>
<th>Neutral</th>
<th>Somewhat Disagree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would like to know <strong>why</strong> the system makes adaptive hints.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Please explain or give examples:*
I would like to know **how** the system makes adaptive hints.

*Please explain or give examples:*

5. When using the system, was there a situation when you needed explanation for adaptive hints? Please explain based off your experience.

**A.3 Post Study Interview (Pilot Study)**

Was it clear in the study that the CSP applet provides hints personalized for you?

Recall the systems “Explain Hint” feature. Was it clear that we were looking for your suggestions on which explanations you would like to see for the hints?

Can you please elaborate on what explanations you would like to see for the system’s hints?

Why would these explanations be valuable to you?

Please give a verbal explanation (to the best of your ability) of why you **think** the system provides hints?

Please give a verbal explanation (to the best of your ability) of how you **think** the system provides hints?

**A.4 Post Study Questionnaires (Full Study)**

Explanation Questionnaire

PID: ______________
**ACSP Hints:** Please rate the degree to which you agree or disagree with each of the following statements about the ACSP's hints, where: 1= Strongly Disagree, 2= Disagree, 3=Agree nor Disagree, 4= Agree, 5= Strongly Agree.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would choose to have the hints again in the future.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I am satisfied with the hints.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>The hints were helpful for me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I understand why hints were delivered to me in general.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I understand why specific hints were delivered to me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I trust the system to deliver appropriate hints.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Given my behavior, I agree with the hints that were delivered to me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>The hints distracted me from my learning task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>The hints were confusing.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>The hints appeared at the right time.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

**ACSP Explanation:** Please rate the degree to which you agree or disagree with each of the following statements about the ACSP's explanation. 1= Strongly Disagree, 2= Disagree, 3=Agree nor Disagree, 4= Agree, 5= Strongly Agree.
<table>
<thead>
<tr>
<th>Statement</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would choose to have the hints again in the future.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am satisfied with the hints.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The hints were helpful for me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I understand why hints were delivered to me in general.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I understand why specific hints were delivered to me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I trust the system to deliver appropriate hints.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given my behavior, I agree with the hints that were delivered to me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The hints distracted me from my learning task.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The hints were confusing.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The hints appeared at the right time.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Open Ended Questions:

1. What did you like about the explanations?

2. What did you dislike about the explanations?

3. What can be improved about the explanations?

4. Please explain (to the best of your ability) of how you think the system provides hints?

5. Do you have any additional comments, questions, or concerns you would like to share?
ACSP Hints: Please rate the degree to which you agree or disagree with each of the following statements about the ACSP's hints, where: 1= Strongly Disagree, 2= Disagree, 3=Agree nor Disagree, 4= Agree, 5= Strongly Agree.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would choose to have the hints again in the future.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I am satisfied with the hints.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>The hints were helpful for me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I understand why hints were delivered to me in general.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I understand why specific hints were delivered to me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I trust the system to deliver appropriate hints.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Given my behavior, I agree with the hints that were delivered to me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>The hints distracted me from my learning task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>The hints were confusing.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>The hints appeared at the right time.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Open Ended Questions:

1. Please describe why you did not access the explanations, using the button "Why was I delivered this hint?".
2. Please explain (to the best of your ability) of how you think the system provides hints?

3. Do you have any additional comments, questions, or concerns you would like to share?