On the Use of Eye-tracking in the Assessment of Self-Explanation

by

Christina Merten

B.Sc, The University of Texas at Dallas, 2002

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Computer Science)

The University of British Columbia

December, 2005

© Christina Merten 2005
Abstract

The accuracy of a user model typically depends on the amount and quality of information available on the user's states of interest. An eye-tracker provides data indicating where a user is looking during interaction with the system. In this thesis we present research on the use of eye-tracking data for on-line assessment of user meta-cognitive behavior during the interaction with an open learning environment. We describe the design of a probabilistic model that processes this information and its formal evaluation. We demonstrate that adding eye-tracking information significantly improves the model accuracy on assessing user exploration and self-explanation behaviors.
Table of Contents

Abstract ................................................................. ii

Table of Contents .................................................. iii

List of Tables ......................................................... vi

List of Figures ......................................................... viii

Acknowledgements ................................................... ix

1 Introduction ......................................................... 1
  1.1 Open Learning Environments .................................. 1
  1.2 Self-Explanation ................................................. 2
  1.3 Thesis Approach: Using Eye-tracking as a Predictor of Self-
       Explanation .................................................. 3
  1.4 Thesis Goals and Contributions ............................... 3
  1.5 Outline ......................................................... 4

2 Related Work ...................................................... 5
  2.1 Eye-tracking Research .......................................... 5
    2.1.1 Retrospective Analysis of Eye Movements .................. 5
    2.1.2 The Online Use of Eye-tracking in Interface Operation .... 6
    2.1.3 Eye-tracking for Online Adaptation of System Interaction . 7
  2.2 Research on Supporting Self-Explanation .................... 8
  2.3 Open Learning Environments .................................. 9

3 ACE and the Addition of Eye-tracking to Assess Self-Explanation ......... 11
  3.1 The ACE Learning Environment ................................ 11
  3.2 Previous Versions of the ACE Student Model .................. 12
    3.2.1 The Original ACE Model .................................. 13
    3.2.2 Extending ACE to Monitor and Support Self-Explanation . 15
# Table of Contents

4 Preliminary User Study ....................................... 18  
4.1 Using Eye-tracking Data to Assess Implicit Self-Explanation ........................................... 18  
4.2 Study Goal .................................................. 19  
4.3 Participants ................................................... 19  
4.4 Experiment Design ........................................... 19  
4.5 Data Analysis ................................................ 21  
4.5.1 Setting the "Gold Standard" ......................... 21  
4.5.2 Time as a Predictor of Self-Explanation ........... 25  
4.5.3 Gaze Shifts as a Predictor of Self-Explanation .. 26  
4.6 Results ...................................................... 28  
4.7 Discussion ................................................... 31  

5 The New ACE Student Model .................................. 34  
5.1 Defining "SE-related Behavior" .......................... 34  
5.2 Adding "SE-related Behavior" to the model ............. 35  
5.2.1 The Naïve Bayesian Classifier ....................... 35  
5.2.2 Setting the Conditional Probabilities ............... 35  

6 Testing the Student Model .................................... 38  
6.1 Initial Tests of the Student Model ....................... 38  
6.2 Additional Training of the Student Model ............... 42  
6.3 New User Study ............................................. 44  
6.4 Testing the Student Model ................................ 45  
6.4.1 Testing the Accuracy of the New Model with Different Priors ........................................... 45  
6.5 Cross-validation Analysis .................................. 51  
6.5.1 Customizing Prior Probabilities ..................... 51  
6.5.2 Comparing the Different Models ..................... 53  
6.5.3 Calculating Optimal Thresholds ..................... 53  
6.6 Testing the Performance of the Models Using Different Evidence of Self-Explanation .......... 55  

7 Pupil Dilation as a Predictor of Self-explanation .......... 57  
7.1 Data Collection ............................................. 57  
7.2 Results and Discussion .................................... 59  

8 Conclusions and Future Work ................................ 65
Table of Contents

8.1 Satisfaction of Thesis Goals ........................................... 65
  8.1.1 User Study to Assess Time and Gaze Shifts as Predictors of Self-Explanation .................................................. 65
  8.1.2 Design of Revised Student Model .................................. 66
  8.1.3 Evaluation of Performance of New Student Model ............. 66
8.2 Summary of Conclusions .................................................. 67
8.3 Limitations ........................................................................ 68
8.4 Future Work ....................................................................... 68

Bibliography .......................................................................... 70

A Pretest .................................................................................. 73

B Posttest ............................................................................... 81
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Experiment schedule.</td>
</tr>
<tr>
<td>4.2</td>
<td>Classification scheme for coding episodes of positive and negative self-explanation.</td>
</tr>
<tr>
<td>4.3</td>
<td>Time elapsed after actions with and without self-explanation.</td>
</tr>
<tr>
<td>4.4</td>
<td>Raw experiment data.</td>
</tr>
<tr>
<td>4.5</td>
<td>Classification accuracy of different predictors.</td>
</tr>
<tr>
<td>6.1</td>
<td>Values of $implicitSE$ nodes corresponding to actions in study data.</td>
</tr>
<tr>
<td>6.2</td>
<td>$ImplicitSE$ accuracies with generic prior probabilities using training data.</td>
</tr>
<tr>
<td>6.3</td>
<td>Exploration node accuracies with generic prior probabilities using training data.</td>
</tr>
<tr>
<td>6.4</td>
<td>Accuracies of implicitSE nodes for different prior probabilities.</td>
</tr>
<tr>
<td>6.5</td>
<td>Areas of ROC curves for implicitSE nodes.</td>
</tr>
<tr>
<td>6.6</td>
<td>Accuracies of exploration nodes for different prior probabilities.</td>
</tr>
<tr>
<td>6.7</td>
<td>Areas of ROC curves for Exploration nodes.</td>
</tr>
<tr>
<td>6.8</td>
<td>Cross-validation results for $implicitSE$ nodes using various prior probabilities.</td>
</tr>
<tr>
<td>6.9</td>
<td>Cross-validation results for exploration nodes using various prior probabilities.</td>
</tr>
<tr>
<td>6.10</td>
<td>Cross-validation results for $implicitSE$ nodes using various student models.</td>
</tr>
<tr>
<td>6.11</td>
<td>Cross-validation results for exploration nodes using various student models.</td>
</tr>
<tr>
<td>6.12</td>
<td>Mean optimal thresholds and standard deviations for $implicitSE$ node assessments across folds of leave-one-out cross-validation.</td>
</tr>
<tr>
<td>6.13</td>
<td>Mean optimal thresholds and standard deviations for exploration node assessments across folds of leave-one-out cross-validation.</td>
</tr>
<tr>
<td>6.14</td>
<td>$ImplicitSE$ accuracies for the ACE model using different predictors as evidence of implicit self-explanation.</td>
</tr>
<tr>
<td>6.15</td>
<td>Exploration accuracies for the ACE model using different predictors as evidence of implicit self-explanation.</td>
</tr>
<tr>
<td>7.1</td>
<td>Sample fragment of pupil size data.</td>
</tr>
<tr>
<td>7.2</td>
<td>Mean normalized pupil sizes with and without self-explanation.</td>
</tr>
<tr>
<td>7.3</td>
<td>Sample fragment of pupil size data using Z-score normalization.</td>
</tr>
<tr>
<td>7.4</td>
<td>Mean Z-score normalized pupil sizes with and without self-</td>
</tr>
</tbody>
</table>
List of Tables

7.5 Mean Z-score normalized pupil sizes with and without self-explanation over only the middle 60% of each time interval. 63
List of Figures

3.1 The Machine Unit and Arrow Unit ......................... 11
3.2 The Plot Unit ........................................... 13
3.3 High level structure of ACE’s student model ............. 14
3.4 Original ACE model with self-explanation ............... 16
4.1 The Eyelink I eye-tracker ................................ 20
4.2 Coding of negative and positive self-explanation and utterance that has yet to be associated with an action ........ 24
4.3 ROC curve for time as filter for self-explanation ......... 26
4.4 Sample gaze shift ........................................ 27
4.5 Excerpt of ACE log file with added gaze shift tags ...... 28
4.6 Excerpt of user data file .................................. 28
4.7 Dual histogram of experiment data ......................... 29
4.8 ROC curve for time and gaze shifts as a combined predictor ............................................. 30
5.1 The ACE student model with a naïve Bayesian classifier component ........................................ 36
5.2 The ACE student model for two time slices ............... 37
6.1 ROC curves for models as predictors of implicit self-explanation ........................................ 40
6.2 ROC curves for models as predictors of sufficient exploration ......................................... 41
6.3 Two time slices of the adjusted ACE student model .... 43
6.4 ROC curves for models as predictors of implicit self-explanation over testing data with generic prior probabilities ........................................ 46
6.5 ROC curves for models as predictors of sufficient exploration over testing data with generic prior probabilities ........................................ 49
7.1 ROC curve for pupil size as a predictor of self-explanation ........................................ 61
7.2 ROC curve for pupil size, normalized with Z-scores, as a predictor of self-explanation ......................... 62
7.3 Pupil size plotted over the time interval when it is largest at the beginning (a) or at the beginning or the end (b). In (b), dotted lines indicate the middle 60% of the time interval ........................................ 63
7.4 ROC curve for pupil size as a predictor of self-explanation, using the middle 60% of the time interval after each action ........................................ 64
Acknowledgements

First, I would like to thank my supervisor, Cristina Conati, for her helpful guidance over the last few years. Special thanks to Kasia Muldner and Andrea Bunt, both for their help with ACE and for their friendship. Giuseppe Carenini graciously agreed to be my second reader. Finally I would like to thank my mother and stepfather, Susan and Don Cundieff, and my father, Jay Merten, for their support and encouragement over the years.
Chapter 1

Introduction

Eye movements have long been used as a source of evidence revealing cognitive state. The active theory of visual form perception relates fixation duration and location to internal events within the mind [21]. Studies in this field have shown that gaze data can be used to infer such higher mental processes as attention, understanding and memory [21].

Metacognitive skills are high-level, domain independent skills which can greatly influence learning [6]. These include self-explanation, self-explanation, analogical reasoning, and the ability to learn through exploration. The metacognitive skill of self-explanation is particularly important to the use of open learning environments which rely on unconstrained exploration of the interface to teach domain material.

The research described in this thesis was carried out to explore the use of eye movements as a predictor of self-explanation in an open learning environment.

1.1 Open Learning Environments

An Intelligent Tutoring System (ITS) is a computer-based tutor which includes knowledge of the domain material being taught (an expert model) and knowledge of the learner (a student model) [27]. With a traditional ITS, students solve problems chosen to serve as effective learning experiences. Throughout this interaction, the student model maintains an assessment of the learner's knowledge of the domain material which the system may use to provide tailored instruction.

Open learning environments are intelligent tutoring systems which depend on free exploration of the interface by the user rather than structured explicit instruction. By exploring the system interface, students can acquire a deeper, more structured understanding of the domain material [27].

However, not all students are able to explore effectively in these environments. In many cases it is because they lack the prerequisite knowledge necessary to understand the domain material. Other students lack the meta-cognitive skills needed to initiate exploratory actions or choose actions which cover all concepts in the target domain. Furthermore, even those who can explore appropriately are sometimes unable to interpret and generalize the results of their actions. In particular, many students are unable
Chapter 1. Introduction

or unwilling to self-explain, i.e. to clarify and elaborate the information presented with respect to underlying domain theory [7, 10].

For these students it is necessary for the system to provide on-line support to encourage effective exploration. While such support can be useful when needed, it can also undermine the unrestrictedness of the open learning environment. Thus it is necessary to model student exploratory tendencies so that intervention can only occur when needed.

1.2 Self-Explanation

Self-explanation is the metacognitive skill of spontaneously explaining to oneself instructional material in terms of the underlying domain theory [7]. Studies have shown that self-explanation can greatly influence learning [6, 23]. For example, students who carefully consider the effects of their actions in terms of the domain material are more likely to learn effectively than those who merely play with the interface in an intelligent tutoring system. It has been shown however, that many students fail to spontaneously self-explain when interacting with an open learning environment. Thus computer-based tools which support self-explanation may be used to enhance exploration in an open learning environment.

Deciding when to hint for self-explanation is a challenging issue in an open learning environment. The hints should interfere as little as possible with the exploratory nature of the interaction, but should also be timely so that even the more reluctant self-explainers can appreciate their relevance. Thus, for an interface to effectively support and encourage self-explanation, it is crucial that the student model be as accurate as possible in its assessment of whether a student is actually engaging in this process. There may be two types of self-explanations which need to be detected: (i) explicit self-explanations, i.e., self-explanation that the student generates using menu-based tools possibly available in the interface; (ii) implicit self-explanations, that students generate in their head. These are the most difficult to detect due to the lack of hard evidence of their occurrence.

Model bandwidth refers to the amount and quality of information available to a user model to assess the user's states of interest [32]. The greater the bandwidth, the less ambiguity exists in mapping observable behaviors onto user internal states, and the more accurate the model's assessment can be.

Time may be used as evidence of self-explanation behavior, under the assumption that students who self-explain take longer between exploratory actions. However, time as a predictor can overestimate self-explanation behavior as user may get distracted by outside factors such as other students.
The research presented in this thesis focuses on the use of eye-tracking to increase the bandwidth in a user model which assesses self-explanation in an open learning environment.

1.3 Thesis Approach: Using Eye-tracking as a Predictor of Self-Explanation

The research presented in this thesis is aimed at exploring if and how eye-tracking data can be used to detect self-explanation in an open learning environment.

In an open learning environment, an essential part of self-explanation involves observing and considering the outcomes of one's actions. Previous studies [21] have shown that a user's gaze may be used as an indication of which parts of an interface hold the user's attention. Thus it is reasonable to assume that the presence or absence of certain gaze patterns may be used to assess self-explanation behavior.

In this thesis, gaze patterns are first determined which indicate that the user is observing the outcomes of her actions. These patterns are then tested as evidence of self-explanation in a student model.

1.4 Thesis Goals and Contributions

The main objective of this thesis is to explore if and how eye-tracking may be used as a detector of implicit self-explanation. To meet this objective, this thesis has the following goals:

1. To create a student model which uses certain gaze patterns as evidence of implicit self-explanation. This involves:
   (a) Performing a user study to test assumptions concerning which gaze patterns indicate implicit self-explanation and collect quantitative data regarding the reliability of these patterns as predictors
   (b) Designing and building a new student model based on the results of this study.

2. To evaluate the performance of the new model as a predictor of implicit self-explanation and sufficient exploration, particularly in comparison with a previously devised student model which does not use eye-tracking data.

The student model discussed in this thesis uses on-line processing of eye movement data in its assessment of whether or not implicit self-explanation
has occurred. Most eye-tracking research to date has involved either the offline processing of gaze data for a greater understanding of the link between eye movement and user cognition or the online use of eye-tracking data for interface evaluation or manipulation by the user. The research described in this thesis involves the design of a student model which uses online processing of eye gaze data to assess self-explanation behaviour and an evaluation of the performance of the new model.

1.5 Outline

Chapter 3 of this thesis presents a review of the design and development of the ACE system and its student model for assessing sufficient exploration. The previous addition of self-explanation is also discussed as well as the potential for eye-tracking data as a predictor of self-explanation. Chapter 4 describes a preliminary user study carried out to determine the correlation between certain gaze patterns and implicit self-explanation and thus establish the value of these gaze patterns as predictors of self-explanation. The results of this study were applied to the design of a new student model which includes eye-tracking data, as explained in Chapter 5. Chapter 6 describes an evaluation of the predictive performance of the new model, as compared to that of the previous models reviewed in Chapter 5. Finally, Chapter 6 contains a summary and conclusions.
Chapter 2

Related Work

This chapter presents a review of previous work related to the research described in this thesis. In section 2.1, earlier eye-tracking research is discussed. Section 2.2 provides a review of previous work on self-explanation. Finally, section 2.3 describes past research involving open learning environments and briefly introduces the ACE system.

2.1 Eye-tracking Research

2.1.1 Retrospective analysis of eye movements

Many areas of science have long utilized the recording and subsequent analysis of eye movement data. For example, eye-tracking has been used in cognitive psychology as a tool to help understand both motor and cognitive processes [18, 21]. In [18], Just and Carpenter analyze the fixations of readers of various fragments of text. They show that better readers directly fixate on fewer words and spend less time on each. Comprehension was also improved for readers who fixated more on content words (e.g., nouns and verbs) than function words (e.g., conjunctions and prepositions). Further, skilled readers were found to fixate on words in the order in which they appear in the text, rather than regress to earlier words.

Williams [33] applied retrospective analysis of eye-tracking data to his study of how people look at pictures and why they choose to fixate on certain areas. He found that color greatly influenced subjects' gaze patterns, while size and shape did not. In addition to fixating on some colors more than others, subjects' gazes were more likely to follow patterns defined by color than by size or shape. Thus color-coding is an effective tool for users to locate items efficiently.

In [29], Sodhi et al. used an eye-tracker to record the gaze patterns of drivers to assess the distraction caused by in-vehicle devices such as cell phones and compact disc players. They performed an experiment in which they had subjects drive for twenty-two miles along a predetermined road while listening to an audio recording of instructions for which devices to use. After collecting the eye movement data, they were able to classify the use of each device as a single-glance or multiple-glance task.
In the field of HCI, eye movement data has been studied post hoc to evaluate usability issues and understand human performance. For instance, Schiessl et al [25] used an eye-tracker to investigate differences in attentional behavior between genders for textual information and pictoral stimuli on websites. They found that women tend to focus more carefully on textual information while men pay more attention to pictures on websites. It should also be mentioned that when the participants of these studies were asked where in the interface they thought they looked, their perceptions often differed from reality. Here accurate eye movement data could only be found with an eye-tracker.

In [11], offline processing of eye-tracking data was used in a study designed to improve the efficient generation of non-photorealistic images in a graphical interface. Here participants were initially asked to stare at different images for five seconds each. Eye fixation data was analyzed to determine which parts of the picture the study subjects found to be most meaningful. Complex algorithms were then designed which were used to draw the most “important” parts of the picture first.

While the research described in this thesis includes the use of retrospective analysis of eye movements to design and test a student model, it differs from these efforts because this new model uses the online processing of eye-tracking data to assess and support student learning.

2.1.2 The online use of eye-tracking in interface operation

There has also been fairly extensive research in using eye gaze as an alternative form of input to allow a user to explicitly operate an interface. In [17], Jakob explores issues surrounding the real-time processing of eye data such as efficient noise reduction and the organization of gaze information into tokens from which relevant data may be extracted. He then discusses the potential of eye-tracking as a tool in several forms of interface manipulation. These include selecting and moving an object, scrolling text, and navigating menus.

Salvucci and Anderson [24] applied these ideas to the design and evaluation of a sample operating-system interface which allows users to execute commands with their eyes. They developed IGO (Intelligent Gaze-added Operating-system), with which users could perform such operations as opening, closing, and dragging windows using their eyes. Experiments showed that after little practice, users were able to perform tasks with their gaze almost as correctly and efficiently as with a mouse. Participants in the study also showed a preference for the gaze modality over the mouse.

Majaranta et al [19] devised a system which allowed users to type with their eyes via an eye-tracker. A picture of a keyboard appeared on the keyboard for users to look at and small fixations were interpreted as key
presses. When these occurred, the system played an audio recording of the name of the character typed. A user study revealed that requiring shorter fixations for key presses resulted in slower typing and more double-letter errors.

In [14], Hornof et al. describe the EyeDraw system, which enables children with severe motor impairments to draw pictures by just moving their eyes. In EyeDraw, longer fixations which exceed a dwell threshold are necessary to "lower" and "lift" a drawing pen which moves across the screen. Three out of four of the children who participated in a small user study were able to quickly and easily draw the requested shapes and recognizable scenes including houses, stick figures, a car, and a butterfly.

Unlike IGO and EyeDraw, the system discussed in this thesis uses real-time processing of a user's gaze to interpret user non-explicit metacognitive behaviours, enabling online adaptation of the interaction.

### 2.1.3 Eye-tracking for online adaptation of system interaction

There is a much smaller body of work on on-line processing of eye movement data to interpret a user's behavior beyond interface operation to allow real-time adaptation of the interaction. Some of this work uses gaze tracking to assess user mental states, i.e. whether the student having difficulty in a reading task [32]. For example, in [28], Sibert et al describe the development and testing of the Reading Assistant, a system for automated reading remediation. Users of the Reading Assistant read text from the screen while an eye-tracker monitors their gaze. Visual and auditory cues are given depending on where the student is looking or whether changes in gaze patterns indicate that the student is having difficulty reading a word. For example, if a student stares at a word for a certain amount of time, an audio recording of the word being spoken will play.

Another example of the use of an eye-tracker in the assessment of final states appears in [15]. Here Iqbal and Bailey perform a study which shows how eye-tracking can be used to determine which type of task the user is performing (e.g., reading email vs. reading a web page). This information can be applied to the design of an attention manager which balances the user's need for minimal disruption with an application's need to deliver necessary information. It also has important implications for interface design in general.

Research has also been carried out in the use of gaze information to assess user mental states such as interest or problem-solving strategies. For example, in [30], Starker and Bolt describe a system which uses eye-tracking to determine the part of a graphical interface at which the user is interested.
More information is then given about this area via visual zooming or synthesized speech.

In [22], Qu and Johnson use eye-tracking in an Attention Tracking Model used to infer the learner's focus of attention while using the Virtual Factory Teaching System (VFTS), an online factory system for teaching engineering concepts and skills. The VFTS system includes a student model which assesses the motivational states of confidence, confusion and effort. This motivational model uses the time that the user spends after reading something but before taking action to assign a level of confusion (high, medium or low) to the user. In addition, the time the user spends reading, making a decision, and then taking an action is used to calculate an effort value relating to each task. A study carried out in [22] showed that this model was able to make accurate assessments about the learner's motivational states. These assessments could then be used by the system to provide tailored, proactive help to the user to motivate learning.

Gluck and Anderson [13] studied the use of eye-tracking in the assessment of problem-solving strategies in the PAT Algebra I tutor. They used gaze patterns to detect cognitive behaviors such as shifts of attention, disambiguation of the problem statements and disambiguation of an error. Other fixation patterns were used to identify when the user failed to read or process error messages or other information critical to problem solving. This information about the user's cognitive behavior could then be used by the student model to adapt to the user's needs and thus create a better interaction.

The work described in this thesis extends this body of research by exploring if and how eye-tracking can help assess mental states related to the meta-cognitive, domain-independent skill of self-explanation.

2.2 Research on Supporting Self-explanation

Several computer-based tutors have been created to provide explicit support for self-explanation. However, each of these systems aims to support and encourage students' self-explanations during highly structured activities which target problem-solving skills. For example, the Geometry Explanation Tutor [1] supports self-explanation as students solve problems in geometry theorem proving. This system prompts students to solve problems and requires that they enter an explanation for each of their solution steps. A user study [1] was also performed which compared this system to a similar version with the self-explanation prompts removed. Students using the tutor which required self-explanation were found to score considerably higher on post-tests than students who used the version which only included problem solving.
Normit-SE [20] was devised as a problem solving environment for learning data normalization. Students using this system are asked to self-explain while they solve problems. However, in contrast to the Geometry Explanation Tutor, Normit-SE only requires self-explanation when an action is performed for the first time or is performed incorrectly. A user study revealed that when students used this system, both their problem solving performance and ability to explain their actions improved.

Neither the Geometry Explanation Tutor nor Normit-SE considers the student who spontaneously self-explains without the need for prompts or other tools. These students are needlessly interrupted, making the interface restrictive and cumbersome. The SE-Coach [8, 9] addresses this problem by modeling the user's self-explanation behavior. This model uses information on both knowledge and reading patterns to assess self-explanation during example studying in the domain of Newtonian physics. Reading patterns are tracked via a poor-man-eye-tracker which requires that students move the mouse to explicitly uncover various parts of the studied example. Thus the SE-Coach only prompts for self-explanation when its student model assesses that the student actually requires it.

In open learning environments, minimizing unnecessary prompts to self-explain is of particular importance. At the same time, the unrestricted nature of the interaction in such an environment makes it difficult to model the user's cognitive state. The research discussed in this thesis involves a system which, like the SE-Coach, includes an assessment of self-explanation. However, while the SE-Coach mainly relies on time spent on interface actions to detect spontaneous self-explanation, this system also uses a cumulative assessment of the student's tendency to self-explain. Thus, it possesses a student model which has a broader set of information to use in determining when a user needs help with self-explanation.

### 2.3 Open Learning Environments

Open learning environments are designed to promote free exploration of the instructional material rather than structured, explicit instruction for student learning. By promoting unconstrained exploration of the interface, they force learners to take a more active role in their instruction. In theory, this type of learning should provide students with a deeper and more structured understanding of the domain material [27, 31].

However, some studies have shown that not all students possess the metacognitive skills necessary for effective exploration of open learning environments [6, 23]. These students require guidance by the system in order to initiate and perform useful experiments and correctly interpret and
Chapter 2. Related Work

generalize the results. At the same time unnecessary interruption of the student must be minimized to preserve the unrestricted nature of open learning environments. Thus the user's exploratory skills should be modeled by the system so that tailored support can be offered.

Smithtown [27] is an open learning environment designed to both improve a student's scientific inquiry skills and provide an environment in which users may discover concepts of macroeconomics. In Smithtown, a distinction is made between the "exploration phase" and the "experiment phase." In the exploration phase, students gather information and make observations about variables in an example economy. In the experiment phase, students perform experiments by changing variables concerning supply and demand and draw conclusions from the outcomes of their actions. While they use the system, their actions are constantly monitored for signs of good or bad behavior and students who perform poorly are coached to become more effective problem solvers. Similarly, users who show evidence of good behavior are congratulated by the coach. However, Smithtown does not assist users in the exploration phase or address the needs of students unable to perform experiments in the first place.

ACE [4] is an open learning environment for the domain of mathematical functions. It uses a student model to assess the student's ability to explore effectively, offering tailored hints when needed. Unlike Smithtown, ACE assists the learner throughout the exploration and provides support to students who appear unable to choose an appropriate action. In addition, ACE monitors the user's tendency to self-explain and supports this self-explanation through the use of customized prompts and other interface tools. This system will be described in detail in the next chapter.
Chapter 3

ACE and the Addition of Eye-tracking to Assess Self-Explanation

This chapter describes the ACE system and the benefits of adding eye-tracking data to the ACE student model. Sections 3.1 and 3.2 provide a review of the ACE interface and previous versions of the student model, respectively. In section 3.3, the need to add gaze data to the assessment of implicit self-explanation is discussed and explained.

3.1 The ACE Learning Environment

ACE [4] is an adaptive learning environment for the domain of mathematical functions designed to support learning through exploration. ACE's activities are divided into units and exercises. Units are collections of exercises whose material is presented with a common theme and mode of interaction. Exercises within units differ in function type and interaction. For completeness, all three units of ACE will be briefly described in this section. However, the third unit, the Plot Unit, will be the focus of all research described in this thesis.

The first unit, the Machine Unit, allows the user to input numbers into a function (represented in the interface by an arrow) by dragging them behind the tail of the arrow. The Machine Unit interface appears in the left hand side.

![Figure 3.1 The Machine Unit (left) and Arrow Unit (right)](image)

The first unit, the *Machine Unit*, allows the user to input numbers into a function (represented in the interface by an arrow) by dragging them behind the tail of the arrow. The Machine Unit interface appears in the left hand side.
This type of exploration gives the user a sense of how functions transform an input value into an output value. This unit features four exercises for different types of functions: a constant function exercise, a linear function exercise, a quadratic monomial exercise and a quadratic polynomial exercise. For functions whose solution involves more than one calculation step (all but the constant function), users must press the step button to cause each operation to occur. By watching the calculations occur one step at a time, the user may acquire a greater understanding of the low-level process.

In the second unit, the Arrow Unit, users may experiment by drawing lines from input values to output values, as shown in the right hand side of figure 3.1. If the correct output is chosen for the input and the function, the line turns green; otherwise the line is red. Like the Machine Unit, the Arrow Unit consists of a constant function exercise, a linear function exercise, a quadratic monomial exercise and a quadratic polynomial exercise. The Arrow Unit is the only unit in ACE for which an action may be classified as “correct” or “incorrect”; thus it is also the only unit that allows students to explicitly demonstrate mastery of a skill or concept. Here students have the opportunity to test their understanding of the material presented in the Machine Unit.

The third and final unit of ACE is the Plot Unit. It will be the focus of all research described in this thesis. In the plot unit, students are presented with a function and its curve plotted in the Cartesian plane as shown in figure 3.2. Here the student may drag the curve around the plane, causing the function below to change to match the curve in its new location. Alternatively, the user may change values in the function equation, causing the curve to change in the plane. This type of exploration is intended to give the users a sense of what plotted curves look like for different functions and what effect each parameter in the function has on the graph. The plot unit includes a constant function exercise, a linear function exercise, and monomial function exercise.

To support the exploration process, ACE also includes a coaching component that provides tailored hints when ACE’s student model predicts that students have difficulties exploring effectively [3]. In the next section we describe previous versions the ACE student model.

### 3.2 Previous Versions of the ACE Student Model

In this section previously developed versions of the ACE student model are reviewed. The original ACE model was based entirely on student actions [4]. It used exploratory behavior alone to assess sufficient exploration. This model was later extended to predict whether or not each of these exploratory actions was self-explained by the student, allowing both for support for self-
ACE and the Addition of Eye-tracking to Assess Self-Explanation

explanation via prompts and other tools and a more informed assessment of sufficient exploration [5].

3.2.1 The Original ACE Model

ACE’s student model uses a Dynamic Bayesian network to manage the uncertainty in assessing students’ exploratory behavior [5]. The main cause of this uncertainty is that neither exploratory behavior nor the related metacognitive skills are easily observable unless students are required to make them explicit. However, forcing students to articulate their exploration steps would clash with the unconstrained nature of open learning environments.

The general structure of ACE’s student model was derived from an iterative design process [4] which yielded a better understanding of what defines effective exploration. Figure 3.3 shows a high-level description of
this structure, which includes several types of nodes used to assess exploratory behaviour at different levels of granularity:

- **Relevant Exploration Cases**: the exploration of individual exploration cases in an exercise (e.g., dragging the -7 square to the left of the arrow in the machine unit or changing the slope of a line to 3 in the Plot Unit)
- **Exploration of Exercises**: the exploration of individual exercises
- **Exploration of Units**: the exploration of individual units
- **Exploration of Categories**: the exploration of groups of relevant exploration cases that appear across multiple exercises (e.g., all the exploration cases involving a positive slope in the Plot Unit)

![Figure 3.3 High Level Structure of ACE'S Student Model](image)

The links among the different types of exploration nodes represent how they interact to define effective exploration [4]. Exploration nodes have binary values which represent the probability that the learner has sufficiently explored the items associated with the node.

ACE's student model also includes binary nodes representing the probability that the learner understands the relevant pieces of knowledge (summarized by the node knowledge in figure 3.3). The links between knowledge and exploration nodes represent the fact that the degree of exploration necessary to understand a concept depends on how much knowledge of that concept a learner already has. Knowledge nodes are updated only through actions for which there is a clear definition of correctness (i.e., actions in the arrow unit). Prior probabilities for these nodes can be estimated from existing information on a student's prior knowledge of the related topic (see Chapter 6 for more details of this process). These nodes are never updated within the Plot Unit since it only supports purely exploratory activities.

Early studies on ACE yielded encouraging evidence that the system based on the model in figure 3.3 could help students learn better from exploration [4]. However, these studies also showed that sometimes the ACE student model overestimated student's exploratory behavior because it regarded interface actions as sufficient evidence of good exploration, without
considering whether or not a student was *self-explaining* the outcomes of these actions. For instance, a student who quickly drags the curve around the screen in the Plot Unit – but never reflects on how these movements change the function equation – performs many exploratory actions but can hardly learn from them because she is not reflecting on (self-explaining) their outcomes. This exact behaviour was apparent in a number of subjects who tended to spend little time on their exploratory actions, and who did not learn the associated concepts.

### 3.2.2 Extending ACE to Monitor and Support Self-explanation

To address the model limitation described above, ACE's interface and student model were extended to track and support self-explanation [5]. The original version of ACE only generated hints indicating that a student should further explore some elements of a given exercise. Augmenting ACE with the means to track self-explanation allows ACE not only to detect when a student's exploration is sub-optimal, but also to recognize if the cause is a lack of self-explanation and provide tailored hints to correct this behavior.

The ACE student model was augmented to assess implicit self-explanation using only time spent on each exploratory action as evidence of implicit self-explanation [5]. The relation between time and self-explanation in the model was based upon the assumptions that (1) no self-explanation can happen if a student switches too rapidly from one exploration case to the next; (2) the longer a student dwells on a case the more likely it is that she is trying to self-explain it.

Figure 3.4 shows two time slices corresponding to implicit self-explanation slices (similar slices capture the occurrence of explicit self-explanation). Nodes representing the assessment of self-explanation are shaded grey. All nodes in the model are binary, except for *time*, which has values *Low/Med/High*. In this figure, the learner is currently exploring exercise 0 (node eo) in the Plot Unit, which has two relevant exploration cases (eoCase0 and eoCase1 in fig 3.4). Each exploration case influences one or more exploration categories (positive and negative intercepts in the figure). In the first slice, the learner performed an action that corresponds to the exploration of eoCase0. In the second slice, the learner performed an action corresponding to eoCase1. In this version of the model, the probability that a learner's action implies effective exploration of a given case depends on both the probability that the student self-explained the action and the probability that she knows the corresponding concept, as assessed by the set of knowledge nodes in the model (summarized in figure 3.3 by the node *Knowledge*). Factors influencing the probability that implicit self-explanation occurs include the *time* spent exploring the case and the *stimuli* that the learner has to self-explain. Low time is always taken as negative evidence for
implicit explanation. The probability that self-explanation happened with longer time on action depends on whether there is a stimulus to self-explain. The probability that a stimulus exists depends on the learner's general tendency to self-explain and on whether the system generated an explicit hint to self-explain.

![Diagram of the Original ACE model with self-explanation]

Figure 3.4. Original ACE model with self-explanation

Time, however, can be an ambiguous predictor for self-explanation. First, it is hard to define for different learners what is insufficient time for self-explanation. Furthermore, a student may be completely distracted during a
long interval between exploration cases. Thus, we chose to explore an additional source of evidence of self-explanation behavior, i.e., the student's attention patterns during the exploration of a given case, as described in the next chapters.
Chapter 4

Preliminary User Study

This chapter describes the design and results of a user study carried out with ACE during the summer and autumn of 2004. The study demonstrates the potential for detection of implicit self-explanation through observation of user attention patterns. In addition, the study serves to compare the performance of latency and gaze patterns as predictors of implicit self-explanation.

This experiment focuses entirely on the plot unit of ACE and while data was collected for all three units, only the students’ explorations within the plot unit are analyzed for evidence of implicit self-explanation.

4.1 Using Eye-tracking Data to Assess Implicit Self-explanation

Self-explanation is more likely to occur if the student actually attends to the parts of the interface showing the effects of a specific exploratory action. Studies have shown that users tend to look at the region of the interface to which they are paying attention [21]. This naturally leads to the assumption that in an interface like ACE’s, eye-tracking data on student attention patterns may be considered as a source of evidence of implicit self-explanation.

As mentioned earlier, the student exploring the Plot Unit may either (i) drag the curve across the graph region, causing the function equation to change or (ii) type new parameters into the function line, leading to a change in the curve’s position or orientation. Thus, it seems reasonable to expect that in order for a student to implicitly self-explain an action, she must make a change in one region (the plane or function line) and then look at the other to observe the outcome.

A user study was then carried out to test these assumptions and collect quantitative data on the relationship between eye movements and implicit self-explanation. This chapter presents analysis intended to compare the performance of the following predictors of self-explanation: (i) time, (ii) gaze shifts and (iii) a combination of the two.
4.2 Study Goal

The purpose of this study was to compare the performance of various predictors of implicit self-explanation. The study was carried out to test assumptions concerning which gaze patterns may be used as evidence of implicit self-explanation. It was further intended to yield data on the correlation between the presence of self-explanation and time taken between exploratory actions.

4.3 Participants

Since participants in the study were going to be using ACE and the interaction was going to be scrutinized for signs of self-explanation, it was essential that the domain knowledge be as new to them as possible. The ideal participants were those with a minimal knowledge of mathematical functions and their graphs but with some initial understanding of the topic. Since ACE does not offer any background instruction, total novices wouldn’t know where to start.

The 19 participants were students at the University of British Columbia who had not yet taken any college level math courses. They varied in the amount and level of mathematics that they recalled from high school. While it was never explicitly stated that users had to be non-science students, advertisements were placed on bulletin boards meant exclusively for psychology students. As a result, participants were all arts and humanities students. ACE participants were paid $10 per hour for their time.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Time taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>15 minutes</td>
</tr>
<tr>
<td>Calibration</td>
<td>10 minutes</td>
</tr>
<tr>
<td>Interaction with ACE</td>
<td>40 minutes</td>
</tr>
<tr>
<td>Post-test</td>
<td>15 minutes</td>
</tr>
<tr>
<td>Total</td>
<td>80 minutes</td>
</tr>
</tbody>
</table>

4.4 Experiment Design

The study consisted of individual sessions lasting approximately 80 minutes each. Table 1 summarizes the structure of each session. At the beginning of each session, the participant would complete a pretest intended to determine
her initial knowledge of mathematical functions and their graphs (see appendix A). This usually took about 15 minutes.

During interaction with ACE, a student’s eye movements were recorded by an Eyelink I eye-tracker, developed by SR Research Ltd., Canada. This is a head-mounted eye-tracker which weighs 600 grams. It is fairly uncomfortable and intrusive to use, due to both its weight and to the presence of a little camera which hangs in the user’s peripheral vision. This is shown in figure 4.1. This particular eye-tracker was used because it was readily available through the psychology department at the University of British Columbia. However, the same data could be obtained through a completely non-intrusive remote eye-tracker, consisting of a small camera which sits on top of the monitor or on some other flat surface (e.g. IView X Red from SensoMotoric Instruments, USA).

![The Eyelink I eye-tracker](image)

Figure 4.1. The Eyelink I eye-tracker

After the eye-tracker was placed on the head of the participant, it had to be calibrated to correctly track her eye movements. For this to occur, the student had to watch a white dot appear at different locations across a blank screen, following it as closely as possible. The process often had to be repeated more than once to achieve acceptable calibration. This phase of the experiment usually lasted less than 10 minutes.

Once calibration was achieved, ACE was started and the participant went through each of the units at her own pace. Each student was told to “think
aloud" during the interaction. In fact the following script was read to each student before opening ACE:

"As you use this program, we would like you to speak aloud and explain what each action means to you. That is, why did you do it and what new information, if any, about functions, does it provide? Tell us whatever is going through your mind – even if it seems unimportant."

This part of the experiment usually lasted about 40 minutes.

After exiting ACE, each participant completed a post-test which was very similar to the pretest and took about 15 minutes (see Appendix B). The post-test only differed from the pre-test in the constants used in the functions and the ordering of different questions.

Due to the use of the eye-tracker, only one subject could participate in the experiment at a time.

In addition to the paper pre-test and post-test, each session yielded a log file from the ACE system. This contained data which included each exploratory action as well as the time at which the action was taken. The eye-tracker also generated a log file containing the coordinates and duration of each fixation. In addition, video and audio recordings of the interaction were created showing the ACE screen and allowing for later analysis of the user’s speech.

4.5 Data Analysis

In order to assess the performance of latency and gaze patterns as predictors of implicit self-explanation, it was necessary to first analyze the audio and video data for each session in order to classify students' self-explanation behavior. In particular, we needed to isolate exploratory actions accompanied by implicit self-explanation and those that were not. Then these classifications could be tested for correlations with those predicted by time and gaze shift data.

4.5.1 Setting the “gold standard”

To understand how attention patterns and time per exploration case relate to implicit self-explanation, it was necessary to obtain from the study data points on actual positive and negative self-explanation episodes. Here, "negative self-explanation" indicates situations in which the students did not self-explain, not situations in which students self-explained incorrectly, consistent with the original definition of self-explanation [7].
Chapter 4. Preliminary User Study

The audio recordings of each interaction were found to be the most useful for detecting the presence or absence of implicit self-explanation. It should be noted that ACE data collected for the purpose of classifying self-explanation behaviors (both positive and negative) consisted of explicit verbalization by the subjects as to their thoughts. Similar to other studies on self-explanation [e.g., 6, 18], using subjects’ verbalizations is acceptable since the participants were instructed to share all of their thoughts and were not told anything about the data analysis process or the actual purpose of the study. Thus the episodes we related to presence or absence of self-explanation in the data can accurately be described as “internal” since they reflect the subjects’ thoughts which are unknown to the ACE system. Further, with existing technology, this is as close as we could come to reading the students’ thoughts in our search for evidence of implicit self-explanation or lack thereof.

To maximize the objectivity in the analysis of the audio data, two observers (the author and another graduate research assistant) independently analyzed the audio data and then created links between the verbal episodes and the corresponding exploration cases in the log files. This turned out to be a much more laborious process than expected, due to two factors.

Table 4.2 Classification scheme for coding episodes of positive and negative self-explanation.

<table>
<thead>
<tr>
<th>Evidence of positive self-explanation</th>
<th>Evidence of negative self-explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Verbalized conclusions about domain-specific principles related to the exploration process (regardless of correctness)</td>
<td>• Simple narration of the interaction</td>
</tr>
<tr>
<td>• Prediction of an action just before it occurred</td>
<td>• Isolated statements of confusion</td>
</tr>
<tr>
<td></td>
<td>• Expressions of inattentiveness</td>
</tr>
</tbody>
</table>

First, we needed to devise a detailed coding scheme in order to objectively convert fragments of audio data into isolated episodes of positive or negative self-explanation. While coding schemes exists for self-explanation study during problem solving, this is the first attempt to evaluate self-explanation during independent exploration. This problem was addressed by having the two observers independently label a subset of the audio data, then compare their classifications, possibly reconcile them and devise the coding scheme based on this discussion.

In the coding scheme, student utterances were classified as self-explanation if they expressed a conclusion about a domain-specific principle
related to the exploration process (e.g., "when I increase the coefficient here, the line gets steeper") regardless of correctness, or if they predicted the result of an action just before it occurred (e.g., "putting a negative sign here will turn the curve upside-down"). Simply narrating the outcome of each action once it happened (e.g., "this number just changed to a 3"), or isolated statements of confusion (e.g., "I don't understand what's happening") were not considered self-explanation. However, tentative explanations followed by expressions of confusion were coded as self-explanation. Obvious statements of inattentiveness (e.g., "I'm just playing") were also coded as evidence of lack of self-explanation. This classification scheme is summarized in table 4.2. It should also be noted that whenever an exploratory action was followed by evidence of both positive and negative self-explanation, the action was considered self-explained.

The coded data for three episodes appears in figure 4.2. The following types of tags were used to describe each episode:

- **ACE_TIME** tags give the system time when the action occurs
- **VID_TIME** is an optional tag giving the time as kept by the video recorder
- **ACTION** tags describe the exploratory action that occurred (e.g. "Moved linear function" or "typed new exponent")
- **SE_TYPE** tags tell whether or not implicit self-explanation occurred as determined by the observers Here a Y (yes), N (no) or ? (inconclusive) always appears
- **SE_DESCRIPTION** tags give the student's statement which led the observers to conclude that self-explanation did or did not occur. If the observers were unable to decide or agree upon whether or not it had occurred, this tag was left blank
- **LOG_PTR** tags give the action's line number in the ACE log file for quick reference

In some cases, it was very difficult to tell from the video recording which action was occurring when certain utterances were made by the student. This was a particular problem when the student was talking and moving the curve around very quickly. The VID_TIME tags were created to solve this problem as objectively as possible. When an utterance was made and it was not possible to tell which action was occurring at the time, this utterance was recorded in an SE_DESCRIPTION tag as in figure 4.2c. The VID_TIME tag gave the time as kept by the video recorder when the speech occurred and the SE_TYPE tag gave the observer's determination of the presence or absence of implicit self-explanation. The ACE_TIME, ACTION, and LOG_PTR tags were left blank as the information they were intended to contain was not yet known.
Next, using the time synchronization line appearing at the start of each coded file (see the top of Figure 4.2) which gave the time at which ACE was started in both forms, a post-processing program filled in the ACE_TIME tags for each of these utterances. This program then used the ACE_TIME information and the ACE log file to determine which action was occurring at the time and fill out the ACTION and LOG_PTR tags accordingly.

(a) <ACE_TIME: 17:07:08> <VID_TIME: 02:38:34> <ACTION: Moved constant function> <SE_TYPE:N> <SE_DESCRIPTION: "I'm not sure what's going on"> <LOG_PTR: 1057>

(b) <ACE_TIME: 17:07:27> <VID_TIME: > <ACTION: Moved linear function> <SE_TYPE:Y> <SE_DESCRIPTION: "moving the line changes the y intercept in the equation"> <LOG_PTR: 1127>

(c) <ACE_TIME: > <VID_TIME: 02:43:65> <ACTION: > <SE_TYPE:Y> <SE_DESCRIPTION: "the exponent doesn't change when the curve is moved"> <LOG_PTR: >

Figure 4.2 Coding of (a) negative and (b) positive self-explanation and (c) utterance that has yet to be associated with an action

Each of the observers individually applied this scheme to code the audio data and then their results were compared. The intercoder reliability was 93% in this phase, which suggests a high level of objectivity in the classification scheme. Only episodes on which the coders fully agreed were used in the rest of the analysis.

A second factor which increased the complexity of data analysis was difficulty in determining which action each coded utterance corresponded to. The observers at first assumed that subjects' utterances always pertained to whatever exploratory action they had just taken. Thus, associating each utterance with an action was as simple as matching up the times at which each occurred. However, subsequent analysis of the video data showed that this was not always the case, particularly for users who showed great reluctance to think aloud. These learners had to be repeatedly prompted by the observers to speak, so some of the conclusions they shared weren't reached when the spoke, but related to self-explanation that occurred a few minutes earlier. The observers solved this problem by studying every coded episode and using its content to match it to its corresponding action. For example, if a user made a comment about even exponents, it was assumed that this self-explanation behavior pertained to an exploratory action which
involved even exponents, even if such an action occurred slightly earlier. If a user made a verbal prediction about the outcome of the next action, this utterance had to be associated to the action that immediately followed. Thirteen coded episodes were discarded because the match was ambiguous.

While both parts of the above coding process resulted in the elimination of data points, the factor that had the greatest impact on the amount of data that could be obtained from the study was students’ willingness to verbalize their thoughts. A number of students were incapable or unwilling to think aloud, even if they were periodically reminded to do so. Without such verbalization, the coders could not tell whether a student had self-explained or not. Thus, of the 567 exploration cases recorded in the log files for all students, only 149 could be classified in terms of associated self-explanation.

Once positive and negative self-explanation episodes were identified and mapped onto specific exploration cases, an analysis could be carried out of the correspondence between these episodes, gaze information, and time students devoted to each case.

### 4.5.2 Time as a predictor of self-explanation

To analyze the relationship between time per exploration case and self-explanation in the plot unit, average times spent on exploration cases with and without self-explanation in this unit were computed and compared. When an exploratory action was accompanied by self-explanation, an average of 24.7 seconds elapsed before the next action. Exploratory actions without self-explanation were only followed by an average of 11.6 seconds of idle time. These results appear in table 4.3. The difference is statistically significant at the 0.05 level, suggesting that time per exploration case is actually a fairly reliable indicator of self-explanation.

<table>
<thead>
<tr>
<th></th>
<th>Self-explanation</th>
<th>No self-explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average time elapsed</td>
<td>24.7 s</td>
<td>11.6 s</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>13.89 s</td>
<td>9.59 s</td>
</tr>
</tbody>
</table>

To use time as a predictor of self-explanation, a threshold $T$ had to be determined so that an action could be classified as self-explained if and only if the student spent more than $T$ seconds on it. To choose the optimal threshold, a Receiver Operating Characteristic (ROC) curve (figure 4.1) was built. The ROC curve is a standard technique used in machine learning to evaluate the extent to which an information filtering system can successfully distinguish between relevant data (episodes the filter correctly classifies as positive, or true positives) and noise (episodes the filter incorrectly classifies...
as positive, or false positives), given a choice of different filtering thresholds. Figure 4.3 shows the ROC curve obtained for time, where each point on the curve represents a different threshold value. Points corresponding to even-numbered thresholds appear as small asterisks on the curve and several of these thresholds are labeled in the figure. As is standard practice, the final threshold was chosen as the point on the curve which corresponds to a reasonable tradeoff between creating too many false positives and creating too few true positives. In this case, this threshold is 16 seconds as labeled by the larger asterisk on the curve in figure 4.3. The reader should recall that both positive and negative self-explanations are verbalized, so higher time for positive self-explanation is not an artifact of verbalization.

![ROC curve for time as a filter for self-explanation. Asterisks indicate points corresponding to even-numbered thresholds](image)

**Figure 4.3 ROC curve for time as a filter for self-explanation. Asterisks indicate points corresponding to even-numbered thresholds.**

### 4.5.3 Gaze Shifts as a predictor of self-explanation

Raw eye-tracker data was parsed by a pattern detection algorithm developed to detect switches of attention ("gaze shifts") among the graph.
panel, the equation area, and any other non-conspicuous areas in the plot unit interface. As mentioned in the previous chapter, these are the gaze patterns hypothesized to be associated with self-explanation in the plot unit. A sample gaze switch appears in figure 4.4. Here a student's eye gaze (shown as the dotted line) starts in some untracked area below the screen, moves to the equation region and then hovers around the graph region above. The data-parsing algorithm uses fixation coordinates from the eye-tracker and matches them to appropriate ACE interface region. Next it searches the data for the pattern of looking at one region and then another, i.e. having a gaze shift. When this pattern is found, a tag is placed in the ACE log file to synchronize the switch with the appropriate exploration case.

An excerpt of such a synchronized log file appears in figure 4.5. Here the user begins by typing a new value for the slope into the equation (line 1 in the figure). Then several gaze shifts occur (lines 2-4). An *indirect gaze shift* happens when the user looks from the graph (or equation) region in the interface to an irrelevant region (e.g. the help menu on the screen or the keyboard on the table) and then to the equation (or graph) region. A *direct gaze shift* involves looking directly from one tracked region to another. Next the user moves the curve without looking down at the function region (line

\[ f(x) = -3 \left( \frac{1}{2} + 1.7 \right)^2 + 1.9 \]

Figure 4.4. Sample gaze shift

---

1 This algorithm was designed and implemented by Dave Ternes, an undergraduate research assistant in the computer science department at UBC.
5). Finally the user moves the curve again and then shifts her gaze down to the function region and back (lines 6-8).

Finally, a program is run which merges the synchronized log file with the coded data from the observers to create a file containing all data for one user in a concise form. An excerpt of such a file appears in Figure 4.6 below.

<table>
<thead>
<tr>
<th>Action</th>
<th>SE?</th>
<th>Gaze shift</th>
<th>Time&gt;16s?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typed new slope</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>moved power function</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>moved power function</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

Figure 4.6 Excerpt of user data file

4.6 Results

The dual histogram in figure 4.7 categorizes the 149 data point into episodes with and without self-explanation (99 circles and 50 triangles, respectively). The vertical line separates the points into those with and without a gaze shift between graph and equation plane in the plot unit. The horizontal line separates points with elapsed time above or below 16 seconds. The raw data also appears in table 4.4.

In order to determine whether or not the addition of gaze shift data caused a change to the optimal threshold set for time analysis, new ROC curve was constructed for time and gaze shifts as combined predictor. Here an exploratory action is classified as self-explained if it is accompanied by either 16 seconds of idle time or a gaze shift. The ROC curve, appearing in figure 4.8 shows that 16 seconds continues to be a reasonable threshold.
Figure 4.7. Dual histogram of experiment data.

Table 4.4 Raw experiment data

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaze shift</td>
<td>61</td>
<td>12</td>
<td>73</td>
</tr>
<tr>
<td>No gaze shift</td>
<td>38</td>
<td>38</td>
<td>76</td>
</tr>
<tr>
<td>Time &gt; 16</td>
<td>71</td>
<td>16</td>
<td>87</td>
</tr>
<tr>
<td>Time &lt;= 16</td>
<td>28</td>
<td>34</td>
<td>62</td>
</tr>
<tr>
<td>Gaze shift or Time &gt; 16</td>
<td>85</td>
<td>19</td>
<td>104</td>
</tr>
<tr>
<td>No gaze shift and Time &lt;= 16</td>
<td>14</td>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>50</td>
<td>149</td>
</tr>
</tbody>
</table>
Table 4.5 shows different measures of self-explanation classification accuracy if the predictor used is: (i) the eye-tracker to detect gaze shifts; (ii) time per exploration case; (iii) a combination of the two. Accuracy is reported in terms of true positive rate (i.e. percentage of positive self-explanation cases correctly classified as such, or sensitivity of the predictor) and true negative rate (i.e. percentage of negative self-explanation cases correctly classified as such, or the specificity of the predictor). A combined measure also appears, which is the average of the two accuracies. This average could also be weighted should it be decided that predicting the occurrence of self-explanation would be more important for ACE pedagogical proposes than predicting lack thereof (or vice versa).

![ROC curve for time and gaze shifts as a combined predictor](image)

Figure 4.8. ROC curve for time and gaze shifts as a combined predictor. Here asterisks indicate points corresponding to even-numbered thresholds

As the table shows, time alone has a higher sensitivity than gaze shift, i.e. the episodes involving self-explanation were more likely to take over 16 seconds than to include a gaze shift. However, the eye-tracker alone has comparably higher specificity, i.e. the cases without self-explanation were
more likely to involve the absence of a gaze shift than shorter time per exploration case. The two predictors have comparable combined accuracy.

Alternate analysis was performed to check if multiple gaze shifts would serve as a good predictor, with or without time. When two gaze shifts were required to indicate self-explanation, the specificity of the eye-tracker alone dropped to 51.7% and the sensitivity only rose to 79.5%, resulting in a combined accuracy of 60.6%. Adding time raised the combined accuracy to 67.9% but this is still lower than the results for a single gaze shift. Requiring more than two gaze shifts continued to lower the sensitivity to unacceptable levels.

Table 4.5 Classification Accuracy of different predictors

<table>
<thead>
<tr>
<th></th>
<th>Eye-tracker</th>
<th>Time</th>
<th>Eye-tracker + Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive Rate (sensitivity)</td>
<td>61.6%</td>
<td>71.7%</td>
<td>85.8%</td>
</tr>
<tr>
<td>True Negative Rate (specificity)</td>
<td>76.0%</td>
<td>68.0%</td>
<td>62.0%</td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>68.8%</td>
<td>69.85%</td>
<td>73.9%</td>
</tr>
</tbody>
</table>

4.7 Discussion

These results seem to suggest at first that the gain of using an eye-tracker is not worth the cost of adding this information to the ACE model. However, there are a few counterarguments to this conclusion.

First, it should be noted that time accuracy here is artificially high. One of the drawbacks of using time as a predictor of self-explanation is that the amount of time elapsed tells the model nothing about a student’s behavior between actions. During a long time spent on a given case, a student may be doing or thinking of something completely unrelated to ACE. This seldom occurs in this data, but it should be kept in mind that students were in a laboratory setting with little available distraction, in the presence of an observer and wearing a rather intrusive device on their head. All of these factors are likely to have made it more difficult for the students’ thoughts to wander from the task at hand, resulting in time being a more reliable indicator of self-explanation than it would be in actual practice.

Second, the sensitivity of the eye-tracker as a predictor may be artificially low due to errors in the eye-tracking data. For some students, calibration of the eye-tracker proved very difficult. This was especially the case for participants with heavy eyelashes or thick glasses. The eye-tracker would function with a reading of “GOOD CALIBRATION” or “POOR CALIBRATION” but not “CALIBRATION FAILED.” For several subjects, “POOR CALIBRATION” was the best that could be achieved. Also,
Chapter 4. Preliminary User Study

calibration could sometimes be compromised when a student sneezed or touched her face. Problems with calibration would make it more difficult for the eye-tracker to detect eye movements and thus some gaze shifts could go unreCORded. In fact, one student's data was entirely discarded immediately after her participation since, according to the eye-tracking data, her eyes stayed fixed on the same spot on the screen throughout her interaction with ACE. During her session with ACE, she performed many exploratory actions and made many comments consistent with self-explanation. Thus it was assumed that the eye-tracker must have been bumped or otherwise gotten shifted, causing a complete loss of calibration. These calibration problems are specific to a head-mounted device and would likely be less of an issue with a desk-mounted one. It should be noted that when we tried to recompute these accuracies using only the data points associated with the eleven students out of nineteen for whom "GOOD CALIBRATION" was achieved, sensitivity increased to 63.8%. Although this does not seem like a substantial increase, the reader should bare in mind that it is based on only 79 episodes - about 50% of the available data points - and it possibly included students whose calibration was compromised during the interaction with ACE.

In episodes where the student appeared to self-explain without making a gaze shift, pre-test performance showed that the student already understood the related concept before using ACE 73% of the time. Thus most of these students may have been merely telling the observer which mathematical phenomenon was relevant to the exploration rather than learning something new through observation and self-explanation.

Finally, combining gaze shift and time into one predictor substantially improves sensitivity. That is, if an action is classified as self-explained when there is either a gaze shift or more than 16 seconds elapsed time, most of the self-explanation episodes (85.8%) are correctly recognized. This increase also causes the combined accuracy to improve. However, as sensitivity increases, specificity is reduced and only 62% of the episodes that lack self-explanation are discovered by the model. This situation is shown in figure 4.3. With the combined model, all data points to the right of the vertical line or above the horizontal time threshold line are classified as self-explained. As a result, most of the episodes with self-explanation are found but many of those without self-explanation are incorrectly classified.

Here a tradeoff appears between sensitivity and specificity. Depending on how the system is used, it may be most important to correctly classify self-explanation when it occurs than to detect the lack thereof. This is the situation when letting natural self-explainers explore without interruption is given highest priority. Here, using the combination of eye-tracker and time data is best. Alternatively, it may be more important to make sure that the system intervenes wherever it is necessary. Then failing to identify lack of self-explanation is a bigger problem than failing to detect it when it occurs.
In this case, the eye-tracker alone is a more appropriate predictor because students who need help will be more likely to get it. However, it also means that more self-explanation episodes will be misclassified and spontaneous self-explainers will be more likely to be needlessly interrupted.

Given the above arguments, we felt that it is worthwhile adding eye-tracker data to the ACE model, and in such way that allows for flexibility in deciding which predictor (or combination of predictors) to use. In the next chapter we describe this stage of the research.
Chapter 5

The New ACE Student Model

This chapter describes the design and implementation of a revised ACE student model which includes the use of eye-tracking data to detect implicit self-explanation. The design of this model is based largely on the results of the user study discussed in the previous chapter.

5.1 Defining “SE-related Behavior”

As discussed in the previous chapter, each of the predictors of self-explanation we investigated has advantages and disadvantages. Depending on the circumstances surrounding the use of ACE, one predictor may be more appropriate than the others.

Time was shown in the last chapter to yield a reasonably high combined accuracy, slightly higher than that of the eye-tracker alone but not as high as that of the combined predictor. Both the specificity and the sensitivity of time are also high, suggesting that it is well-suited to situations where detecting positive and negative self-explanation are given equal importance. When it is more important to catch the absence of self-explanation and intervene, the eye-tracker alone has proven to be a better predictor. In other settings, it may be most important to detect self-explanation when it occurs so that natural self-explainers are not needlessly interrupted. Then the combined predictor is most appropriate.

Time is always simple and efficient to measure. In some situations an eye-tracker might not be available due its cost or other factors. Then time would be the only predictor available. However, there may also be settings in which ACE users are surrounded by such high levels of distraction (e.g. a noisy classroom) that time would perform exceptionally poorly. Then an eye-tracker would be needed to present a much clearer picture of the focus of the student’s attention.

These results suggest the need for a high degree of flexibility in the structure of the part of the ACE student model which assesses implicit self-explanation.
5.2 Adding "SE-related behavior" to the model

This section describes the structure of the portion of the student model which assesses student implicit self-explanation behavior. The choice of structure is based on the study results described earlier. The conditional probabilities associated with this part of the model are also determined from the results of the user study.

5.2.1 The naïve Bayesian classifier

The revised student model appears in figure 5.1 with a dashed triangle drawn around the part of the student model which was modified to include evidence of self-explanation. The gaze shift node has a binary value which indicates whether or not a gaze shift has occurred. Time is also a binary node indicating whether a time longer than the threshold M identified from the study data has elapsed, indicating sufficient time for self-explanation. All other nodes remain the same as shown in figure 3.4. Nodes representing the assessment of self-explanation are shaded grey.

The main advantage to this approach is that it is highly modular, allowing the gaze shift and time nodes to be included in the model or left out as needed. The disadvantage is that it assumes independence between time and the presence of gaze shifts, which is not necessarily true. In fact, our data actually suggests a slight positive correlation between the two. However, similar assumptions in pure naïve Bayesian classifiers have been shown in practice to perform surprisingly well even when this independence cannot be guaranteed.

5.2.2 Setting the Conditional Probabilities

Another advantage to using a naïve Bayesian classifier is the ease with which the associated conditional probabilities can be obtained from the data. In this part of the model, each node only has one parent, reducing the number of conditional probabilities that need to be determined. These can easily be calculated as the frequencies in the data from the user study described in the previous chapter:

\[
P(\text{time} > 16 \text{ seconds} | \text{implicit SE}) = 0.71
\]

\[
P(\text{time} > 16 \text{ seconds} | \text{no implicit SE}) = 0.32
\]

\[
P(\text{gaze shift} | \text{implicit SE}) = 0.61
\]
P(gaze shift | no implicit SE) = 0.24.

Further, these probabilities remain the same even as the gaze shift and time nodes are added and removed.

Figure 5.1 shows the revised student model over two time slices. In this model, after an exploratory action (e.g. the action represented by the node $e_0\text{Case}_1$ in slice T) occurs, time is kept and eye movements are monitored until the next action. When the next action is carried out a new slice is created; in parallel, if the new action is not an action indicating implicit self-explanation (using a selection of predefined self-explanations in an interface menu), the implicitSE node for the previous action is updated according to the stimuli to SE, time and gaze shift nodes. An implicitSE node is created for the previous action in slice T, along with time and gaze shift nodes with the appropriate values. These new nodes are used to assess the effectiveness of the exploration case, updating the corresponding node, $e_0\text{Case}_1$, in figure 5.2.
This update further propagates to the exploration of exercise and knowledge nodes.

Figure 5.2 The ACE student model for two time slices
Chapter 6

Testing the Student Model

This chapter discusses data collection and analysis carried out to test the performance of the revised ACE student model. First, section 6.1 presents early tests done to check the performance and consistency of the new student model. Section 6.2 describes the setting of additional conditional probabilities in the model using the data from the previous user study. Then section 6.3 describes the collection of further user data for use in tests of the new model. This was largely a continuation of the previous user study with some slight changes. Section 6.4 discusses early tests of the accuracy of the new student model. Finally, section 6.5 gives the results of cross-validation tests performed to check the stability of the new model and get a better estimate of the model's accuracy.

6.1 Initial tests of the new model

This section discusses early tests of the new model to check its performance and consistency. These tests were done primarily to make sure that the model was correctly implemented and that it behaved as expected. They were also intended to assess whether or not the revised structure was appropriate.

Here the data from the previous user study was used to evaluate the model. This training data was used to test basic model performance. The interaction data from each student (including eye-tracking data) was given to the model as input using a simulated student program. Then the resulting implicitSE node and exploration node probabilities were compared to the previously used self-explanation data points and post-test scores.

To get a better idea of the value of adding eye-tracking to the ACE student model, the two other previously developed models are also tested and compared in this chapter: (i) the one which does not assess self-explanation at all (model 1 from now on) [4] and (ii) the one which uses time only to detect implicit self-explanation (model 2 from now on) [5]. For purposes of consistency, the latter of these models was also assigned a time threshold of 16 seconds.

In order to test the new model (model 3 from now on), the SE tendency node and each of the knowledge nodes were assigned a generic prior probability of 0.5. Thus the model began with an identical assessment of the
knowledge and behavior of each student. Generic prior probabilities were also used.

To test model accuracy in assessing implicit self-explanation, we needed a threshold probability to use in deciding when an implicitSE node predicts the occurrence of self-explanation. This was derived from data from the previous study as follows. First the interaction data for each student was run through each of the models which assess self-explanation (i.e., models 2 and 3) using a simulated student program. The probabilities of implicitSE nodes were then compared against the coded data points from the first study. Each data point corresponds to a user action which the experts determined was or was not self-explained. The implicitSE node in each model (see Figures 3.4 and 5.2) also yielded probabilities that self-explanation occurred at the time of this action. These probabilities were compared to expert assessments to test the predictive performance of each model. For purposes of clarification, a small fragment of this data is given in Table 6.1 below.

Table 6.1 Values of implicitSE nodes corresponding to actions in study data

<table>
<thead>
<tr>
<th>Action</th>
<th>Experts SE assessment</th>
<th>Model 2 (time only)</th>
<th>Model 3 (time and gaze shifts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
<td>0.698</td>
<td>0.723</td>
</tr>
<tr>
<td>2</td>
<td>N</td>
<td>0.287</td>
<td>0.180</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>0.409</td>
<td>0.645</td>
</tr>
</tbody>
</table>

For each model, an optimal threshold M had to be determined such that the model showed that a student self-explained an action if and only if the implicitSE node associated with the action returned a probability over M. To find the best threshold, ROC curves were constructed. These appear in figure 6.1.

Each of these curves yielded an optimal threshold which was used to find the accuracies of each model. For the model which only uses time to detect self-explanation, this threshold was found to be 0.56 and for the new model which adds gaze shifts as evidence, the optimal threshold was 0.45.

Table 6.2 shows the sensitivity and specificity in each case. A combined measure, the average of the two accuracies, is also reported. Here the addition of the eye tracker causes an increase in each of the measures. It should also be noted that the largest increase is in specificity. This is consistent with the assumption that time as a predictor causes too many false positives by assuming that the user spends all of the time between her actions considering the outcomes of the actions.
All three models were then evaluated as predictors of sufficient exploration over the training set. Recall that in the preliminary user study, participants completed a post-test of mathematical concepts immediately after interacting with ACE. A correspondence was then created between questions on this post-test and domain concepts. These questions were then
use to determine each student’s aptitude in each of these concepts (e.g., positive intercepts and even exponents) at the end of the experiment. In addition, when a student’s log file was run through any of the three models, the final probabilities of the exploration of categories nodes (e.g., nodes positive intercepts and negative intercepts in figure 3.4) represent the model’s assessment that the student understands these concepts at the end of the interaction (i.e., that the student had adequately explored this material). These assessments could then be compared with the corresponding post-test scores to evaluate model accuracy over effective exploration. That is, if the model returned a high probability that the user understood a concept and the user correctly answered problems on the post-test that dealt with the concept, then the model was assumed to be a good predictor of sufficient exploration of the concept.

For each of the three models, a ROC curve was used to determine the best threshold at which an exploration node could be found to indicate adequate exploration – and thus knowledge – of the material. All three of these ROC curves are given in figure 6.2. The optimal thresholds found for the model without self-explanation, the model which only uses time to assess it, and that which uses eye tracking data as well are 0.66, 0.58 and 0.44, respectively.

Figure 6.2 ROC curves for models as predictors of sufficient exploration
Using these thresholds, the accuracies of the exploration nodes in each of the models could be computed. These appear in Table 6.3. These accuracies show that the new model is well-suited to assess adequate exploration in the training data. Also, each of the accuracies increased with each successive model, suggesting that the addition of self-explanation and gaze shift data were improvements to the model. These results also suggest that an increase in the accuracy of implicit self-explanation detection does in fact cause an increase in the accuracy of exploration assessment.

Table 6.3: Exploration node accuracies with generic prior probabilities using training data

<table>
<thead>
<tr>
<th></th>
<th>Model without SE</th>
<th>Model with time only For SE</th>
<th>Model with time and gaze shifts for SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True Positive rate</strong></td>
<td>68.8%</td>
<td>74.3%</td>
<td>77.4%</td>
</tr>
<tr>
<td><em>(sensitivity)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>True Negative rate</strong></td>
<td>59.4%</td>
<td>76.3%</td>
<td>78.3%</td>
</tr>
<tr>
<td><em>(specificity)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Combined Accuracy</strong></td>
<td>63.7%</td>
<td>75.3%</td>
<td>77.9%</td>
</tr>
</tbody>
</table>

It should also be noted that the increase in accuracy caused by the addition of the eye-tracker is higher for the implicitSE nodes than for the exploration nodes. This is due to the difference in the way each is measured. Each implicitSE probability is taken at the time that the actions occurs while only the probabilities of the exploration nodes at the end of the interaction are used in the analysis. Thus, while the implicitSE nodes represent the state of the user at a certain time and are strongly effected by the presence or absence of a gaze shift, the exploration nodes’ final probabilities are the result of many actions throughout the interaction and are influenced by other factors. Given these results, it can be concluded that the primary benefit in adding eye-tracking versus time only is the more accurate assessment of implicit self-explanation, which allows ACE to generate more precise real-time interventions during student interaction with the system.

### 6.2 Additional Training of the Student Model

After the revised model was tested for consistency, the data from the previous user study was used to set additional conditional probabilities in the student model.
When the user study was carried out, the ACE coach hints were turned off. Therefore, no information was gathered about the effectiveness of coach hints as a stimulus to self-explain. Further, no data was available concerning whether or not a stimulus to self-explain actually caused the student to self-explain.

In order to better fit this part of the model to the available data, the coach hint and stimuli to SE nodes were removed and the direct relationship between tendency to self-explain and implicit self-explanation was modeled instead. Two time slices of the resulting model appear in figure 6.3.

The 18 study participants were then divided into self-explainers – those who self-explained at least 20% of the time - and non-self-explainers – those who did not. This resulted in 14 self-explainers and 4 non-self-explainers. It was then found from the data that the self-explainers self-explained 79.8% of the time and the non-self-explainers self-explained 13.3% of the time. These frequencies were then used to set the following conditional probabilities in the model:

\[
P(\text{implicitSE} | \text{SE Tendency}) = 0.80
\]

\[
P(\text{implicitSE} | \text{no SE Tendency}) = 0.13.
\]
It should also be mentioned here that the pretest and post-test scores were compared for each student and a mean increase of 24% from the pretest to the post-test was found for the self-explainers while the non-self-explainers achieved a mean increase of only 5.7%. The increase in the self-explainers' improvement over the non-self-explainers' improvement was also determined to be statistically significant at the p<0.05 level, confirming that self-explanation has a significant effect on overall learning.

6.3 New User Study

Since the training data was used to derive the conditional probabilities used in the student model, new user data was needed to reliably test the accuracy of the model. Thus, we conducted a second study during which we collected data from 18 new students interacting with ACE as before and this study yielded the same form of data as in the previous study. The new participants were university students who had not taken any college-level math, also as before. However, two changes were made in an effort to solve earlier problems and improve the data overall.

First, the previous experiment showed that user's reluctance to think aloud made it difficult to glean episodes of positive and negative self-explanation from the study data. We thought this could be due in part to the fact that by the time these users got to the plot unit, they rushed through it, speaking as little as possible, because they were tired and eager to have the heavy and intrusive eye-tracker taken off. In order to get more data points, new study participants were asked to skip the arrow unit. Since the arrow unit serves as more of a test of the student's knowledge than as a chance for the student to explore and observe, it contributes little to a study of self-explanation behavior during exploration anyway. It was hoped that if students got to the plot unit sooner, they would stay longer and be more willing to think aloud. The machine unit was left in the experiment as it gave the users a chance to get used to thinking aloud before arriving at the plot unit.

Second, participants in the new study were asked to turn off the monitor and stare at a blank screen for several seconds at the end of the experiment. This was done in order to collect a baseline pupil size against which pupil size measurements throughout the interaction could be compared. This work is described in greater detail in Chapter 7.
6.4 Testing the New Student Model

This section discusses research carried out to evaluate the performance of the new student model with the revised structure as shown in Figure 6.3. Here analysis is performed to test the sensitivity of the model assessments to prior probabilities of the knowledge and SEtendency nodes. For simplicity, all tests are carried out over the new user data – the testing set.

6.4.1 Testing the Accuracy of the New Model with Different Priors

This section describes different tests of the accuracy of the student model using the new user study data - the testing set. First the SEtendency node and each of the knowledge nodes are assigned a prior probability of 0.5. Then tailored prior probabilities are used for knowledge nodes and in later tests, the prior probability of the SEtendency node is customized as well.

Throughout this subsection, all “training” of the model was done using the user data from the preliminary study of Chapter 4. As described in Chapter 5 and section 6.2, frequencies in this data were used to set conditional probabilities in the implicit self-explanation part of the new model. In addition, each accuracy presented here was based on a probability threshold (i.e., a value M for which a node was said to return a YES if it had a probability greater than M) which was derived from a ROC curve over the preliminary user data.

As in section 6.1, the exploration and SEtendency nodes were assigned generic prior probabilities of 0.5. Thus each model began with no initial information about each particular student. Generic prior probabilities were also used in the model which did not assess self-explanation and that which only used time.

Next the accuracy of the implicitSE nodes over the testing set was calculated for each model. The first two rows of Table 6.4 show the sensitivity and specificity in each case. A combined measure, the average of the two accuracies, is reported as well. Here the addition of the eye tracker causes a substantial increase in each of the measures. As in section 6.1, the largest increase is in specificity. This is consistent with the assumption that the use of eye tracking data will catch many of the false positives inherent in the use of time as a predictor.

In order to gain a more complete picture of the model’s implicitSE assessment over the testing data, a ROC curve was created for each model over this data. These curves appear in Figure 6.4. As discussed in [12], the area under a ROC curve is equal to the probability that a randomly selected positive case will be given a higher probability by the model than a randomly
selected negative case. Thus ROC curves with larger areas correspond to better predictors over the data.

Table 6.4 Accuracies of implicitSE nodes for different prior probabilities

<table>
<thead>
<tr>
<th>Generic prior probabilities for knowledge nodes</th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts for SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive rate (sensitivity)</td>
<td>65.1%</td>
<td>71.6%</td>
</tr>
<tr>
<td>True Negative rate (specificity)</td>
<td>62.6%</td>
<td>74.3%</td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>63.9%</td>
<td>73.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Customized prior probabilities for knowledge nodes</th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts for SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive rate (sensitivity)</td>
<td>65.6%</td>
<td>72.2%</td>
</tr>
<tr>
<td>True Negative rate (specificity)</td>
<td>62.8%</td>
<td>74.6%</td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>64.2%</td>
<td>73.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Customized prior probabilities for knowledge nodes and the SE tendency node</th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts for SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive rate (sensitivity)</td>
<td>68.8%</td>
<td>74.2%</td>
</tr>
<tr>
<td>True Negative rate (specificity)</td>
<td>65.2%</td>
<td>77.6%</td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>67.0%</td>
<td>75.9%</td>
</tr>
</tbody>
</table>

Figure 6.4 ROC curves for models as predictors of implicit self-explanation over testing data with generic prior probabilities
The first two rows of table 6.5 give the area of each curve as well as the standard error for each area. These values show improved performance with the addition of the eye-tracker. This improvement was found to be statistically significant at the z > 1.96 level, as indicated by the dotted line in the table.

Table 6.5 Areas of ROC curves for implicitSE nodes. Here the dotted lines indicate statistically significant differences.

<table>
<thead>
<tr>
<th>Generic prior probabilities for knowledge nodes</th>
<th>ROC curve area</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model with time only for SE</td>
<td>0.62</td>
<td>0.042</td>
</tr>
<tr>
<td>Model with time and gaze shifts for SE</td>
<td>0.75</td>
<td>0.036</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Customized prior probabilities for knowledge nodes</th>
<th>ROC curve area</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model with time only for SE</td>
<td>0.73</td>
<td>0.037</td>
</tr>
<tr>
<td>Model with time and gaze shifts for SE</td>
<td>0.84</td>
<td>0.029</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Customized prior probabilities for knowledge nodes and the SE tendency node</th>
<th>ROC curve area</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model with time only for SE</td>
<td>0.76</td>
<td>0.035</td>
</tr>
<tr>
<td>Model with time and gaze shifts for SE</td>
<td>0.88</td>
<td>0.025</td>
</tr>
</tbody>
</table>

For each model, the accuracies of the exploration nodes were then computed. These appear in the first three rows of table 6.6. Each of the accuracies increased with each successive model, suggesting that the addition of self-explanation and gaze shift data were in fact improvements. It also confirms here that an increase in the accuracy of implicit self-explanation detection does in fact cause an increase in the accuracy of exploration assessment.

For each model, ROC curves were then created for the exploration node assessments over the testing data. These appear in Figure 6.5. The areas under the curves and the standard error for each area are displayed in the first two rows of table 6.6. The differences between each of the areas were also found to be statistically significant at the z > 1.96 level, as indicated by the dotted lines in the table.

Before each student in the user study used ACE, she completed a pre-test which covered all of the relevant mathematical concepts. For the second test of the new model, these results were used to pick prior probabilities for the knowledge nodes for each student. If the student answered the corresponding pretest items correctly, a prior probability of 0.85 was assigned to the
corresponding knowledge node. Otherwise a probability of 0.15 was given. These values were chosen as reasonable estimates since they are close to 1 and 0 but they still allow for those students who guess correctly or make errors even though they understand the concept. Also, early analysis also showed that small changes in the prior probabilities (e.g., using 0.9 and 0.1 instead) have very little effect on the resulting model assessments.

Table 6.6 Accuracies of exploration nodes for different prior probabilities

<table>
<thead>
<tr>
<th></th>
<th>Model without SE</th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts for SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generic prior probabilities for knowledge nodes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positive rate (sensitivity)</td>
<td>62.7%</td>
<td>70.4%</td>
<td>73.9%</td>
</tr>
<tr>
<td>True Negative rate (specificity)</td>
<td>55.2%</td>
<td>71.5%</td>
<td>76.3%</td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>59.0%</td>
<td>71.0%</td>
<td>75.1%</td>
</tr>
<tr>
<td><strong>Customized prior probabilities for knowledge nodes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positive rate (sensitivity)</td>
<td>64.3%</td>
<td>72.2%</td>
<td>76.2%</td>
</tr>
<tr>
<td>True Negative rate (specificity)</td>
<td>58.8%</td>
<td>74.9%</td>
<td>78.4%</td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>61.6%</td>
<td>73.6%</td>
<td>77.1%</td>
</tr>
<tr>
<td><strong>Customized prior probabilities for knowledge nodes and the SE tendency node</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positive rate (sensitivity)</td>
<td>----</td>
<td>74.5%</td>
<td>78.1%</td>
</tr>
<tr>
<td>True Negative rate (specificity)</td>
<td>----</td>
<td>76.7%</td>
<td>79.6%</td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>----</td>
<td>75.6%</td>
<td>78.9%</td>
</tr>
</tbody>
</table>

With these customized prior probabilities, the accuracies of the implicitSE and exploration nodes were computed for each model. These accuracies appear in the middle three rows of Tables 6.4 and 6.4.

Each set of accuracies increases with the customization of the prior probabilities of the exploration nodes. However, the increases for the implicitSE nodes are smaller than for the exploration nodes as knowledge effects the assessment of sufficient exploration more directly than it effects the presence of implicit self-explanation in the model. As with the case for generic prior probabilities, each new model causes an increase in accuracy.

In addition, ROC curves were created for the implicitSE and exploration assessments of each model over the testing data. The middle two rows of tables 6.4 and 6.6 give the areas under each curve as well as the standard errors for each of these areas. For each model, customizing the knowledge
prior probabilities leads to an increase in area which is statistically significant at the $z > 1.96$ level, as indicated by the dotted line in the tables.

Figure 6.5 ROC curves for models as predictors of sufficient exploration over testing data with generic prior probabilities

As was done in section 6.2 for the original study data, the new study participants were divided into self-explainers – those who self-explained at least 20% of the time – and non-self-explainers – those who did not. This classification was used to customize the SE tendency node prior probabilities for each student. For self-explainers, this node was assigned a prior probability of 0.85 and for non-self-explainers, a value of 0.15 was used. As with the knowledge nodes, these values were chosen as they are close to 1 and 0 and tests revealed that the overall model has little sensitivity to small changes in these prior probabilities.

With the tailored tendency to SE nodes, the implicitSE and exploration node accuracies were determined for each model. These values appear in the lower three rows of Tables 6.4 and 6.6. In each of the models which included self-explanation, tailoring the prior probabilities of the SE tendency node caused an improvement of the performance in assessing both implicit self-explanation and sufficient exploration. It should also be noted that the accuracies were further improved by the addition of the eye-tracker.
Table 6.7 Areas of ROC curves for Exploration nodes. Dotted lines indicate statistically significant differences

<table>
<thead>
<tr>
<th></th>
<th>Model without SE</th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts for SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generic prior probabilities for knowledge nodes</strong></td>
<td>ROC curve area</td>
<td>0.56</td>
<td>0.68</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.036</td>
<td>0.033</td>
<td>0.029</td>
</tr>
<tr>
<td><strong>Customized prior probabilities for knowledge nodes</strong></td>
<td>ROC curve area</td>
<td>0.64</td>
<td>0.73</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.034</td>
<td>0.031</td>
<td>0.024</td>
</tr>
<tr>
<td><strong>Customized prior probabilities for knowledge nodes and the SE tendency node</strong></td>
<td>ROC curve area</td>
<td>0.79</td>
<td>0.88</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.027</td>
<td>0.021</td>
<td></td>
</tr>
</tbody>
</table>

For each model, ROC curves were the constructed for the implicitSE and exploration assessments over the testing data. The last two rows of tables 6.4 and 6.6 give the areas under each curve as well as the standard errors for each of these areas. For each model, the increase in area which resulted from customizing the SEtendency prior probabilities failed to achieve statistical significance.

This is perhaps due to the fact that while SE Tendency is maintained in the model as a probability that changes over time and always depends on previous values, the dependence between the current tendency to self-explain and the presence or absence of self explanation is very strong (recall that P(implicitSE | SE Tendency) = 0.80 and P(implicitSE | no SE Tendency) = 0.13). Thus, for each action, SEtendency is greatly effected by such factors as time and gaze shifts and the effects of the prior probability can dissipate over time. It should also be mentioned that it is difficult to select an appropriate value for this prior probability as many students seem to self-explain about half the time. Then assigning a prior probability of 0.85 is actually less accurate than the originally used value 0.5.

In summary, the addition of the eye-tracker causes a significant improvement in the performance of the model as a predictor of implicit self-explanation and sufficient exploration over the testing data. Customizing the prior probabilities of the knowledge nodes also significantly increases model accuracy. However, customizing the prior probabilities of the SEtendency node fails to produce a significant improvement.
6.5 Cross-validation Analysis

This section describes cross-validation analysis carried out to test the overall stability of the new model and get a better sense of the strength of the accuracy results of the last section.

Three different types of cross-validation were performed in each of the experiments described in this section. The first was leave-one-out in which the model was trained — i.e. the conditional probabilities were set from frequencies – on the data from all but one student. Then the model was tested over the data from the remaining student. This was done for each of the 36 students and results from each student were averaged. This type of cross-validation is the most important because it gives the best sense of how the model behaves for one student.

The second type of cross-validation was leave-half-out in which the students were randomly divided in half into a training set and a testing set. Frequencies from the first set were used to set conditional probabilities in each model, which was then tested over the latter set. This was performed 40 times and average accuracy over the student was calculated. This type of cross-validation shows the mean behavior of the model across a group of students.

Finally leave-half-points-out involved treating all of the data points as belonging to one student and then randomly dividing them into a testing set and a training set. This was done because some students had more data points than others and this ensures that the data is truly divided in half. In addition, it gives an alternate view of the performance of the models over the data independent of which students are being used. As with leave-half-out, 40 folds of the cross-validation were performed and the accuracies were averaged.

6.5.1 Customizing prior probabilities

The first set of tests was performed to determine if and how the prior probabilities for exploration nodes should be assigned. The decision was made at this point to always use generic prior probabilities for SEtendency nodes both because it is unrealistic in practice to assume that a student’s self-explanation tendencies have been tested before using ACE and because the previous tests showed that the custom priors do not significantly improve accuracy. However, mathematical competence can easily be tested through the use of pretests or some other means. Since we found that customized SEtendency priors failed to significantly improve accuracy, we chose not to explore this type of customization further.
Here we ran model 3 with three types of prior probabilities for knowledge nodes: (i) the generic setting of 0.5, (ii) a population prior probability based on the average pre-test performance of the test set and (iii) a probability (0.15 or 0.85) based solely on the pre-test score of the individual student.

Table 6.8 Cross-validation results for implicitSE nodes using various prior probabilities

<table>
<thead>
<tr>
<th></th>
<th>Generic prior probabilities</th>
<th>Population prior probabilities</th>
<th>Custom prior probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean accuracy</td>
<td>Std dev</td>
<td>Mean accuracy</td>
</tr>
<tr>
<td>Leave-one-out</td>
<td>71.6%</td>
<td>7.9%</td>
<td>72.5%</td>
</tr>
<tr>
<td>Leave-half-out</td>
<td>70.4%</td>
<td>9.7%</td>
<td>73.6%</td>
</tr>
<tr>
<td>Leave-half-pts-out</td>
<td>71.7%</td>
<td>8.5%</td>
<td>73.8%</td>
</tr>
</tbody>
</table>

Each type of cross-validation was carried out and the accuracy results for the implicitSE nodes are shown in table 6.8. Here only the combined accuracies are given for purposes of brevity. Each type of cross-validation shows that performance improves as the prior probabilities of the knowledge nodes are increasingly customized. ANOVA tests showed statistical significance in the differences within the set of results but T-tests were performed to find that, in each case, only the increase in accuracy from generic priors to custom priors is statistically significant to the p<0.05 level. Customizing prior probabilities to individual students also improves model stability, as shown by the decreases in standard deviation.

It should also be noted that these results are very similar across the different types of cross-validation. This is due to the fact that each value is the average of many accuracy values which were based from model assessments from the same group of students.

Table 6.9 Cross-validation results for exploration nodes using various prior probabilities

<table>
<thead>
<tr>
<th></th>
<th>Generic prior probabilities</th>
<th>Population prior probabilities</th>
<th>Custom prior probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean accuracy</td>
<td>Std dev</td>
<td>Mean accuracy</td>
</tr>
<tr>
<td>Leave-one-out</td>
<td>71.6%</td>
<td>11.6%</td>
<td>72.6%</td>
</tr>
<tr>
<td>Leave-half-out</td>
<td>72.7%</td>
<td>9.4%</td>
<td>73.2%</td>
</tr>
<tr>
<td>Leave-half-pts-out</td>
<td>70.7%</td>
<td>10.6%</td>
<td>72.9%</td>
</tr>
</tbody>
</table>

The results for the exploration nodes appear in table 6.9. As with the implicitSE nodes, accuracy increases and prior probabilities go from generic...
to population to custom. However, only the increase caused by changing the priors from generic to custom is statistically significant. Stability also increases as prior probabilities are customized.

6.5.2 Comparing the different models

Here the student model without self explanation, the model which only uses time to assess it, and the new model which uses gaze shifts as well are tested and compared. In each case, individual prior probabilities are used for the knowledge nodes in each model.

Table 6.10 Cross-validation results for implicitSE nodes using various student models

<table>
<thead>
<tr>
<th></th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean accuracy</td>
<td>Std dev</td>
</tr>
<tr>
<td>Leave-one-out</td>
<td>65.8%</td>
<td>7.4%</td>
</tr>
<tr>
<td>Leave-half-out</td>
<td>66.3%</td>
<td>8.3%</td>
</tr>
<tr>
<td>Leave-half-pts-out</td>
<td>65.7%</td>
<td>7.2%</td>
</tr>
</tbody>
</table>

Each type of cross-validation was carried out and the accuracy results for the implicitSE nodes and exploration nodes are shown in tables 6.10 and 6.11, respectively. For each type of node, the addition of self-explanation and then eye-tracking increased model accuracy and stability. ANOVA tests also showed statistical significance within the set and pairwise T-test confirmed that each increase was significant at the p < 0.05 level.

Table 6.11 Cross-validation results for exploration nodes using various student models

<table>
<thead>
<tr>
<th></th>
<th>Model without SE</th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean accuracy</td>
<td>Std dev</td>
<td>Mean accuracy</td>
</tr>
<tr>
<td>Leave-one-out</td>
<td>57.3%</td>
<td>11.6%</td>
<td>65.3%</td>
</tr>
<tr>
<td>Leave-half-out</td>
<td>56.8%</td>
<td>12.6%</td>
<td>66.3%</td>
</tr>
<tr>
<td>Leave-half-pts-out</td>
<td>57.8%</td>
<td>11.9%</td>
<td>67.3%</td>
</tr>
</tbody>
</table>

6.5.3 Calculating Optimal Thresholds

As previously noted, the leave-one-out cross-validation was essential because it gave the best sense of how the model would perform for a single user. In this process, one student would first be chosen as the one to be “left out.”
Then all necessary training would be carried out using the remaining data. This is included setting frequencies in the model and determining optimal thresholds via ROC curves. Thus, in addition to the stability results just described, this analysis yielded an optimal threshold for each student and each model.

This information was used to choose a “final” threshold for each type of model. Since a training set would not be appropriate or available for ACE system use in practice, it was necessary to make such a selection. This was carried out as follows: for each model and type of prior, leave-one-out cross-validation resulted in 36 optimal thresholds (one per student). These thresholds were then averaged and these means appear in tables 6.12 and 6.13 with standard deviations.

Table 6.12 Mean optimal thresholds and standard deviations for implicitSE node assessments across folds of leave-one-out cross-validation

<table>
<thead>
<tr>
<th></th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Threshold</td>
<td>Std dev</td>
</tr>
<tr>
<td>Generic knowledge node priors</td>
<td>0.49</td>
<td>0.0048</td>
</tr>
<tr>
<td>Custom knowledge node priors</td>
<td>0.52</td>
<td>0.0055</td>
</tr>
</tbody>
</table>

As the low standard deviations in the tables show, the optimal thresholds chosen in each fold of the cross-validation are quite similar. This is due to the fact that set of thresholds results from ROC curves formed using model assessments over almost identical data sets (i.e., each data set differs by one student). The low standard deviation also confirms that for new users of the system, these mean thresholds are appropriate to use in interpreting model assessments.

Table 6.13 Mean optimal thresholds and standard deviations for exploration node assessments across folds of leave-one-out cross-validation

<table>
<thead>
<tr>
<th></th>
<th>Model without SE</th>
<th>Model with time only for SE</th>
<th>Model with time and gaze shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Threshold</td>
<td>Std dev</td>
<td>Mean Threshold</td>
</tr>
<tr>
<td>Generic knowledge node priors</td>
<td>0.59</td>
<td>0.0084</td>
<td>0.52</td>
</tr>
<tr>
<td>Custom knowledge node priors</td>
<td>0.65</td>
<td>0.0077</td>
<td>0.56</td>
</tr>
</tbody>
</table>
6.6 Testing the Performance of the Models using Different Evidence of Self-explanation

As previously discussed, the main advantage to the structure of the new model is that it is highly modular, allowing the time and gaze shift nodes to be used or ignored as needed. This section presents tests performed to assess the accuracy of the model using time alone and using gaze shifts alone to detect implicit self-explanation.

First the ACE student model was adjusted to ignore the gaze shift node. Note that while this model uses time alone for implicit self-explanation, it differs in structure from model 2 (figure 3.4). Another version of the model was then similarly created which only uses gaze shifts as evidence of implicit self-explanation.

As previously done, the data from the preliminary user study was sent into each model using a simulated student program. Then ROC curves were created for the implicitSE nodes of each model over this data. These curves yielded optimal thresholds which could then be used to test the performance of each model over the new user data. These results appear in table 6.14 below. For purposes of comparison, the accuracies for the model which uses time and gaze shifts as evidence are also presented.

Table 6.14. ImplicitSE accuracies for the ACE model using different predictors as evidence of implicit self-explanation

<table>
<thead>
<tr>
<th></th>
<th>Model with time only</th>
<th>Model with eye-tracking only</th>
<th>Model with time and eye-tracking</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive rate</td>
<td>67.9%</td>
<td>62.3%</td>
<td>71.6%</td>
</tr>
<tr>
<td>(sensitivity)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True negative rate</td>
<td>64.8%</td>
<td>67.8%</td>
<td>74.3%</td>
</tr>
<tr>
<td>(specificity)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>66.3%</td>
<td>65.1%</td>
<td>73.0%</td>
</tr>
</tbody>
</table>

As shown in the table, the model which uses time alone has a higher sensitivity than the model with gaze shifts alone while the latter model returns a higher specificity. These findings match those of the original user study. They are also consistent with the assumption that time overestimates self-explanation behavior by assuming that the user spends all idle time considering the exploration. For each measure, the combined predictor outperforms either on its own.

As before, ROC curves were created for the exploration nodes of each model over the original study data. The resulting thresholds were then used to calculate the accuracies of each model over the new data. These results are given in table 6.15.
As with the *implicitSE* nodes, the model with time alone has a higher sensitivity than the model which only uses gaze shifts. However, the model with gaze shifts alone achieves higher specificity. These predictors combine to yield the highest accuracy for each measure. This is due to the fact that accuracy improves with more evidence used. It should also be noted that each single predictor seems to succeed where the other fails so this complimentary behavior likely contributes to the high accuracy of the combined predictor.

Table 6.15 Exploration accuracies for the ACE model using different predictors as evidence of implicit self-explanation

<table>
<thead>
<tr>
<th></th>
<th>Model with time only</th>
<th>Model with Eye-tracking only</th>
<th>Model with time and eye-tracking</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive</td>
<td>71.2%</td>
<td>69.8%</td>
<td>73.9%</td>
</tr>
<tr>
<td>rate (sensitivity)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True negative</td>
<td>72.9%</td>
<td>73.4%</td>
<td>76.3%</td>
</tr>
<tr>
<td>rate (specificity)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined Accuracy</td>
<td>72.1%</td>
<td>71.6%</td>
<td>75.1%</td>
</tr>
</tbody>
</table>

It should also be noted that the accuracies generated by the new model when only time information is used are comparable to (although slightly higher than) the accuracies of model 2, despite the differences in structure and method of CPT definition (data-based for model 3 and expert-based for model 2).
Chapter 7

Pupil Dilation as a Predictor of Self-Explanation

Previous research has demonstrated a positive correlation between a person’s pupil dilation and cognitive load in a wide variety of tasks [2]. People performing tasks requiring more attention have been shown to have greater pupil size.

Recent work has been done to evaluate the performance of pupil dilation as an indicator of cognitive load of users of adaptive systems, with controversial results. Schultheis and Jameson [26] led a study to analyze the pupil sizes of users of adaptive hypermedia systems. Participants were asked to read text samples of varying difficulty and their pupil sizes were examined throughout the task. The difference in text difficulty – and thus cognitive load – was not apparent in pupil diameter. So in this study pupil size failed to be a reliable indicator of cognitive load.

Similarly, Iqbal et al [16] examined the sensitivity of pupil size to cognitive load as users performed different tasks, including file manipulation and the reading of text. As in [26], pupil size failed to be an accurate indication of cognitive load during the reading task. However, it was also found that pupil size was sensitive to task difficulty only during certain subtasks of the file management task. Unlike the file management task, the reading of text could not be decomposed into smaller, independent subtasks. Thus, while pupil size may be sensitive to cognitive load during some tasks, it had to be concluded that it is not a suitable indicator of difficulty during the reading of text.

This chapter describes research performed to determine whether pupil size may be used as an additional predictor of self-explanation behavior in users of the ACE system. Section 7.1 details the process of trying to collect reliable pupil dilation data during the user studies and in section 7.2, the analyses and results are presented.

7.1 Data Collection

In the user studies described in Chapter 4 and 6, each session yielded a log file containing data from the eye-tracker. In addition to gaze data such as the time and location of each of the user’s fixations, these log files included the
diameter of the user’s pupil throughout the interaction. However, as the preliminary study in Chapter 4 was focused entirely on testing the merit of gaze patterns as predictors of self-explanation, no effort was taken to attend to differences in pupil behavior between study participants or environmental changes between sessions. It was thus impossible to perform any reliable analysis of this data.

Several factors have long been known to influence pupil size [2]. For example, pupils tend to dilate in darker rooms and contract in the presence of more light. Pupil size also naturally fluctuates with the size of the eye itself from one person to another. Even nonvisual stimuli such as sound can affect pupil dilation.

During the study, pupil sizes results were given by the eye-tracker in units of number of pixels in the pupil image. Thus the pupil size readings for each student were also influenced by the placement of the eye-tracker itself: if the camera was placed closer to the participant’s eye, all pupil size readings increased.

While it is not possible to keep environmental conditions adequately constant from one study session to the next or to ensure that users have similar pupils, steps may be taken to normalize against these differences and get a pure correlation between pupil size and cognition across all study participants. One method which has been used and accepted [26] involves taking a baseline measurement – i.e., measuring the pupil dilation at a time in the session when all users should have the same cognitive state. This baseline encapsulates all information regarding the pupil attributes of the user and environmental conditions such as lighting. For each user, the baseline may be subtracted from all other pupil size values to yield a normalized measure which may then be compared to those of users from other sessions.

As mentioned in section 6.3 a second user study was performed in order to gather additional data for testing of the ACE student model and address problems that appeared in the preliminary study (e.g., users’ reluctance to think aloud). Efforts were also taken during this study to collect reliable pupil size data using the baseline normalization method described above.

To get a baseline pupil size value, participants were asked to turn off the computer monitor and stare at the blank screen for a few seconds after completing their interaction with ACE. This baseline was chosen for several reasons. First, staring at a blank screen would ensure that each user had the same visual stimuli at this point in the session. It was also assumed that, while sitting idle at the close of the ACE interaction, the participants shared the same cognitive state. Finally, turning off the eye-tracker while the user was staring at the blank screen caused the baseline value to be the last pupil size measurement recorded by the eye-tracker. While not necessary, this eliminated any error that might occur in trying to find the moment of baseline measurement in the eye-tracking data.
Chapter 7: Pupil Dilation as a Predictor of Self-Explanation

With a baseline value obtained for each of the 18 participants in the new study, analysis could be carried out to explore the correlation between pupil dilation and self-explanation. This is discussed in the next section.

7.2 Results and Discussion

Recall from Chapter 6 that the second user study resulted in a set of exploration cases in which observers determined whether self-explanation was definitely present or absent. For each of these data points, the user’s average pupil size after the action but before the next action was calculated. This was carried out as follows: during the time interval between actions, the learner made a series of eye fixations, each of which was recorded by the eye-tracker. In addition to location coordinates, the data for each fixation included the length of the fixation in milliseconds and the user’s pupil size (given as image area in pixels) when it occurred. For each of these fixations, the pupil size and duration were multiplied. These values were then summed and divided by the total elapsed time, resulting in a weighted average pupil size between actions. Finally, the baseline measurement was subtracted, resulting in a normalized value. A sample of this data appears in Table 7.1. Note the each normalized value in the table is negative. While one might expect that users had low cognitive load – and thus small pupil size - at the moment the baseline recording, this could no be guaranteed. For the purposes of a baseline it was only necessary that students be in the same cognitive state. Further, students would naturally have larger pupils when staring at a blank monitor than when looking at a lit screen.

Table 7.1. Sample fragment of pupil size data

<table>
<thead>
<tr>
<th>User #</th>
<th>Self-Explanation?</th>
<th>Mean pupil size after action</th>
<th>User baseline</th>
<th>Normalized value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Y</td>
<td>1321</td>
<td>1391</td>
<td>-70</td>
</tr>
<tr>
<td>7</td>
<td>Y</td>
<td>1359</td>
<td>1391</td>
<td>-32</td>
</tr>
<tr>
<td>8</td>
<td>N</td>
<td>550</td>
<td>606</td>
<td>-56</td>
</tr>
<tr>
<td>8</td>
<td>Y</td>
<td>584</td>
<td>606</td>
<td>-22</td>
</tr>
</tbody>
</table>

After determining a normalized average pupil size value for each of these exploration cases, calculations were done to determine whether there was a significant difference in the pupil size of users who were self-explaining and those who were not. These results are shown in Table 7.2. Here we see that users had a mean normalized pupil size of -56 when they self-explained and -59 when they did not. However, due to the large standard deviations, this
difference fails to achieve statistical significance (as measured by a two-tailed T-test).

Table 7.2. Mean normalized pupil sizes with and without self-explanation

<table>
<thead>
<tr>
<th>Self-Explanation?</th>
<th>Mean normalized pupil size</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>-56</td>
<td>14.4</td>
</tr>
<tr>
<td>No</td>
<td>-59</td>
<td>10.5</td>
</tr>
</tbody>
</table>

A ROC curve was also created for these points to give a graphical representation of the performance of pupil size as a predictor of self-explanation for different normalized pupil size thresholds. This curve appears in figure 7.1 below. The area under this curve is 0.43, less the area of 0.5 given by random chance. Thus, over this data, pupil size is definitely not an acceptable predictor of self-explanation.

There are several possible causes for the failure of pupil size as a reliable detector of self-explanation in this data. It was assumed earlier that all users who stared at the blank screen after interacting with ACE were at approximately the same cognitive state. This might not have been a correct assumption, in which case the baseline used may not have been appropriate. To test this theory, Z-scores were used to normalize pupil sizes between participants. Each mean pupil size was normalized according to the following formula:

\[
X_k' = \frac{X_k - m}{s}
\]

where \(X_k'\) is the normalized mean pupil size after action \(k\), \(X_k\) is the mean pupil size after action \(k\) (these values appear in the third column of tables 7.1), \(m\) is the average value of \(X_k\) over all actions by the user, and \(s\) is the standard deviation corresponding to \(m\). While the mean pupil size \(m\) encapsulates all information which would effect the user's pupil size (i.e., lighting conditions and eye size), it is not specific to any single moment during the experiment and thus does not require the assumption that we identify the point at which all users are in the same state, as is required for the baseline method. A sample fragment of this data appears in table 7.3.
Chapter 7. Pupil Dilation as a Predictor of Self-Explanation

Figure 7.1. ROC curve for pupil size as a predictor of self-explanation

As before, calculations were done to determine whether there was a significant difference in the pupil size of users who were self-explaining and those who were not. These results appear in Table 7.4. Here we see that when users had a mean normalized pupil size of 0.928 when they self-explained and 0.812 when they did not. However, due to the large standard deviations, this difference also fails to achieve statistical significance.

Table 7.3. Sample fragment of pupil size data using Z-score normalization

<table>
<thead>
<tr>
<th>User #</th>
<th>SE?</th>
<th>Mean pupil size After action</th>
<th>Overall mean pupil size for user</th>
<th>Standard Deviation</th>
<th>Normalized value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Y</td>
<td>1321</td>
<td>1318</td>
<td>26.4</td>
<td>0.114</td>
</tr>
<tr>
<td>7</td>
<td>Y</td>
<td>1359</td>
<td>1318</td>
<td>26.4</td>
<td>1.553</td>
</tr>
<tr>
<td>8</td>
<td>N</td>
<td>550</td>
<td>562</td>
<td>17.3</td>
<td>-0.693</td>
</tr>
<tr>
<td>8</td>
<td>Y</td>
<td>584</td>
<td>562</td>
<td>17.3</td>
<td>1.272</td>
</tr>
</tbody>
</table>

A ROC curve was also constructed for these points to give a graphical representation of the performance of pupil size as a predictor of self-
explanation for different Z-score normalized pupil size thresholds. This curve appears in figure 7.2. The area under this curve is 0.49, about the same as that given by random chance. Thus, applying the Z-score normalization, pupil size is still not an acceptable predictor of self-explanation.

Table 7.4. Mean Z-score normalized pupil sizes with and without self-explanation

<table>
<thead>
<tr>
<th>Self-Explanation?</th>
<th>Mean normalized pupil size</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.928</td>
<td>0.37</td>
</tr>
<tr>
<td>No</td>
<td>0.812</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Figure 7.2. ROC curve for pupil size, normalized with Z-scores, as a predictor of self-explanation

It was then hypothesized that perhaps positive self-explanation caused the greatest increase in cognitive load – and thus pupil size – immediately after the outcome of the corresponding exploratory action. Here the user would have just noticed the change in the interface and would not yet be planning the next action. In this case, examining the weighted mean pupil size over the whole time interval between actions would fail to capture this behavior as the pupil size increase would be dissipated by the smaller pupil size values at the end of the time interval. Figure 7.3(a) shows a plot of the pupil size in this case. Another possibility considered was that the greatest increase in
cognitive load occurred at the end of the time interval, after the user had noticed the action outcome and had time to consider its meaning. These two cases appear together in figure 7.3(b). In either case, taking the mean pupil size over only the middle portion of the interval, as shown in the figure, would best capture the pupil size increase.

![Figure 7.3](image)

**Figure 7.3.** Pupil size plotted over the time interval when it is largest at the beginning (a) or at the beginning or the end (b). In (b), dotted lines indicate the middle 60% of the time interval.

To test this theory, normalized weighted pupil sizes were calculated (using Z-score normalization) over only the middle 60% of each time interval. Analysis was then performed to test for differences in pupil sizes in the presence or absence of self-explanation. These results appear in table 7.5.

**Table 7.5.** Mean Z-score normalized pupil sizes with and without self-explanation over only the middle 60% of each time interval

<table>
<thead>
<tr>
<th>Self-Explanation?</th>
<th>Mean normalized pupil size</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.861</td>
<td>0.41</td>
</tr>
<tr>
<td>No</td>
<td>0.794</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Here users had a mean normalized pupil size of 0.861 when self-explaining and a mean pupil size of 0.794 otherwise. However, T-tests failed to establish statistical significance in this difference given the large standard deviations. Figure 7.4 shows a ROC curve for this data. The area under this curve was found to be 0.41, again less than that which would have resulted from random chance.

Ambient sounds (e.g., people walking in the hall outside of the research laboratory) were not something that could be controlled and may have skewed the results. It is also possible that there may have been cases when a
user was not self-explaining but their cognitive load was relatively high because they were thinking about something other than the domain material. A third explanation is that for some of the participants, the spontaneous self-explainers in particular, self-explanation comes quite naturally and thus does not involve an increase in cognitive load at all.

Figure 7.4. ROC curve for pupil size as a predictor of self-explanation, using the middle 60% of the time interval after each action
Chapter 8

Conclusions and Future Work

8.1 Satisfaction of Thesis Goals

The ultimate goal of this thesis is to improve the monitoring and support of effective exploration in an open learning environment. By using eye-tracking to gain a more accurate assessment of implicit self-explanation, we improve the system's ability to provide tailored support for self-explanation in the form of gentle prompts and intervention. In addition, an improved assessment of self-explanation brings about a better assessment of sufficient exploration, which improves the system's ability to support effective exploration. Together these lead to a richer, more adaptive exploratory experience.

Recall from section 1.4 that this research began with a user study to test the reliability of elapsed time and gaze shifts as predictors of self-explanation. A new student model was then designed and built around these results and analysis was perform to assess the accuracy of this new model.

8.1.1 User Study to Assess Time and Gaze Shifts as Predictors of Self-Explanation

The research described in this thesis began with a preliminary user study to test whether elapsed time and gaze shifts could be used as reliable indicators of self-explanation behavior. Participants in this experiment interacted with the ACE system, an exploratory learning environment for mathematical functions, while their exploratory behavior and eye movements were captured in log files. In addition, video recordings were made which included the ACE interface screen and the learner's speech throughout the interaction.

Data analysis was then performed to classify each utterance, if possible, as evidence of the presence or absence of implicit-self-explanation. This was done using a detailed coding scheme developed for this purpose. While coding schemes exist for self-explanation study during example studying and problem solving, this was the first attempt to evaluate self-explanation during independent exploration. Those utterances that were definitively classified were then matched to the exploratory action to which they pertained. Each of these actions was then examined for time taken before the next action and the
presence or absence of a gaze shift. This allowed an assessment of the reliability of these as predictors of implicit self-explanation.

Accuracies were then calculated to show that gaze shifts alone had the highest specificity, i.e., that they served as a better detector of the absence of self-explanation. On the other hand, time and gaze shifts as a combined predictor had higher sensitivity, i.e., they performed best together at identifying self-explanation when it occurred. Thus it would be better to use time and gaze shifts together when priority is given to minimizing unnecessary interruption by the system. Similarly, gaze shifts alone are the best predictor when it is more important to catch the absence of self-explanation and intervene as needed.

8.1.2 Design of Revised Student Model

As suggested by the results of the preliminary study, gaze shifts and time combine to serve as the best predictor of implicit self-explanation in some situations and in others it is preferable to use gaze shifts on their own. In addition, certain environmental factors may exist to influence the choice of predictor. For example, in situations where an eye-tracker is not available, due to cost or other reasons, it is necessary to use time on its own. There may also exist situations (e.g., a busy office) where distractions are great enough to render time useless as a predictor. Then gaze shifts would need to be used alone. Thus some structural flexibility is necessary in the part of the student model which assesses implicit self-explanation.

To address these concerns, the original ACE student model was modified to include a highly modular component very similar to a naïve Bayesian classifier. This revised structure allows time and gaze shifts to be easily added or excluded from the model. In addition, the conditional probabilities were easily derived from frequencies in the user data.

8.1.3 Evaluation of Performance of New Student Model

In order to get an initial assessment of the behavior and consistency of the new ACE student model, this model was first tested using the user data from the original user study. This data was run through the new model (model 3) as well as an earlier model which did not include self-explanation (model 1) and one which only used time to assess it (model 2).

Accuracy results showed that the addition of self-explanation and then eye-tracking caused improvements in model performance. The improvement was particularly high for sensitivity, the measure of the models' abilities to detect the absence of implicit self-explanation. This is consistent with the theory that time as a predictor overestimates self-explanation behavior by
assuming that the user spends all the time between actions considering their outcomes.

After these initial tests, the student model was further refined to allow training of the part which assesses a user’s tendency to self-explain. The original version of this part [5] was based on subjective intuition and reasonable estimates rather than user data. Frequencies in the data were then used to determine conditional probabilities in the revised model.

An additional user study – almost identical to the first – was carried out to collect test data to assess the performance of the revised ACE model. Here each of the three models was tested in its ability to accurately assess implicit self-explanation (when applicable) and sufficient exploration.

These models were tested over the new data using two methods. First, this data was run through each model using a simulated student program and model assessments were compared to observed cases of self-explanation and post-test performance. This resulted in accuracy rates for each model. Second, ROC curves were constructed for each student model over the data. These curves provided a graphic picture of the behavior of each model and the area under each curve served as a measure of model performance.

Each type of analysis showed that accuracy increased with the addition of self-explanation and then eye-tracking. Adding the eye-tracker was found to significantly improve the accuracy of self-explanation assessment over using time only. In addition, the improved accuracy of self-explanation prediction resulted in a significant improvement in the assessment of student exploratory behavior. Customizing the prior probabilities of the knowledge nodes was also shown to significantly improve model performance. While customizing the prior probabilities of the SEtendency nodes increased model accuracy, this improvement failed to achieve statistical significance.

In order to test the stability and reliability of these results, cross-validation was performed for each model. Here the leave-one-student-out, leave-half-students-out and leave-half-data-points-out methods were each used. Each type of analysis showed that model accuracy increased with the addition of self-explanation and the eye-tracking data. In addition, T-tests revealed that these increases were statistically significant at the p<0.05 level.

8.2 Summary of Contributions

In this thesis, research was presented on the use of real-time eye-tracking data for the on-line modeling of user meta-cognitive behaviors during interaction with an open learning environment. It was shown that the addition of eye-tracking caused a significant improvement in the model’s ability to assess implicit self-explanation and thus sufficient exploration. A more
accurate student model allows the environment to better provide adaptive support to improve these behaviors and consequent student learning.

The main contribution of this thesis is a formal evaluation demonstrating that the student model which includes eye-tracking data provides a more accurate assessment than a model using only time as evidence of implicit self-explanation. The evaluation also confirms that the inclusion of self-explanation in the model leads to a more accurate assessment of sufficient exploration.

8.3 Limitations

First, there is the obvious limitation of the need for the eye-tracker itself. While eye-trackers are decreasing in cost, they are still prohibitively expensive, especially for individual use. While monitor-mounted eye-trackers (e.g. IView X Red from SensoMotoric Instruments, USA) lack the intrusiveness of their head-mounted counterparts, it can still be difficult to achieve and maintain calibration. Many factors, such as the presence of eye-makeup and the color and thickness of the user's eyelashes, can complicate the calibration process by making it more difficult for the eye-tracker to isolate the user's pupil. Later, during interaction with the system, the user must continue to face the monitor or risk a loss of calibration and a resulting break in the data stream.

Second, as mentioned above, any intelligent system which uses gaze shifts as evidence of implicit self-explanation must have interfaces with distinct, well-defined regions where actions occur and effects appear. For example, suppose the Plot Unit were changed so that when the user drops the curve, the appropriate equation appears in the plane with the curve. Then it would no longer be possible to define disjoint regions between which gaze shifts could occur.

Third, recall from Chapter 4 that the time threshold of 16 seconds was established using data collected from user study participants who expressed their self-explanation behavior or lack thereof by vocalizing their thoughts as they explored. Another problem is how to transfer this threshold to one more appropriate for real usage, when users do not need to speak to self-explain.

8.4 Future Work

While eye-tracking has been shown to enhance the ability of the ACE student model to assess implicit self-explanation and sufficient exploration, it remains to determine whether this leads to an improvement in user performance. The next step of this research is to collect this empirical...
evidence. A variety of interface tools have been designed and implemented which allow ACE to provide different levels of prompting for both exploration and self-explanation by relying on the assessment of the student model described in this thesis. A user study to test the effectiveness of these adaptive tools is currently being designed.

A long term research step is to consider the inclusion in the model of other meta-cognitive skills involved in effective exploration, such as the ability to form hypotheses and to monitor one’s own progress in the learning task.

While all work in this thesis pertains to the Plot Unit in ACE, the idea of using gaze shifts as evidence of implicit self-explanation could be applied to any system in which the interface is divided into distinct regions for actions and outcomes. It would simply be a matter of deciding, for each type of action, where the action would occur and then where the user would have to look to observe its effects. Other approaches include the use of unsupervised learning to detect relevant gaze patterns.
Bibliography


Appendix A

Pretest

Part 1: Questionnaire

A package gets dropped off at your doorstep, containing a brand new scooter. The only problem is that assembly is required. You
a) read the manual carefully, and then commence assembly
b) skim the manual quickly, and start to put the thing together
c) ask an expert to help you do it
d) trash the manual and start right away – you will figure it out as you go

When working with others, you
a) are very vocal about your ideas - you often have a good idea of how to do things
b) contribute some ideas, but don’t like to be the main person in charge
c) are mainly quiet – you prefer for others to take charge

When learning how to use a new computer application for an assignment, you
a) only figure out use those things that are totally necessary to get the assignment done
b) figure out how to use it to get the assignment done, plus learn a few extra things that caught your attention
c) fully explore the application – you are very curious about the various things you may discover in there

When it comes time to decide on what and how to do a school project, you
a) prefer to have the teacher tell you
b) prefer to make one up yourself
c) a mix of a and b: the teacher initially helps, but you have the final say

When learning something, you like to
a) get an overall picture of the thing you are learning, and then figure out the details
b) figure out the details first, and then get an overall picture
c) a mix of a and b
d) other

You are in a strange new city. You
a) get a map right away and use it frequently to plan your routes
b) get a map but use it only when absolutely necessary
c) don’t bother with a map –if lost, you’ll ask someone

You play computer games
a) almost every day
b) once or twice a week
c) less often that once a week

You use the Web to find information that you need (like a phone number, shopping, etc)
 a) almost every day
 b) once or twice a week
 c) less often that once a week
On a scale of 1-5, where 1 means very much, you like to surf the web
1) Very much  2) so-so  3) Neutral  4) not very much  5) not at all

On a scale of 1-5, where 1 means strongly like, your feelings about math are at:
1) Strongly like  2) like  3) Neutral  4) dislike  5) Strongly dislike

Part 2: Recognizing the output of a function.

For the following function equations specify the output of the function with the given inputs. Please show you work in the spaces provided.

1)  \( f(x) = 3 \)

What is the output of \( f(0) \) ?

a) 1  b) 0  c) 3  d) -3

What is the output of \( f(4) \) ?

a) 4  b) 3  c) -4  d) 0

2)  \( f(x) = 3x + 2 \)

What is the output of \( f(4) \) ?

a) 4  b) 5  c) 2  d) 14

What is the output of \( f(-4) \) ?

a) -4  b) 5  c) -10  d) 2

3)  \( f(x) = 5x^2 \)

What is the output of \( f(-1) \) ?

a) -1  b) -5  c) 5  d) 1

What is the output of \( f(1) ? \)

a) -1  b) -5  c) 5  d) 1

4)  \( f(x) = 3x^3 + 2x + 1 \)

What is the output of \( f(0) \)?

a) 0  b) 1  c) -1  d) 6
What is the output of \( f(2) \)?

a) 2  b) 29  c) 23  d) 6

**Part 3: Graph Properties**

For each question, circle the appropriate response.

1) The y-intercept of the graph in this picture is: Positive/Negative/Zero

![Graph 1](image)

2) The y-intercept of the graph in this picture is: Positive/Negative/Zero

![Graph 2](image)

3) The slope of the graph in this picture is: Positive/Negative/Zero

![Graph 3](image)
4) The slope of the graph in this picture is: Positive/Negative/Zero

5) The slope of the graph in this picture is: Positive/Negative/Zero

6) The exponent in the function equation graph in this picture is: Even/Odd

7) The exponent of the function equation graph in this picture is: Even/Odd
8) The graph has been scaled by: a Positive Number / a Negative Number / Zero

9) Which graph has been scaled by a larger number?: Graph A / Graph B

10) This graph has been shifted by a: positive number / negative number

11) This graph has been shifted by a: positive number / negative number
Part 4: Graph/Equation Properties

The function $f(x) = 2x + 4$ may be best described by the graph:

a)  

b)  

c)  

d) None of these graphs

The function $f(x) = x - 7$ may be best described by the graph:

a)  

b)  

c)  

d) None of these graphs

The function $f(x) = 3$ may be best described by the graph:

a)  

b)  

c)  

d) None of these graphs
The function \( f(x) = |x| \) may be best described by the graph:

a) ![Graph A](image1)

b) ![Graph B](image2)

c) ![Graph C](image3)

d) ![Graph D](image4) None of these graphs

The function \( f(x) = -2x + 5 \) may be best described by the graph:

a) ![Graph A](image5)

b) ![Graph B](image6)

c) ![Graph C](image7)

d) ![Graph D](image8) None of these graphs

The function \( f(x) = 4x^2 - 5 \) may best described by the graph:

a) ![Graph A](image9)

b) ![Graph B](image10)

c) ![Graph C](image11)

d) ![Graph D](image12) None of these graphs

The function \( f(x) = 1x^3 \) may be best described by the graph:

a) ![Graph A](image13)

b) ![Graph B](image14)

c) ![Graph C](image15)

d) ![Graph D](image16) None of these graphs
Part 5: Equation Properties

The function $f(x) = 3$ has
a) a positive slope  
   b) a negative slope  
   c) a zero slope

The function $f(x) = 3x + 7$ has
a) a positive slope  
   b) a negative slope  
   c) a zero slope

The function $f(x) = -1x + 10$ has
a) a positive slope  
   b) a negative slope  
   c) a zero slope

The function $f(x) = -2x + 5$ has
a) a positive y-intercept  
   b) a negative y-intercept  
   c) a zero y-intercept

The function $f(x) = 3x - 6$ has
a) a positive y-intercept  
   b) a negative y-intercept  
   c) a zero y-intercept

The function $f(x) = 5$ has
a) a positive y-intercept  
   b) a negative y-intercept  
   c) a zero y-intercept

The function $f(x) = -7$ has
a) a positive y-intercept  
   b) a negative y-intercept  
   c) a zero y-intercept
Appendix B

Post-test

Part 1: Questionnaire

Comments can be added in the spaces provided.

I found the information in the hints helpful (please circle one)

a) Strongly agree   b) Agree   c) Neutral   d) Disagree   e) Strongly disagree

I understood why the computer suggested I stay in an exercise (please circle one)

a) Strongly agree   b) Agree   c) Neutral   d) Disagree   e) Strongly disagree

I found it helpful when the computer told me to explore an exercise more than I had (please circle one)

a) Strongly agree   b) Agree   c) Neutral   d) Disagree   e) Strongly disagree

I found it helpful when the computer prompted me to answer multiple-choice questions (please circle one)

a) Strongly agree   b) Agree   c) Neutral   d) Disagree   e) Strongly disagree

I found it annoying when the computer prompted me to answer multiple-choice questions (please circle one)

a) Strongly agree   b) Agree   c) Neutral   d) Disagree   e) Strongly disagree

The amount of guidance provided by the computer was (please circle one)

a) Not enough   b) OK, but I would have liked more   c) Just Right   d) A bit too much   e) Way too much
Appendix B. Post-test

The computer gave me enough freedom to move around between activities (please circle one)
a) Strongly agree  b) Agree  c) Neutral  d) Disagree  e) Strongly disagree

I found the information in the help pages useful (please circle one)
a) Strongly agree  b) Agree  c) Neutral  d) Disagree  e) Strongly disagree

I prefer to learn this way over using a text book (please circle one)
a) Strongly agree  b) Agree  c) Neutral  d) Disagree  e) Strongly disagree

My favorite unit was: Machine (unit 1)  Switchboard (unit 2)  Plot (unit 3)

I feel that I learned a lot using ACE (please circle one)
a) Strongly agree  b) Agree  c) Neutral  d) Disagree  e) Strongly disagree

I feel that wearing the eye tracker made the experience less enjoyable. (please circle one)
a) Strongly agree  b) Agree  c) Neutral  d) Disagree  e) Strongly disagree

I feel that wearing the eye tracker detracted from my ability to learn. (please circle one)
a) Strongly agree  b) Agree  c) Neutral  d) Disagree  e) Strongly disagree

The thing I liked the most about ACE:

The thing I liked the least about ACE:

What I would like to see added to ACE:
Part 2: Recognizing the output of a function.

For the following function equations specify the output of the function with the given inputs. Please show you work in the spaces provided.

1) \( f(x) = 8 \)

What is the output of \( f(0) \) ?

a) 5   b) 0   c) 8   d) -8

What is the output of \( f(2) \) ?

a) 2   b) 8   c) -2   d) 0

2) \( f(x) = 3x + 3 \)

What is the output of \( f(5) \) ?

a) 18   b) 5   c) 6   d) 14

What is the output of \( f(-5) \) ?

a) -5   b) 3   c) -12   d) -15

3) \( f(x) = 2x^2 \)

What is the output of \( f(-2) \) ?

a) -1   b) -4   c) 8   d) -8
What is the output of \( f(2) \)?

a) 2  
b) 4  
c) 1  
d) 8

4) \( f(x) = 2x^3 + 3x - 1 \)

What is the output of \( f(0) \)?

a) 0  
b) 1  
c) -1  
d) 4

What is the output of \( f(1) \)?

a) 1  
b) 4  
c) 5  
d) 6

**Part 3: Graph Properties**

For each question, circle the appropriate response.

1) The **y-intercept** of the graph in this picture is: Positive / Negative / Zero

![Graph 1](image)

2) The **y-intercept** of the graph in this picture is: Positive / Negative / Zero

![Graph 2](image)
3) The **slope** of the graph in this picture is: Positive / Negative / Zero

4) The **slope** of the graph in this picture is: Positive / Negative / Zero

5) The **slope** of the graph in this picture is: Positive / Negative / Zero

6) The **exponent** in the function equation graph is this picture is: Even / Odd
7) The exponent in the function equation graph in his picture is: Even / Odd

8) The graph has been scaled by: A Positive Number / A Negative Number / Zero

9) Which graph has been scaled by a larger number? Graph A / Graph B

10) This graph has been shifted by a: Positive Number / Negative Number
11) This graph has been shifted by a: Positive Number / Negative Number

\[ f(x) = 2x - 6 \]

\[ f(x) = x + 7 \]

\[ f(x) = 5 \]

**Part 4: Graph / Equation Properties**

Which graph does the function \( f(x) = 2x - 6 \) best describe?

- a)
- b)
- c)
- d) None of these graphs

Which graph does the function \( f(x) = x + 7 \) best describe?

- a)
- b)
- c)
- d) None of these graphs

Which graph does the function \( f(x) = 5 \) best describe?

- a)
- b)
- c)
- d) None of these graphs
Which graph does the function $f(x) = 5x$ best describe?

a) b) c) d) None of these graphs

Which graph does the function $f(x) = -2x + 6$ best describe?

a) b) c) d) None of these graphs

Which graph does the function $f(x) = 3x^2 + 5$ best describe?

a) b) c) d) None of these graphs

Which graph does the function $f(x) = -x^3$ best describe?

a) b) c) d) None of these graphs
Part 5: Equation Properties

The function $f(x) = -4$ has
a) a positive slope  b) a negative slope  c) a zero slope

The function $f(x) = -5x + 2$ has
a) a positive slope  b) a negative slope  c) a zero slope

The function $f(x) = 4x + 6$ has
a) a positive slope  b) a negative slope  c) a zero slope

The function $f(x) = -4x + 5$ has
a) a positive y-intercept  b) a negative y-intercept  c) a zero y-intercept

The function $f(x) = 3x - 4$ has
a) a positive y-intercept  b) a negative y-intercept  c) a zero y-intercept

The function $f(x) = -7$ has
a) a positive y-intercept  b) a negative y-intercept  c) a zero y-intercept

The function $f(x) = 3$ has
a) a positive y-intercept  b) a negative y-intercept  c) a zero y-intercept