

## ABSTRACT

SABOURIN, JENNIFER LYNNE. Stealth Assessment of Self-Regulated Learning in Game-Based Learning Environments. (Under the direction of Dr. James C. Lester.)

Game-based learning environments have been touted as a powerful mechanism for keeping twenty-first century students motivated and engaged in learning tasks. These environments capitalize on important game features such as autonomy, control, and interactivity to maintain student interest. However, they may introduce new challenges by requiring students to engage in more advanced self-regulatory skills such as goal-setting and monitoring, complex problem solving, and critical thinking. Students without these skills are unlikely to realize the benefits of game-based learning and may struggle in systems where they are provided with too much autonomy. Consequently, real-time identification of self-regulated learning (SRL) is a problem of critical importance for providing adaptive scaffolding targeted to an individual student's needs.

The goal of the research presented in this dissertation is to explore the use of a theoretically grounded machine learning framework for the real-time stealth assessment of SRL skills. In this framework, prominent psychological theories are used to guide the selection and application of machine learning and data mining techniques. Empirical models are trained on a corpus of data from students interacting with a game-based learning environment, CRYSTAL ISLAND. These models are evaluated to determine if the predictive models are benefited by the theoretically grounded framework and if they are also sufficiently robust to offer real-time trace-based assessment of SRL.

The work consists of four phases of theory and data-driven investigation: (1) identification of SRL, (2) feature selection and creation, (3) model creation and (4) model evaluation. First, SRL skill measures of particular relevance to the CRYSTAL ISLAND environment are identified using a variety of techniques. Each measure is designed to be representative of one of the three components of SRL: cognition, metacognition and motivation. Students are classified based on their demonstration of these skills. Next, feature selection is used to identify simple personal attribute and in-game actions that are tied to SRL classifications. A differential sequence mining approach is then used to identify patterns of behavior

that are indicative of each component of SRL. These patterns are used to create complex event features guided by SRL theory. In the model creation phase, these features are integrated in naïve, static and dynamic Bayesian networks with structures that are increasingly informed by process models of SRL. The models are then empirically learned from a corpus of student interaction data. Finally, the models are evaluated to determine the value of the theoretically grounded empirical framework. Specifically, the evaluation seeks to quantify usefulness of the features identified in Phase 2 and the structures devised in Phase 3. Furthermore, the models are evaluated to identify how early into the interaction SRL skill can be accurately assessed and how well the models generalize to unseen populations.

Overall, the results indicate that the theoretically-grounded machine learning framework is effective for developing real-time trace based assessment models of SRL. The differential sequence mining approach yields patterns of behavior that are both meaningful and improve the predictive accuracy of the models. Dynamic Bayesian models with temporal structures representing the cyclic nature of SRL outperformed those without this representation. Finally, the framework was effective at predicting the cognitive, metacognitive and motivational components of SRL within the first 20% of the interaction. These results highlight the potential of the framework for guiding the design of real-time trace-based SRL models.

© Copyright 2013 by Jennifer Lynne Sabourin

All Rights Reserved

Stealth Assessment of Self-Regulated Learning in Game-Based Learning Environments

by  
Jennifer Lynne Sabourin

A dissertation submitted to the Graduate Faculty of  
North Carolina State University  
in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2013

APPROVED BY:

---

Dr. James C. Lester  
Committee Chair

---

Dr. Kristy E. Boyer

---

Dr. Tiffany M. Barnes

---

Dr. John L. Nietfeld

## **DEDICATION**

To my husband Philippe,  
who challenges me to be my best,  
supports me when I am at my worst,  
and whose love carries me through it all.

## BIOGRAPHY

Jennifer Lynne Robison was born in Alexandria, Virginia on September 5, 1986 but has spent most of her life living in Cary, North Carolina. She attended William G. Enloe High School in Raleigh, NC. Here, she took her first Computer Science class under the insistence of her parents. Though hesitant, she grew to love it and when she graduated from Enloe in 2004 she chose to pursue Computer Science as her major at North Carolina State University. During her time at N.C. State, Jennifer also minored in Cognitive Science and Psychology. She also interned at SAS Institute starting in 2005. She participated in the University Scholars Program, the Computer Science Honors Program, and the STARS Student Leadership Corps. Through these programs she had the opportunity to take graduate level classes and pursue undergraduate research projects under the direction of James Lester and her mentor Scott McQuiggan. These experiences left her confident in her decision to pursue a Ph.D. at N.C. State.

She graduated from her undergraduate program as Valedictorian in 2008 and immediately began the graduate program the following semester. She was awarded the NSF Graduate Research Fellowship, allowing her great flexibility in guiding her graduate career. She remained active with the STARS SLC and helped devise outreach programs and other efforts aimed at broadening participation in computing. She also engaged in many activities focused around improving her teaching skills including the Certificate for Accomplishment in Teaching program, a Mentored Teaching Assistantship and the Preparing the Professoriate program. She continued to work as an intern at SAS, now in the Education Practice, working on innovative educational technology. She accepted a full-time position with SAS in June 2013 as a Research Scientist and Software Developer.

Jennifer is lucky to work at SAS with amazing colleagues, many of whom were part of the same graduate program at NC State. Jennifer also works with her husband Philippe Sabourin, who she met in Computer Science class her freshman year and later married in June 2010.

## ACKNOWLEDGMENTS

I would first like to thank my family who has made everything possible. My parents Connie and Russ Robison taught me that I was capable of anything I put my mind to and sometimes even more. My grandparents were always a strong source of support, especially my grandfather Dr. O.W. Robison who was proud to watch his oldest grandchild follow in his academic footsteps. My brothers Chris and Ben have always teased me for being a nerd, and I have yet to prove them wrong. I also have the privilege of having married into a fun, crazy, and large family who have always welcomed me with open arms and encouraged me throughout my graduate career.

I would like to thank my husband, Philippe, who has been with me throughout this whole journey. He has been unfailing in his ability to help me power through when the road was tough and to celebrate the successes big and small. I want to thank all of my friends who have supported me as well, especially Meisha Gourley. She has never doubted me in anything and has given me more strength than she knows.

I would also like to thank everyone who has guided me academically. First and foremost, my advisor Dr. James Lester. His curiosity, enthusiasm and professional character have inspired me greatly throughout this journey. I thank him for all the trust he put in me in choosing my path and all the countless hours he spent reviewing my work and guiding my research. I would like to thank Dr. Scott McQuiggan, who mentored me when I first began my career and continues to mentor me to this day. He provided me with an incredible model of scholarly achievement without which I would not be where I am today. I will always value his friendship, guidance and enthusiastic conversations about all sorts of topics. I would also like to thank Mr. Michael Downey, my first computer science teacher. He ignited my passion for Computer Science and taught me to always fight for myself and never be intimidated by those who didn't believe in me.

I would like to thank everyone I have had the pleasure of working with in my time at N.C. State. My committee members, Dr. Kristy Boyer, Dr. Tiffany Barnes, and Dr. John

Nietfeld have given me such excellent guidance academically, professional and personally. I am also thankful for the guidance of Jonathan Rowe, who has always selflessly made time to talk through interesting research ideas or discuss professional or personal development, even though he has so little time to give. I am thankful to Lucy Shores for being a great companion throughout this journey and helping to keep me on track to the very end.

I am also thankful for all of the others I have had the opportunity to work with throughout the years. The members of Intellimedia, including Alok Baikadi, Veronica Catete, Stephen Cossa, Kirby Culbertson, Julius Goth, Joe Grafsgaard, Eunyoung Ha, Sarah Hegler, Karoon McDowell, Chris Mitchell, Bradford Mott, Sam Leeman-Munk, Eleni Lobene, Wookhee Min, Marc Russo, Rob Taylor, Andy Smith, and Donnie Wrights, have always been a great source of collaboration and camaraderie. The Faculty and Staff of the Department of Computer Science, including Barbara Adams, Carol Allen, Suzanne Balik, Sarah Heckman, Linda Honeycutt, Thomas Honeycutt, Purush Iyer, Margery Page, Susan Peaslee, Douglas Reeves, Robert St. Amant, Ken Tate, and David Thuente have always been there for assistance in navigation of the academic waters. I would also like to thank all of the amazing researchers outside of NCSU whose conversations and research have guided me, including Roger Azevedo, Ryan Baker, Gautam Biswas, Cristina Conati, Sidney D’Mello, Joao Dias, Jon Gratch, Jason Harley, John Kinnebrew, Val Shute, and Phil Winne.

The work reported in this dissertation was supported by the National Science Foundation, under grants REC-0632450, DRL-0822200, IIS-0812291, DRL-1114655, and CNS-0739216. This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

## TABLE OF CONTENTS

|  |      |
|--|------|
| LIST OF TABLES .....   | viii |
| LIST OF FIGURES .....  | ix   |
| CHAPTER 1 Introduction.....  | 1    |
| 1.1 Problem .....  | 2    |
| 1.2 Approach .....   | 3    |
| 1.3 Thesis Statement and Hypotheses .....  | 7    |
| 1.4 Contributions .....  | 8    |
| 1.5 Organization .....   | 9    |
| CHAPTER 2 Background and Related Work.....   | 11   |
| 2.1 Fostering Student Engagement.....  | 11   |
| 2.2 Self-Regulated Learning .....  | 16   |
| 2.3 Measuring SRL .....  | 21   |
| 2.4 Modeling and Supporting Student SRL Behaviors .....                            | 24   |
| 2.5 Stealth Assessment.....  | 26   |
| CHAPTER 3 Student Behavior Corpus.....   | 28   |
| 3.1 CRYSTAL ISLAND .....   | 28   |
| 3.2 Corpus Collection.....   | 29   |
| CHAPTER 4 Prior Findings on Student Engagement and Learning in CRYSTAL ISLAND .... | 33   |
| 4.1 Affect and Motivation .....  | 33   |
| 4.2 Engagement and Disengagement .....   | 36   |
| 4.3 Inquiry Behaviors.....   | 38   |
| CHAPTER 5 Identifying Self-Regulated Learning .....                                | 40   |
| 5.1 Cognitive/Behavioral Classifications .....                                     | 40   |
| 5.2 Metacognitive Classifications .....  | 47   |
| 5.3 Motivation Classification .....  | 52   |
| 5.4 Discussion .....   | 56   |
| CHAPTER 6 Feature Selection and Creation.....                                      | 59   |
| 6.1 Feature Selection .....  | 59   |
| 6.2 Feature Creation .....   | 64   |

|  |     |
|--|-----|
| 6.3 Discussion .....   | 74  |
| CHAPTER 7 Predictive Models of Self-Regulated Learning .....         | 76  |
| 7.1 Bayesian Approach .....  | 76  |
| 7.2 Results .....  | 85  |
| 7.3 Discussion .....   | 100 |
| CHAPTER 8 Generalization.....  | 101 |
| 8.1 Population.....  | 101 |
| 8.2 Predictive Models.....   | 105 |
| 8.3 Discussion .....   | 112 |
| CHAPTER 9 Conclusion .....   | 114 |
| 9.1 Hypotheses Revisited.....  | 115 |
| 9.2 Summary .....  | 118 |
| 9.3 Limitations .....  | 120 |
| 9.4 Future Directions.....   | 121 |
| 9.5 Concluding Remarks .....   | 122 |
| REFERENCES .....   | 124 |
| APPENDICES .....   | 134 |
| Appendix A – SRL Tagging Protocol.....                               | 135 |
| Appendix B – Content Test for CRYSTAL ISLAND Study.....              | 137 |
| Appendix C – Pre-Experiment Materials for CRYSTAL ISLAND Study.....  | 141 |
| Appendix D – Post-Experiment Materials for CRYSTAL ISLAND Study..... | 147 |
| Appendix E – Handout Materials for CRYSTAL ISLAND Study.....         | 154 |

## LIST OF TABLES

|   |     |
|---|-----|
| Table 1. Frequency and proportion of emotion self-reports .....             | 34  |
| Table 2. Affective states correlated with motivation outcome measures ..... | 35  |
| Table 3. Class differences on cluster features .....                        | 46  |
| Table 4. Metacognition tagging scheme .....                                 | 48  |
| Table 5. IMI subscale differences by motivation class.....                  | 53  |
| Table 6. Classification details.....  | 57  |
| Table 7. Feature selection for cognition .....                              | 61  |
| Table 8. Feature selection for metacognition.....                           | 62  |
| Table 9. Feature selection for motivation .....                             | 63  |
| Table 10. Differential patterns for cognition .....                         | 68  |
| Table 11. Differential patterns for metacognition .....                     | 69  |
| Table 12. Differential patterns for motivation .....                        | 69  |
| Table 13. Identified patterns of behavior.....                              | 73  |
| Table 14. Personal features to SRL processes .....                          | 80  |
| Table 15. Occurrence features to SRL processes .....                        | 81  |
| Table 16. Created contingency event features to SRL processes .....         | 84  |
| Table 17. Predictive model performance for cognition .....                  | 89  |
| Table 18. Predictive model performance for metacognition .....              | 92  |
| Table 19. Predictive model performance for motivation .....                 | 95  |
| Table 20. Secondary corpus classification details .....                     | 102 |
| Table 21. Generalized model performance for cognition .....                 | 106 |
| Table 22. Generalized model performance for metacognition .....             | 108 |

## LIST OF FIGURES

|   |    |
|---|----|
| Figure 1. Overview of proposed framework.....   | 6  |
| Figure 2. Components of self-regulated learning .....   | 16 |
| Figure 3. Three phases and subprocesses of self-regulated learning.....   | 18 |
| Figure 4. Posttest score by self-regulated learning score using learner (LC) or program(PC) controlled software ..... | 21 |
| Figure 5. CRYSTAL ISLAND learning environment.....  | 28 |
| Figure 6. Self-report device .....  | 31 |
| Figure 7. Histogram of off-task behavior .....  | 36 |
| Figure 8. Learning differences by problem-solving cluster.....  | 45 |
| Figure 9. Testing behavior by problem-solving cluster.....  | 45 |
| Figure 10. Histogram of metacognitive score.....  | 49 |
| Figure 11. Learning differences by metacognitive class .....  | 50 |
| Figure 12. In-game behaviors by SRL classification.....   | 51 |
| Figure 13. Histogram of IMI Score .....   | 54 |
| Figure 14. Learning differences by motivation class.....  | 55 |
| Figure 15. Learning gains by classification .....   | 57 |
| Figure 16. Targeted behaviors .....   | 67 |
| Figure 17. Structure of naïve Bayesian network using <i>All Features and Feature Selection</i> .. .....               | 79 |
| Figure 18. Structure of naïve Bayesian network using <i>Feature Selection + Event Feature Creation</i> .....          | 79 |
| Figure 19. Structure of static Bayesian network using <i>Feature Selection</i> .....                                  | 82 |
| Figure 20. Structure of static Bayesian network using <i>Feature Selection + Event Feature Creation</i> .....         | 83 |
| Figure 21. Structure of dynamic Bayesian network using <i>Feature Selection</i> .....                                 | 86 |
| Figure 22. Structure of dynamic Bayesian network using <i>Feature Selection + Event Creation</i> .....                | 87 |
| Figure 23. Predictive accuracy of cognition over time.....  | 90 |
| Figure 24. Predictive accuracy of metacognition over time .....   | 93 |
| Figure 25. Predictive accuracy of motivation over time .....  | 96 |

|   |     |
|---|-----|
| Figure 26. Distributions of accuracy by feature set .....                         | 98  |
| Figure 27. Distributions of accuracy by model type .....                          | 99  |
| Figure 28. Interaction of feature set and model type .....                        | 99  |
| Figure 29. Secondary corpus learning differences by cognition .....               | 102 |
| Figure 30. Secondary corpus learning differences by metacognition.....            | 104 |
| Figure 31. Secondary corpus learning differences by motivation .....              | 104 |
| Figure 32. Generalized model predictive accuracy of cognition over time .....     | 107 |
| Figure 33. Generalized model predictive accuracy of metacognition over time ..... | 109 |
| Figure 34. Generalized model predictive accuracy of motivation over time.....     | 111 |

## CHAPTER 1

### Introduction

One-to-one tutoring has long been considered the gold standard of effective instruction by demonstrating significant improvements in student learning over the typical classroom setting (Bloom, 1984). This benefit is suspected to be due to high levels of interactivity as well as highly individualized attention and feedback. While it is not feasible for each student to have his or her own tutor, the intelligent tutoring systems community has been trying to bridge this gap by endowing computers with the same interactive and individualized tutoring capabilities (Vanlehn, 2006; Woolf, 2009). These efforts have led to significant improvements in educational software's effectiveness, and in some cases, intelligent tutoring systems have produced learning gains comparable to human tutors (Vanlehn, 2006). However, as with any educational tool, the software's ability to produce learning gains for students depends heavily on students' engagement and motivation with the system as well as their ability to use the software effectively.

One common approach for encouraging engagement involves increasing student autonomy and allowing each individual student to guide their own learning (Easterday, Alevan, Scheines, & Carver, 2011; Young, 1996). The insight behind this approach is that students will be able to focus on tasks and topics that fit within their own learning goals and interests (Pintrich, 2004). However, while this approach has gained popularity, there is increased evidence that not all students are successful at guiding their own learning (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Kirschner, Sweller, & Clark, 2006; Young, 1996). To be successful, students must be capable of setting meaningful learning objectives. They must then identify activities, behaviors, and strategies that may achieve these goals, monitor and evaluate their progress, and alter their behavior and strategies accordingly. Unfortunately, there is evidence that not all students are capable of guiding their own learning in this way (Ellis & Zimmerman, 2001) and may consequently experience limited success with systems that re-

quire these skills (Azevedo, Moos, Greene, Winters, & Cromley, 2008; Easterday et al., 2011; Young, 1996).

The ability to set learning goals, identify successful strategies and evaluate personal success is the hallmark of a self-regulated learner. Students who exhibit self-regulated learning (SRL) skills are able to drive their own learning and are often more successful in learning tasks and academic settings (Zimmerman, 1990). While SRL skills can be taught and often improve with practice (Kostons, van Gog, & Paas, 2012), students who have not yet developed appropriate SRL strategies are more likely to flounder in self-guided learning systems. However, there is evidence that with appropriate scaffolding these environments can be beneficial in improving learning and interest as well as aid in development of SRL skills (Alevén, McLaren, Roll, & Koedinger, 2006; Azevedo, Cromley, Winters, Moos, & Greene, 2005). While there is much debate on how to properly scaffold SRL (Davis, 2003; Fiorella & Mayer, 2012; Ifenthaler, 2012; Kauffman, 2004; White & Frederiksen, 1998), there is consensus that appropriate scaffolding involves a delicate balance of allowing autonomy but providing support when necessary (Koedinger & Alevén, 2007). To do this successfully, teachers and tutoring systems must be able to accurately identify a student's SRL skill level which may vary based on the specific learning context (Schraw, 2010).

## **1.1 Problem**

This dissertation aims to investigate the issue of real-time SRL assessment in an open-ended learning environment for middle grade microbiology, CRYSTAL ISLAND. This game-based learning environment encourages independent inquiry to solve a science mystery while learning related domain knowledge. Prior work has identified that, in general, students experience positive learning gains and engagement benefits from interactions with the CRYSTAL ISLAND game. However, further evidence suggests that students experience a broad range of success during their interactions, with some students demonstrating ineffectual behaviors and reduced learning outcomes that suggest they are floundering. It is hypothesized that these students lack the SRL skills necessary to identify and evaluate learning goals in the open-ended

game environment. Consequently, it is the goal of the proposed dissertation to identify students who are less able to regulate their own learning behaviors in CRYSTAL ISLAND.

However, assessment of SRL presents a substantial challenge. Questionnaires and other static measures may not reliably indicate a student's ability to regulate their learning in a specific context (Schraw, 2010) and do not capture real-time behaviors (Winne, 2010). Other online measures such as think-aloud protocols or prompting are intrusive and may alter a student's natural behaviors (Winne & Perry, 2000). Consequently, developing real-time unobtrusive assessments of self-regulated learning in computer-based environments has been recognized as an important and challenging objective (Greene & Azevedo, 2010; Schraw, 2010; Winne, 2010).

## 1.2 Approach

This dissertation focuses on the use of theory-driven machine learning and data mining techniques to investigate real-time SRL assessment in the CRYSTAL ISLAND game-based learning environment. In this framework, accepted psychological theories of self-regulated learning are used to select appropriate data mining techniques and develop useful empirically-learned computational models. Here we will describe the psychological theories used to guide the framework and then specifically how these theories informed the computational aspects of the work.

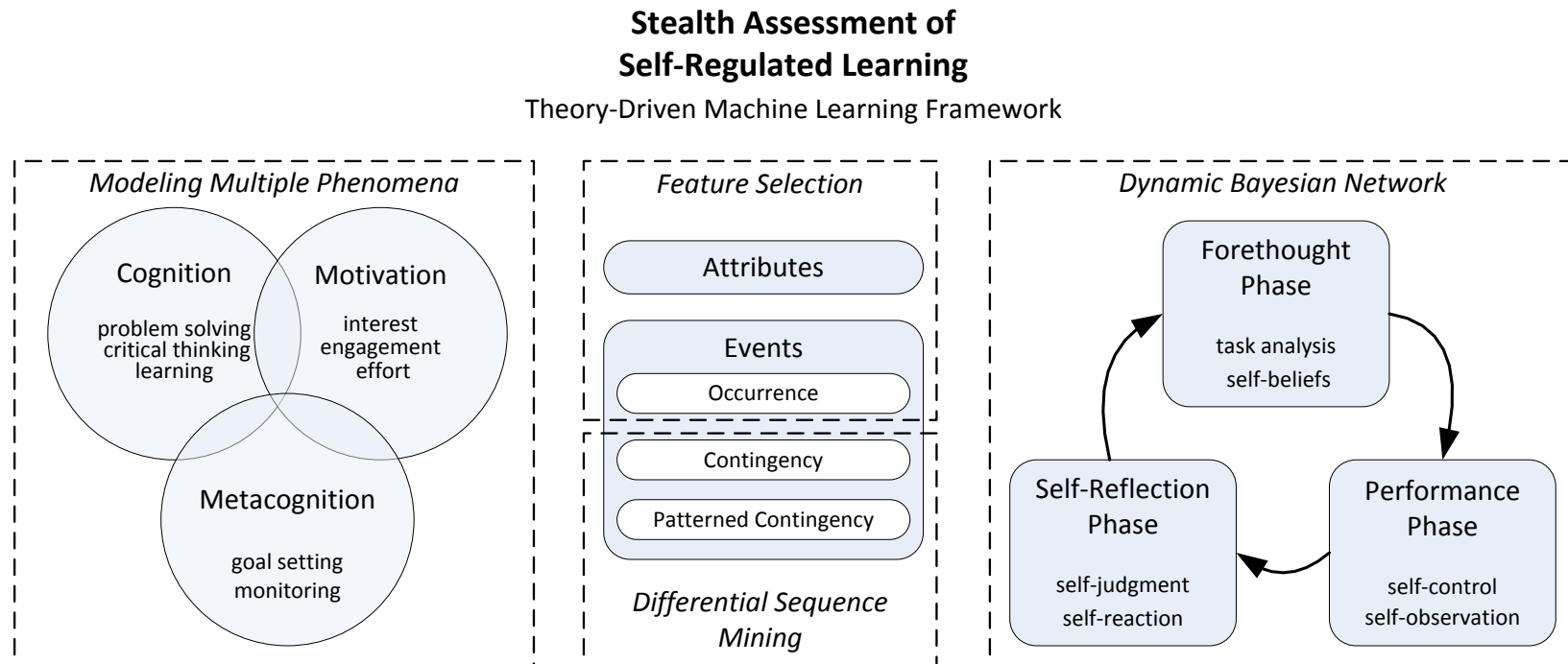
**Theoretical Grounding.** One issue with developing a theory-driven framework for computational representations of self-regulated learning is the abundance of existing cognitive and metacognitive models and theories. Presently there is no universally accepted theory of self-regulated learning so it is difficult to identify which models should be utilized. This work takes the approach of utilizing theories that (1) are commonly used in practice, (2) represent different components and phenomena of SRL, and (3) have straightforward ties to computational approaches. Specifically, three descriptive theories of SRL are integrated in the current framework:

- *Components of self-regulated learning:* According to Zimmerman, “although definitions of self-regulated learning involving specific processes often differ on the basis of researchers’ theoretical orientations, a common conceptualization of these students has emerged as metacognitively, motivationally, and behaviorally active participants in their own learning.” (Zimmerman, 1990, p. 4). This provides the distinction of the three primary components of self-regulated learning: cognition/behavior, metacognition, and motivation (Schraw, Crippen, & Hartley, 2006; Zimmerman, 1990). The cognitive/behavioral component involves students having the necessary strategies and skills to be effective learners. The metacognitive component focuses on how students recognize their own knowledge and monitor their learning. Finally, the motivation component describes self-regulated learners’ high self-efficacy, interest, and engagement with the learning material.
- *Cyclical nature of self-regulated learning:* Researchers commonly describe self-regulated learning as a cycle of constant reflection on and adaptation of behaviors. Though there are multiple cyclical models (DiBenedetto & Zimmerman, 2013; Winne & Hadwin, 1998; Zimmerman, 2000) most involve at least three phases including planning, performance, and reflection. These phases cycle constantly with planning impacting performance driving reflection, which in turn alters future planning.
- *Instantiation of self-regulated learning:* Finally, researchers tend to agree that self-regulated learning can be viewed as either a general aptitude of the learner or as the actual occurrences or events of SRL behaviors in a particular learning scenario (Greene & Azevedo, 2010; Winne & Perry, 2000). Winne (2010) further identifies three levels of events that should factor into SRL assessment: *occurrence*, *contingency* and *patterned contingency*. Here, *occurrence* refers to how often a student took a particular action, such as taking a note. *Contingency* describes both an action and its predecessor. For example, the student took a note after reading about a disease. *Patterned contingency* represents the pattern of such contingent behaviors. In general, it

is accepted that an accurate representation of SRL will need to incorporate both attribute and event information (Schraw, 2010; Winne, 2010).

**Theory-Driven Machine Learning Framework.** These three theoretical components are used to guide the machine learning and data mining approaches for modeling self-regulated learning (Figure 1). In each case the theory guides the structure or selection of computational techniques. However, specific measures, patterns, and relationships are empirically learned using the machine learning and data mining techniques. Theory and machine learning are integrated in the following ways:

- *Feature selection:* Forward-selection logistic regression is used to identify the attribute and occurrence event level features most predictive of self-regulated learning in CRYSTAL ISLAND. Attribute features are taken from standard validated questionnaires given prior to students' interactions with the environment. Occurrence level features are taken as the frequency of specific actions within the environment.
- *Differential sequence mining:* A differential sequence mining approach is used to identify more robust patterns of actions that are demonstrated by students with more advanced self-regulated learning skills. These patterns are then used to create features representing contingency and patterned contingency events.
- *Dynamic Bayesian networks:* Dynamic Bayesian networks are used to represent the cyclical and dynamic nature of self-regulated learning. Specifically, the model seeks to capture the planning, performance, and reflection phases of SRL. The dynamic model further represents the cyclical nature across time.
- *Modeling multiple phenomena:* Finally, the approaches described above are applied to several phenomena representative of the cognitive, metacognitive, and motivational components of SRL. Each phenomenon is modeled independently as students may have strengths in one component and not in another (Schraw et al., 2006; Zimmerman, 2000).



**Figure 1.** Overview of proposed framework

### 1.3 Thesis Statement and Hypotheses

This theory-driven machine learning framework incorporates commonly accepted psychological theories of self-regulated learning and utilizes them to guide empirical machine learning and data mining approaches for stealth assessment of self-regulated learning in a game-based learning environment. Along these lines, this dissertation investigates the following thesis:

*The proposed theory-driven framework can guide the investigation of how machine learning and data mining techniques can be leveraged for stealth assessment of self-regulated learning.*

To effectively evaluate this thesis, the individual components of the framework should be explored. Specifically, we investigate how the novel applications of existing psychological theories and machine learning impact the real-time predictive accuracy of computational models of self-regulated learning. This evaluation centers on the two primary machine learning and data mining techniques utilized: differential sequence mining and dynamic Bayesian networks. The following hypotheses explore the usefulness of these techniques for real-time SRL assessment:

- **Hypothesis 1:** Differential sequence mining can identify patterns of student behavior that:
  - (a) are indicative of self-regulated learning.
  - (b) improve the predictive accuracy of real-time predictive models of self-regulated learning.
- **Hypothesis 2:** Bayesian networks designed with a structure to encode the key processes of self-regulated learning will yield increased predictive power. Specifically,

- (a) Static Bayesian networks which encode key SRL processes will achieve greater predictive accuracy than naïve Bayesian networks with no such representation.
- (b) Dynamic Bayesian networks which encode the cyclic relationship between key processes across time will achieve greater predictive accuracy than static Bayesian networks with no such temporal representation.
- **Hypothesis 3:** The proposed theory-driven machine learning techniques can be used to assess the
  - (a) cognitive/behavioral
  - (b) metacognitive
  - (c) motivational
 aspects of self-regulated learning in real-time.

With the investigation of these hypotheses we hope to show how psychological theory can effectively guide the development of empirically-learned models by improving predictive accuracy and interpretability.

## 1.4 Contributions

The work reported in this dissertation makes the following contributions:

- An empirical account of the relationship between learning and engagement in an open-ended inquiry- and game-based learning environment. (Sabourin, Rowe, Mott, & Lester, 2013; Sabourin, Rowe, Mott, & Lester, 2012a, 2012b; Sabourin & Lester, in press).
- A clustering approach for identifying behavioral indicators of problem-solving and strategy use in CRYSTAL ISLAND that are meaningful and predictive of student learning outcomes (Sabourin & Lester, in press).

- An annotation protocol for analyzing student goal setting and monitoring behavior that is capable of classifying students' metacognitive self-regulatory skills (Sabourin, Shores, Mott, & Lester, 2012; Sabourin, Shores, Mott, & Lester, in press)
- An empirical account of the role of self-regulated learning as it relates to engagement and learning outcomes in an inquiry- and game-based learning environment (Sabourin, Shores, et al., 2012; Sabourin et al., in press).
- An empirical investigation of the patterns of behavior and strategies associated with cognitive, metacognitive and motivational components of self-regulated learning in CRYSTAL ISLAND (Sabourin, Mott, & Lester, 2013).
- A theoretical and empirical framework for real-time unobtrusive assessment of self-regulated learning in computer-based learning environments.
- A suite of predictive models of self-regulated learning incorporating both empirical and theoretical knowledge to identify self-regulatory skills at real-time during students' interaction (Sabourin, Shores, et al., 2012; Sabourin, Mott, & Lester, 2012, 2013).

## 1.5 Organization

The remainder of this dissertation is organized as follows. Chapter 2 provides a discussion of related work in the areas of intelligent tutoring systems, self-regulated learning, and approaches to measuring and modeling SRL. Chapter 3 presents CRYSTAL ISLAND, the learning environment at the focus of this work as well as a corpus collection that provides the data for the empirical studies described throughout this document. Chapter 4 discusses early findings related to learning and engagement in CRYSTAL ISLAND. Chapter 5 presents the three components of self-regulated learning being explored in CRYSTAL ISLAND and how students were classified into SRL skill groups. Chapter 6 discusses the feature selection and feature creation techniques that were used to create the feature sets for the predictive modeling. Chapter 7 describes the Bayesian approach to empirically learned predictive models and presents a de-

tailed evaluation of the models. Chapter 8 explores how well the learned models generalize to an unseen population as a measure of robustness. Finally, Chapter 9 revisits the hypotheses in relation to the presented findings and offers some discussion and concluding remarks.

## CHAPTER 2

### Background and Related Work

This dissertation focuses on examining issues of modeling and assessing self-regulated learning in an intelligent game-based learning environment. This section begins with a brief overview of efforts to promote student engagement and motivation within intelligent tutoring systems, focusing on two specific approaches to foster student interest, game-based and inquiry-based learning. The following section will describe the role of self-regulated learning in these systems as well as in general academic settings, and will motivate the need to understand and support these processes during learning activities. It will discuss modern approaches to measuring SRL and recent concerns about the limitations of many of these approaches. These concerns highlight the need for real-time unobtrusive assessment of SRL that is afforded by computer-based learning environments. The next section discusses recent advances in this area and describes research on modeling and identifying SRL processes in intelligent tutoring systems. The final section describes the inspiration behind stealth assessment and its application in other game-based learning environments.

#### 2.1 Fostering Student Engagement

The focus on encouraging student engagement and motivation has been growing rapidly in recent decades in both computer and classroom-based instruction. This attention is driven by empirical findings that students' feelings of interest and motivation towards an activity, domain, or learning in general has a powerful influence on how long they will persist with a task and how willing they are to initiate an activity (Baker, D'Mello, Rodrigo, & Graesser, 2010; Kanfer & Ackerman, 1989; Pekrun, Goetz, Titz, & Perry, 2002; Picard et al., 2004).

**Game-based Learning.** Game-based learning has been proposed as an approach to encourage positive affect, engagement and motivation in learning activities by utilizing game-like

features and environments (Gee, 2003; Kapp, 2012; Shaffer, 2006). This work draws on empirical evidence that games are highly motivating and have natural ties with how people learn (Gee, 2003; Kapp, 2012; McNamara, Jackson, & Graesser, 2009). In recent years games have been devised for a broad range of skills and domains including scientific inquiry (Clarke & Dede, 2009; Rowe, Shores, Mott, & Lester, 2011), mathematics principles (Conati, 2002), negotiation skills (Kim, Hill, Durlach, Lane, Forbell, Core, Marsella et al., 2009), foreign languages (Hallinen, Walker, Wylie, Ogan, & Jones, 2009; Johnson, 2010), policy argumentation (Easterday et al., 2011) and critical reasoning (Millis et al., 2011).

In addition to incorporating engaging game features, many of these systems also intelligently model student behavior. For example, *Prime Climb*, a game developed to teach factorization includes models of intelligent hinting behavior as well as student's affective states (Conati, 2002). *Project ARIES* integrates natural dialog intelligent tutoring technologies in a game where students must learn critical reasoning skills in order to prevent an alien invasion (Millis et al., 2011). *FearNot!* is a game aimed at teaching young children strategies to handle bullying and involves robust affective models that drive the characters in the game and adapts the game scenario (Paiva, Dias, Sobral, Aylett, Woods, Hall, & Zoll, 2005). The *BiLAT* system involves rich interactions with in-game characters and is primarily used in teaching cross-cultural negotiation skills to military personnel (Kim, Hill, Durlach, Lane, Forbell, Core, Marsella et al., 2009). This game incorporates a variety of intelligent models that guide virtual character behavior as well as delivery of hints and explanations.

Overall, game-based learning systems have been shown to be effective in increasing student knowledge and skills, as well as fostering engagement and positive affect. A study by Hallinen *et al.* indicated that incorporating game-like features into a learning system resulted in equivalent quality of learning but resulted in significant increases in engagement (Hallinen et al., 2009). Meanwhile, other work has shown that measures of engagement are strongly correlated with learning in game-based systems, as in traditional tutoring systems (Rowe, Shores, Mott, & Lester, 2010).

However, game-based learning has recently fallen under scrutiny for having features that are superfluous to the learning task (Mayer, & Johnson, 2010). A major concern is that the game-play aspects that are designed to encourage interest and motivation may also introduce many distractions or “seductive details” that draw student attention away from the learning tasks (Harp & Mayer, 1998; Sabourin, Rowe, Mott, & Lester, 2011). For example, students may become distracted by the characters and objects that are present in the world, or may spend time playing with aspects of the physics engine that underlies the gameplay. There is evidence that while these features may offer positive benefits for engagement, not all students are capable of regulating their behavior in a way that takes advantage of these features without harming their learning outcomes (Sabourin, Rowe, et al., 2011).

Another concern of game-based learning environments is that students interacting with these systems may receive too little direct guidance (Easterday et al., 2011). These environments often focus on allowing greater student control in order to increase engagement and interest, but, depending on the design of the learning environment, this may be to the detriment of learning. According to Fiorella and Mayer, "One of the challenges associated with the design of educational games is that the features intended to motivate students may result in extraneous processing" (Fiorella & Mayer, 2012, p. 1076) Students interacting with these systems are often expected to identify and select their own learning tasks, requiring additional resources and skills which students may not yet possess.

**Inquiry Based Learning.** Another type of learning system that uses autonomy and student control to promote engagement and motivation is inquiry-based learning. Inquiry-based learning has been a focus of recent attention in both traditional classrooms (Alfieri et al., 2011; Kirschner et al., 2006) and intelligent tutoring systems (Ketelhut, 2007; Roll, Alevan, & Koedinger, 2010; Veermans, van Joolingen, & de Jong, 2000; Woolf et al., 2005). It has achieved this attention in large part because of its use of authentic problem solving scenarios and the fact that students are put in control of their own learning. As part of inquiry-based learning activities, students are expected to play an active part in “making observations, for-

mulating hypotheses, gathering and analyzing data, and forming conclusions from that data” (Ketelhut, 2007, p. 100).

A variety of approaches to inquiry-based learning have been explored in the intelligent tutoring systems community. For example, Woolf *et al.* have developed the inquiry environment *Rashi*, which supports inquiry skills in a variety of different domains including biology and geology (Woolf *et al.*, 2005). Students are able to use tools such as the inquiry notebook and hypothesis editor to record their observations, reason about findings and support or reject hypotheses. In this system, students are not taught specific inquiry processes; instead, they are guided through inquiry by a “coach.” This approach allows students to experience inquiry without needing direct instruction.

*SimQuest* is another inquiry-based system designed for the domain of physics (Veermans *et al.*, 2000). In this system students manipulate simulations in order to conduct experiments and test hypotheses. Veermans *et al.* have examined mechanisms to support inquiry in *SimQuest*, such as encouraging students to reason about the correctness of hypotheses and providing details on how experiments can be used to validate them. Results indicate that students whose inquiry is prompted in this way interact with the environment in a more reflective way, by completing more experiments and spending more time on each problem. Though no learning differences were found based on the inquiry scaffolding, it was found that, for students who had the inquiry support, learning was correlated with the number of experiments they conducted, whereas for unscaffolded students, learning was correlated with how many assignments they completed. This presents an interesting dichotomy between the quality versus quantity of interactions.

Another system for promoting inquiry is *Invention Lab*, in which students are encouraged to “invent” equations that explain the relationships between variables (Roll *et al.*, 2010). They are presented with cases where one or more variables are modified and are encouraged to discover the equation that explains the differences between the cases. Early evaluation of the *Invention Lab* suggests that scaffolding both domain skills as well as inquiry skills is important for enabling students to arrive at high-quality solutions.

*River City* presents inquiry as a scientific mystery in which students are encouraged to gather information about patient symptoms and possible contaminants in an open-ended environment (Ketelhut, 2007). Analysis of student inquiry behaviors in *River City* has demonstrated that students tend to engage in more inquiry behaviors after repeated interactions with the environment. This is particularly true for female students and students with low self-efficacy. This last finding is especially important since low self-efficacy students initially demonstrate fewer inquiry behaviors at the time of the first interaction, but are completing equivalent numbers to the high self-efficacy students by the last interaction.

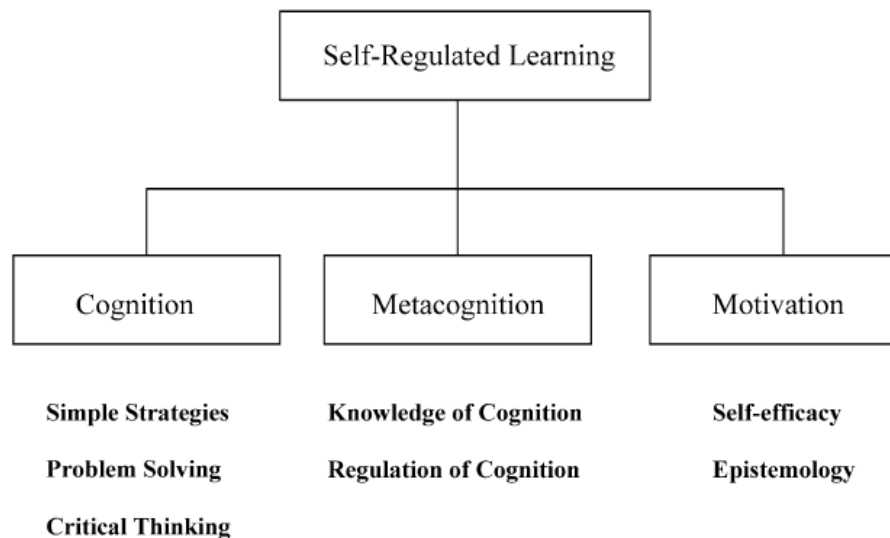
Many of these investigations of inquiry-based learning environments indicate that this approach may only be effective in particular circumstances. For instance, there is evidence that students need to have a reasonable level of background knowledge in order to learn new material in an inquiry-based setting (Alfieri et al., 2011; Kirschner et al., 2006). For students without sufficient prior knowledge, inquiry environments can require significant cognitive resources resulting in poor learning. Students should also be guided through the steps of inquiry to avoid floundering (Alfieri et al., 2011; Ketelhut, 2007; Kirschner et al., 2006). However, there is evidence that through initial guidance of appropriate inquiry behaviors students show improved inquiry skills in the future (Ketelhut, 2007). The finding that inquiry-based instruction can improve inquiry skills is important for motivating this as an effective method of teaching (Cuevas, Lee, Hart, & Deaktor, 2005).

Both inquiry and game-based learning utilize features of student control and autonomy to promote interest in the learning activity. However, there is evidence that not all students are capable of guiding their learning in this way. One of the key features of students who are successful in these systems is the ability to identify and adopt appropriate learning goals. This type of behavior is one of the key components of self-regulated learning.

## 2.2 Self-Regulated Learning

Self-regulated learning (SRL) is a term used to describe the behaviors of students who actively control their learning goals and outcomes (Schunk & Zimmerman, 2003). Among other things, SRL involves students actively setting goals and making conscious choices to measure and evaluate their progress towards them. Self-regulated learners deliberately reflect on their knowledge and learning strategies and make adjustments based on past success and failure. They are actively engaged in learning tasks and demonstrate persistence in the face of challenges (Zimmerman, 1990).

**Components of Self-Regulated Learning.** While many models of SRL have been proposed, there are three main components (Figure 2) that have been identified as central to SRL processes: *cognition/behavior*, *metacognition* and *motivation* (Schraw et al., 2006; Winne & Hadwin, 1998; Zimmerman, 1990). The *cognitive/behavioral* component of SRL involves the skills and behaviors that students demonstrate to support their own learning. This in-



**Figure 2.** Components of self-regulated learning (Schraw et al., 2006)

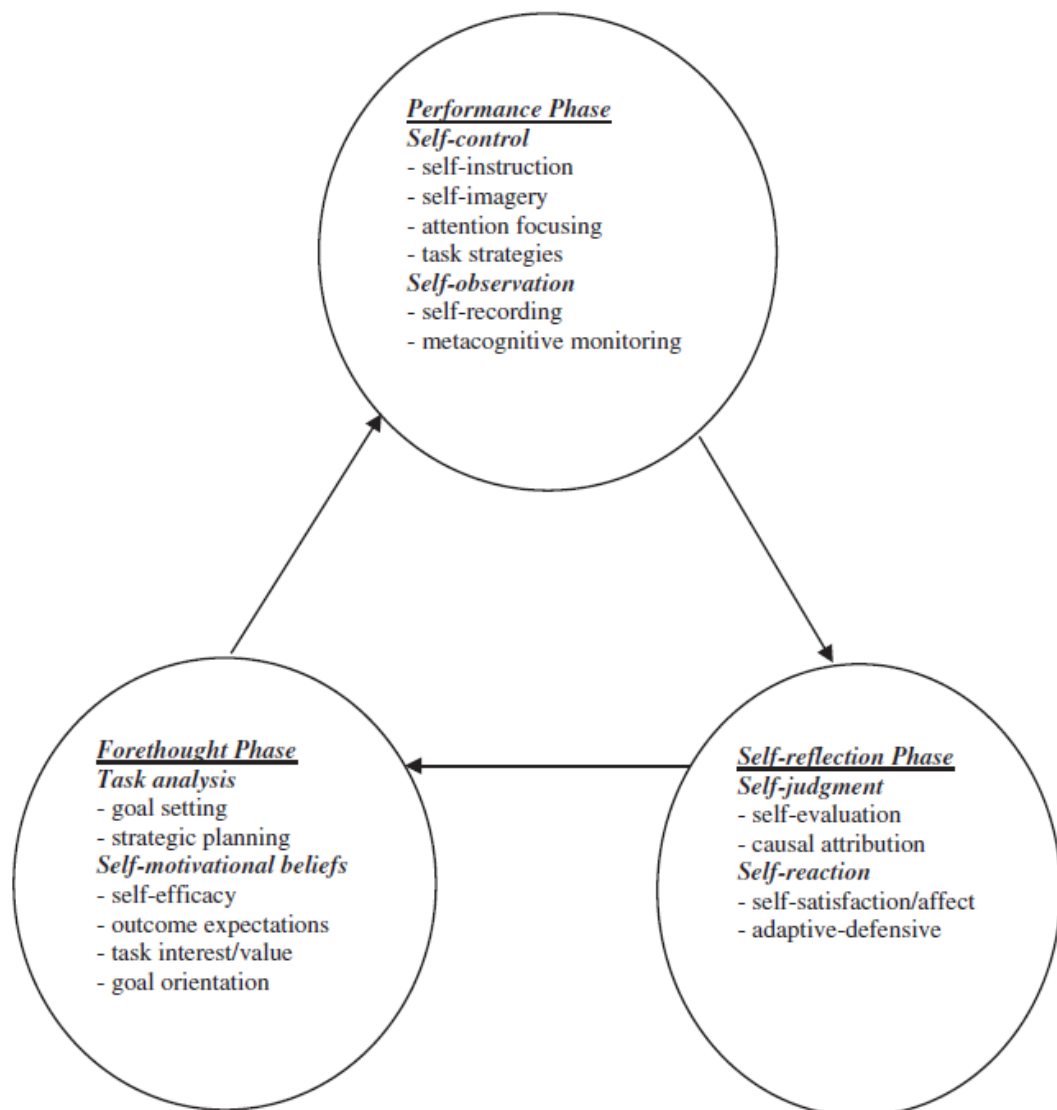
cludes problem solving strategies such as decomposing large problems into smaller steps, making use of careful observations, and analyzing outcomes. Students also demonstrate critical thinking skills such as analyzing information for credibility, making ties with existing knowledge, and synthesizing information from a variety of sources. Another key behavior is the ability to create positive learning environments by reducing distractions and seeking out sources of appropriate help.

The *metacognitive* component of SRL involves the regulation of cognitive activities. Self-regulated learners engage in metacognitive behaviors such as setting appropriate learning goals, planning strategies to accomplish them, monitoring and evaluating progress and adapting or refining behaviors as necessary. These students are more consciously attuned to their level of knowledge while engaged in a task and are better able to organize thoughts and open questions. They can then use this information to make efficient decisions about the best approaches for learning or completing a task.

Finally, the *motivational* component of SRL involves students' beliefs about the value of the task and their own abilities to succeed. Self-regulated learners typically have intrinsic interest in learning tasks, persist in the face of difficulty, and demonstrate high self-efficacy. They tend to have epistemological beliefs that support the idea that they can learn with the appropriate effort and opportunity. These students are ready and willing to learn.

Together, these three components result in learners who are interested and capable of engaging and succeeding in learning tasks. While each component is powerful and important, all three must be present for effective self-regulated learning (Schraw et al, 2006). A student may have the metacognitive ability to recognize issues with a problem solving task, but if they lack the necessary cognitive/behavioral skills they will likely still struggle. Similarly, a student may have the cognitive and metacognitive skills to learn effectively, but if they lack the motivation to do so then they are less likely to succeed. Consequently, it is important to consider and support each of these components of self-regulated learning.

**Cyclical Nature of Self-Regulated Learning.** Most theories of self-regulated learning describe it as a cyclical process, with each phase impacting the next (DiBenedetto & Zimmerman, 2013; Schraw et al., 2006; Winne & Hadwin, 1998; Zimmerman, 2000). Three phases are commonly described as part of this cycle: planning, performing, and then reflecting (Figure 3). Zimmerman (2000) further describes the subprocesses that are part of each



**Figure 3.** Three phases and subprocesses of self-regulated learning. (DiBenedetto & Zimmerman, 2013)

phase. The *planning* phase is made up of task-analysis and self-motivational beliefs. During task-analysis the student identifies possible outcomes of the learning activity and specifically identifies the outcomes that are most desirable and consequently arrives at a concrete goal. The student can then plan how this goal can be achieved by identifying the most appropriate strategies. During this phase self-motivational beliefs also come into play. A student's goal orientation and interest in the task can drive what goals they are more likely to focus on. The student's beliefs about his or her ability will also impact what goals are chosen and how hard the student is willing to work to achieve them.

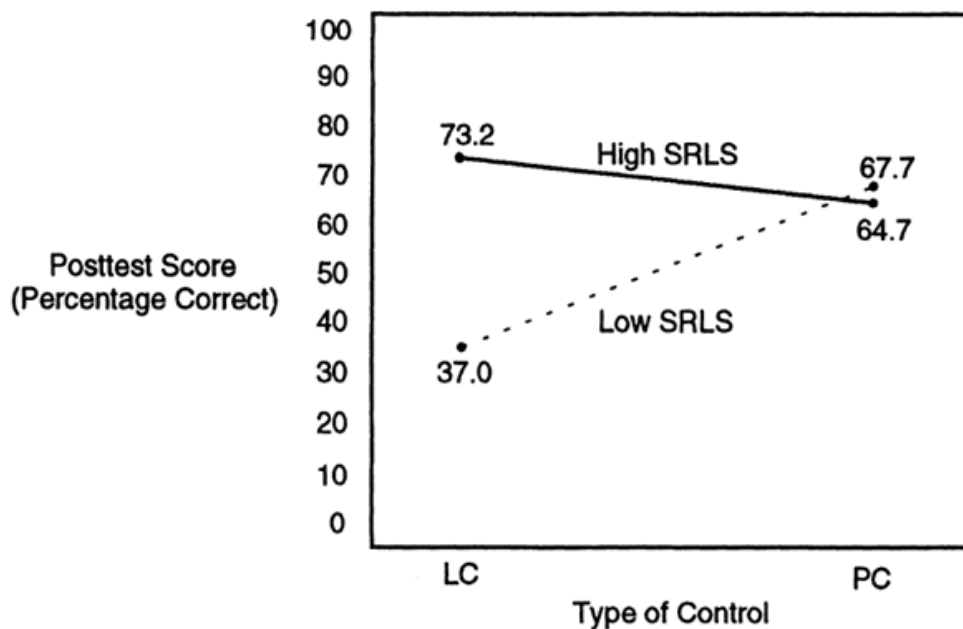
Students next enter the *performance* phase where they execute the plans and utilize the strategies identified in the planning phase. Zimmerman (2000) identifies two processes associated with this phase: self-control and self-observation. Self-control involves actively focusing attention on important material and working to avoid distractions. Students also engage in practices to make the most of learning such as identifying key points, integrating information, breaking down tasks, and constructing mental models of learned information. The self-observation processes involve students attending to the strategies they use and the resulting outcomes. To be most effective, self-observations must be timely, accurate, structured, and ideally focused on the positive aspects of performance rather than the negative. Students may record observations externally or just mentally note observations about their performance. These observations may also lead to experimentation where students recognize they may need to try something new to see if it more effective.

The third phase is *reflection*, where students evaluate their performance based on the strategies used and observations of the effectiveness. This phase involves self-judgment and self-reaction. During self-judgment students compare their observations of performance to their initial goals or other standards. This judgment may involve correctness of a solution, improvement over prior performance, or success relative to others. Students' reactions to these judgments depend on how they attribute cause. For example, a student may attribute failure to an innate inability or they may believe they selected the wrong strategy for accomplishing their goal. The attribution will impact how they feel about themselves and how they

may approach the task moving forward. Students who blame the task will identify new strategies as they re-enter the planning phase. Students who blame themselves may defensively alter their goals or may simply have reduced interest and self-efficacy impacting the next planning phase. The student's beliefs, intentions, and behaviors will continually change as students repeatedly cycle through the three phases of self-regulated learning.

**Individual Differences in Self-Regulated Learning.** Understanding the processes underlying SRL, it is unsurprising that individuals who are better able to regulate their learning in intentional and reflective ways are more likely to achieve academic success (Zimmerman, 1990). However, while it seems all students apply self-regulatory behaviors during learning, the degree of competency is unfortunately broad, even among students of the same age (Schunk & Zimmerman, 2003). Research suggests that low self-regulation during learning activities is simply due to students not having sufficient opportunities to practice and these skills are often not highlighted or required during typical classroom learning (Schunk & Zimmerman, 2003; Schunk & Ertmer, 2000; Zimmerman & Kitsantas, 2002). However, these skills can be improved with practice and with targeted interventions (Kostons et al., 2012; Schunk & Swartz, 1993a, 1993b; Schunk & Zimmerman, 2003).

Self-regulated learning can also vary with context. According to Schraw (2010) “many students are not sufficiently self-regulated, and even good learners experience trouble regulating learning in unfamiliar domains or challenging circumstances” (p. 258). This may be especially true in open-ended computer-based learning environments that are both novel and cognitively demanding (Azevedo & Witherspoon, 2009; Young, 1996). There is evidence that these open-ended environments may also exacerbate any differences in self-regulated learning skills by offering additional benefits to skilled learners (Kostons et al., 2012; Young, 1996). In a study by Young et. al (1996), students with high and low levels of self-regulation performed equally well in a strictly guided learning environment. However, when interacting with an open-ended environment with more student control and autonomy, high self-regulated learners had improved outcomes over the guided system while low self-



**Figure 4.** Posttest score by self-regulated learning score using learner (LC) or program(PC) controlled software. (Young, 1996)

regulated learners performed significantly worse when they were required to guide their own learning (Figure 4). This highlights the role these environments may play in exploring individual differences in self-regulated learning strategies and skills.

### 2.3 Measuring SRL

SRL behaviors can vary between individuals and contexts and have significant impacts on learning and performance outcomes. Consequently, the ability to measure and evaluate SRL is important for forming a full picture of the learning process both in traditional classrooms and in intelligent tutoring systems. Additionally, research suggests that SRL scaffolding should be tailored to a student's existing demonstration of SRL skills (White & Frederiksen, 1998; Young, 1996). Too much or too little scaffolding can harm learning or engagement outcomes (Koedinger & Alevan, 2007). To accomplish this, accurate and meaningful measures of SRL behaviors are critical.

**Current Approaches.** Measures of SRL typically operate under one of two views of SRL: as an *aptitude* or as an *event* (Winne & Perry, 2000). When SRL is measured as an *aptitude* a single metric represents an individual's propensity and skill at regulating their learning behaviors. Measures of this type often include questionnaires where students are asked whether or how often they engage in specific SRL behaviors such as asking for help or finding a quiet place to study. These measures rely on students' honesty and clear recollection of their typical study behaviors. SRL may also be measured as an aptitude using teacher evaluations or structured interviews. These methods present a holistic view of SRL as a trait of an individual learner based on typical behaviors and response patterns (Winne & Perry, 2000).

An alternative view is measuring SRL as an *event*. In this view, SRL is dependent on the specific learning context and measurements reflect individual actions of SRL behaviors. Measures of SRL as an event include external observations of performance, think aloud protocols, and trace methodologies. These methods of measurement seek to examine how individuals respond to specific stimuli and identify both the presence or absence of SRL behaviors as well as the varieties of skills and responses demonstrated. In this view, SRL is represented by a single action or pattern of actions in a given context and is very specific to the task at hand (Winne, 2010).

While these two views of SRL measurement conceptualize SRL differently, researchers agree that it is important to consider both aspects in order to get a full picture of SRL (Winne, 2010). While some students are more skilled at regulating their own learning, the successful use of these skills depends highly on the specific learning context (Schraw, 2010). Furthermore, SRL skills may develop and change over time based on independent experiences or instruction.

**Challenges.** Managing the conceptualization of SRL as an event and as an aptitude is just one example of the many challenges of measuring SRL. Other issues with traditional measurement strategies are common in measurement of other psychological phenomena as well. First, the act of measuring something may alter the very aspect that is being measured. For

example, asking a student about their current learning goal may cause the student to rethink their current activities and perhaps identify a new goal. Secondly, many measures are invasive and interrupt or put additional demands on learning tasks. For example, think aloud protocols require students to verbalize their actions that may cause additional cognitive load and reduce their ability to work on the task at hand. Finally, many measures require subjects to personally recall or qualify their activities. Questionnaires and interviews that ask about typical behaviors depend on individuals' ability to accurately categorize their past behaviors which may impact the outcome of the measure.

SRL researchers have lamented these challenges of measuring SRL for decades. However, with the increasing availability of computer-based learning environments, new mechanisms for measuring and evaluating SRL have recently come to light. In 2010, a special issue of *Educational Psychologist* featured leading SRL researchers discussing the possibilities and challenges of measuring SRL using trace data generated from computer-based learning environments.

**Trace-based Assessment of SRL.** The rich trace data provided by computer-based learning environments presents novel opportunities for assessing SRL that was previously unavailable. Overall, there are three primary benefits to measuring SRL from trace data:

*Rich, dynamic event data.* In computer-based learning environments, student actions can be logged at a granularity that is infeasible using other online, real-time measures such as think-aloud protocols. These actions can be analyzed to provide a rich picture of individual SRL events and allow “researchers to capture and model the dynamic nature of these processes as they are continually deployed and adjusted during learning” (Greene & Azevedo, 2010, p. 205).

*Unobtrusive measurement.* An additional benefit is that the logging of this information is unobtrusive and does not interfere with the students learning activities. Students are more likely to engage in naturalistic behaviors than when they are being clearly monitored or asked to report on some phenomena (Winne & Perry, 2000). This type of “stealth assess-

ment” has been gaining acceptance because of its ability to capture information about the student without interruption or interference (Shute, 2011).

*Contextualization.* SRL behaviors are always in response to a specific set of stimuli in a particular context (Winne, 2010). Measuring SRL without accounting for these variables misses out on key components of how students react to learning scenarios. Trace assessment naturally focuses on how students engage in SRL in context; specifically, the context of the computer-based learning environment with which they are presently interacting. Additionally, the context also includes the individual learner and the skills, beliefs and aptitudes that they bring to bear on the learning process.

Together, these features allow a richer and more meaningful understanding of SRL processes as they unfold during a learning task. Furthermore, computer-based learning environments often put users in control of their own learning, necessitating the use of SRL strategies. The need for understanding SRL in these environments as well as the measurement benefits provided by these systems has motivated recent examination of these phenomena.

## **2.4 Modeling and Supporting Student SRL Behaviors**

Recent work has examined issues of modeling and supporting SRL in a variety of different computer-based learning environments. For example, in *MetaTutor*, a hypermedia environment for learning biology, think-aloud protocols have been used to examine which strategies students use, while analysis of students’ navigation through the hypermedia environment helps to identify profiles of self-regulated learners (Azevedo, Johnson, Chauncey, & Burkett, 2010). Further studies have examined the role of prompts encouraging students to set and monitor specific learning goals (Azevedo et al., 2012). The results of these investigations identified the importance of providing timely prompts as well as offering feedback on students’ responses to the prompt.

Aleven *et al.* have examined modeling and scaffolding the self-regulatory process of help-seeking in the *Cognitive Tutor* systems (Aleven et al., 2006). In these environments, students may request help at any time though many students do not use this feature correctly.

Some students may request help too early or frequently without spending time on the problem, while others may avoid help even when it would be beneficial. Alevén *et al.* developed a model of ideal help seeking behavior to identify students who were not using the feature correctly. Using this model, the system delivered feedback either encouraging students to use or avoid help. While there were no learning differences between students who received feedback and those who did not, the authors did find that the feedback improved help-seeking behavior in future interactions with the system.

Similarly, researchers have identified patterns of behavior in the *Betty's Brain* system that are indicative of low and high levels of self-regulation (Biswas, Jeong, Roscoe, & Sulcer, 2009). Prompting students to use SRL strategies when these patterns of behavior occur has shown promise in improving student learning. They have also used Hidden Markov Models to identify behavior differences between students who engaged with a version of a system where they received SRL feedback and those who received traditional feedback (Biswas, Jeong, Kinnebrew, Sulcer, & Roscoe, 2010).

While previous work has focused primarily on examining SRL in highly structured problem-solving and learning environments, there has also been work on identifying SRL behaviors in open-ended inquiry environments. For example, work by Shores *et al.* has examined early prediction of students' cognitive tool use in order to inform possible interventions and scaffolding (Shores, Rowe, & Lester, 2011). This work focused in on a single component of SRL rather than considering the extremely large scope of phenomena that can be investigated in an open-ended learning scenario.

Each of these lines of work represents a step closer to real-time assessment of SRL in computer-based learning environments. However, while data-mining techniques have identified possible SRL behavior patterns in some of these systems (Azevedo & Witherspoon, 2009; Biswas *et al.*, 2010), this information has not been incorporated into a model for real-time assessment. Alternatively, researchers who have developed real-time predictive models focus on a single component of SRL such as tool use or help seeking (Alevén, Roll, McLaren, & Koedinger, 2010; Shores, Rowe, & Lester, 2011). To achieve a run-time model

for assessing SRL it will be important to bridge the gap between these two approaches by both focusing on data-mined behavior patterns as well as run-time modeling.

## 2.5 Stealth Assessment

Stealth assessment is a form of real-time empirical modeling that has been the subject of increasing attention, particularly in the context of open-ended game-based environments (Shute, 2011; Shute & Ventura, 2013). The primary objective of stealth-assessment is to seamlessly measure low-level features of student knowledge, behaviors, or characteristics. The assessment is hidden from the student who is engaged in the learning activity. The system tracks fine-grained, naturalistic actions of the student that can be integrated in a model of skills or competencies of interest. The resulting assessment can be used to provide formative feedback with the objective of improving the learning or skill acquisition of the student. Stealth assessment is designed to measure important characteristics without interrupting feelings of engagement and flow. Since the student is unaware they are being measured they may also demonstrate significantly less anxiety than in traditional assessment measures.

Stealth assessment has been heavily explored in open-ended game-based learning environments. Games are of particular interest for a variety of reasons (Shute, 2011; Shute & Ventura, 2013). First, games are engaging and students may be more willing to interact with them than other learning systems. Second, games often involve “learning by doing” (Shute, 2011, p. 506) which is expected to lead to improved learning outcomes. Third, open-ended learning environments require a wide array of skills and provide opportunities to identify student strengths and weaknesses in these areas. Finally, games offer unique opportunities to provide feedback and improve skills and learning so the student receives additional benefit.

Stealth assessment has been successfully used to measure a variety of phenomena. For example, *Newton’s Playground*, a game-based environment for physics learning has been used as a testbed for exploring the validity of stealth assessment (Shute & Ventura, 2013). In this environment, students are asked to create drawings of physical objects that adhere to laws of motion in order to accomplish a specified task. Stealth assessment has been used to

measure persistence, or how long will students continue to work on a problem when they are struggling. It has also been used to measure creativity, or how unique are the solutions of an individual student. Finally, it has been used to measure physics concept knowledge, or how well students understand each of the relevant laws of motion. The outcomes of the stealth assessments have been compared to external measures and have been found to be a valid assessment of these characteristics. These investigations highlight the role that games and stealth assessment can play in recognizing and further supporting student's learning and self-regulatory skills.

## CHAPTER 3

### Student Behavior Corpus

This proposal investigates self-regulated learning in the game-based learning environment CRYSTAL ISLAND (Figure 5). This system includes rich game features and storylines aimed at encouraging student interest and motivation. Additionally, it is an inquiry-based system in which students must gather information, form hypotheses and develop and run tests in order to solve the overarching problem. In the following section the details of the CRYSTAL ISLAND environment will be described along with a corpus collection involving over four hundred students interacting with the system.

#### 3.1 CRYSTAL ISLAND

CRYSTAL ISLAND features a science mystery designed to support the North Carolina 8<sup>th</sup> grade microbiology curriculum. The premise of the mystery is that the student arrives on an island to discover that the research team that has been established there has fallen ill. The camp nurse explains that they have not been able to identify the cause or type of illness and asks



**Figure 5.** CRYSTAL ISLAND learning environment

for the student's help. The student then works to collect clues by talking with virtual characters, running tests on objects in the world, and reading related books and posters. Once the student arrives at the correct source and type of illness and proposes a diagnosis, they have solved the mystery and completed the game.

There are a variety of activities students engage in to learn the material and solve the mystery. Students are encouraged to read books and posters with the related microbiology content. They must converse with characters to understand the symptoms of the ill camp members and to acquire information about what types of things the camp members have been eating. They must gather and run tests on relevant food items to determine what objects might be contaminated. Once a contaminated object has been found they may examine it further with a microscope to distinguish with what it is contaminated. They are also encouraged to keep track of all their findings and hypothesis by taking notes and recording data in their "diagnosis worksheet."

## 3.2 Corpus Collection

Data was collected from students at two North Carolina middle schools interacting with CRYSTAL ISLAND as part of their school day. A primary corpus was collected from the first middle school. This data is used for the analyses of self-regulated learning and models of student behavior presented in subsequent sections of this proposal. A secondary corpus was collected from the second, smaller middle school. This corpus is used as a validation set to ensure that the models and findings from the first corpus extend to novel populations.

**Participants.** The first, primary corpus collection was part of a study in which 296 eighth grade students from local middle school interacted with the CRYSTAL ISLAND environment. After removing instances of incomplete data, the final corpus included data from 260 students. Of these, there were 129 male and 131 female participants. The average age of the students was 13.4 years ( $SD = 0.57$ ). Approximately 1% of the participants were American Indian or Alaska Native, 1% were Asian, 17% were Black or African American, 8% were His-

panic or Latino, 63% were Caucasian, and 10% were of mixed or other races. At the time of the study, the students had not yet completed the microbiology curriculum in their classes.

The second, validation corpus was collected from a study involving 154 eighth grade students from a middle school in same school district as the first study. After cleaning the data, 140 students' data remained in the validation corpus. Of these, there were 64 male and 76 female students. The average age of the students was 13.6 years ( $SD = 0.69$ ). Approximately 1% of the participants were American Indian or Alaska Native, 10% were Black or African American, 21% were Hispanic or Latino, 66% were Caucasian, and 1% were mixed or other races.

**Measures.** A week prior to the interaction, students completed a set of pre-study questionnaires including a test of prior knowledge, as well as several measures of personal attributes. *Personality* was measured using the Big 5 Personality Questionnaire, which describes personality along five dimensions: openness, conscientiousness, extraversion, agreeableness and neuroticism (McCrae & Costa, 1993). *Goal orientation*, which refers to the extent that a student values mastery of material and successful performance outcomes when engaged in learning activities, was also measured using the Achievement Goals Questionnaire (Elliot & McGregor, 2001). Finally, students' *emotion regulation* strategies were measured with the Cognitive Emotion Regulation Questionnaire (Gernefski & Kraati, 2006), which measures the extent to which each of nine common strategies are used by an individual student. Students also completed a researcher-generated curriculum test to measure their level of knowledge before interacting with CRYSTAL ISLAND.

Immediately after completing their interaction with CRYSTAL ISLAND, students were given a post-interaction curriculum test with questions identical to that of the pre-test. Students also completed two questionnaires aimed at measuring students' interest and involvement with CRYSTAL ISLAND. A shortened form of the Intrinsic Motivation Inventory (McAuley, Duncan, & Tammen, 1989) was used to measure student motivation along five factors: interest/enjoyment, perceived competence, effort/importance, pressure/tension, and

value/usefulness. Student engagement and presence in the system was measured using the Presence Questionnaire (Witmer, & Singer, 1998), which includes several subscales such as a sense of immersion and involvement. Students also completed a questionnaire related to their understanding of the CRYSTAL ISLAND mystery, though these measures are outside the scope of this discussion.

**Procedure.** Having completed the pre-test materials a week prior to the study, the students were seated at a station, which included a laptop, mouse, headphones and a set of explanatory materials. Students listened to an introductory presentation by a researcher that included a brief description of the purpose of the study and details about the game controls. The explanatory materials at each student station included this same information printed so they could reference it throughout the study.

During the study, students interacted with CRYSTAL ISLAND for 55 minutes or until they completed the mystery. Student's in-game behaviors were logged in detail by the system. During their interaction they also received an in-game prompt asking them to report on



**Figure 6.** Self-report device

their emotional state (Figure 6). This prompt was described to students as being part of an “experimental social network” that was being used on CRYSTAL ISLAND. Students selected from one of seven emotional states: *anxious*, *bored*, *confused*, *curious*, *excited*, *focused*, and *frustrated*. They were also asked to type a short “status update.”

Once they completed the mystery, or after 55 minutes of interaction, students were directed to a computer lab where they completed the post-interaction materials.

## CHAPTER 4

### Prior Findings on Student Engagement and Learning in CRYSTAL ISLAND

The primary purpose of the CRYSTAL ISLAND learning environment is to increase microbiology domain knowledge with additional emphasis on improving inquiry skills and fostering engagement. Consequently, initial work sought to examine how these phenomena are represented in the collected corpus. First, analyses were conducted to ensure that students experienced learning gains from interacting with CRYSTAL ISLAND. Using data from both corpora, paired t-tests comparing student's pretest ( $M = 6.6$ ,  $SD = 2.3$ ) and posttest ( $M = 8.6$ ,  $SD = 3.4$ ) scores indicated that students' learning gains from using CRYSTAL ISLAND were statistically significant,  $t(399) = 12.5$ ,  $p < 0.0001$ .

#### 4.1 Affect and Motivation

Initial work sought to examine the overall affective experiences of students interacting with CRYSTAL ISLAND (Table 1). It is hypothesized that the open-ended features promote positive affect and engagement, which can in turn contribute to better learning and motivational outcomes. It was found that positive, learning-focused affective states such as focused (23%) and curious (19%) accounted for the majority of student's self-reported emotions. Confusion (16%) and frustration (16%) were the next most frequent emotional states. These states are expected to result from the open-ended nature of the CRYSTAL ISLAND environment. The environment does not tell students specifically what they should be doing at any given time, which may be different from their classroom learning experiences. The somewhat high levels of these emotional states suggest that there may be some students who need increased levels of guidance, though other students may benefit from exploring the environment on their own. Excitement (13%) occurred somewhat frequently while highly negative emotions such as, boredom (8%) and anxiety (4%) were relatively infrequent.

**Table 1.** Frequency and proportion of emotion self-reports

| Emotion      | Primary Corpus |          | Secondary Corpus |          | Total       |          |
|--------------|----------------|----------|------------------|----------|-------------|----------|
|              | Freq           | Per      | Freq             | Per      | Freq        | Per      |
| anxious      | 86             | 4.6%     | 41               | 4.0%     | 127         | 4.4%     |
| bored        | 159            | 8.5%     | 84               | 8.2%     | 243         | 8.4%     |
| confused     | 300            | 16.1%    | 167              | 16.3%    | 467         | 16.2%    |
| curious      | 347            | 18.6%    | 203              | 19.8%    | 550         | 19.1%    |
| excited      | 251            | 13.5%    | 126              | 12.3%    | 377         | 13.1%    |
| focused      | 417            | 22.4%    | 252              | 24.6%    | 669         | 23.2%    |
| frustrated   | 303            | 16.3%    | 150              | 14.7%    | 453         | 15.7%    |
| <b>Total</b> | <b>1863</b>    | <b>-</b> | <b>1023</b>      | <b>-</b> | <b>2886</b> | <b>-</b> |

A repeated measure ANOVA indicated that the states occurred at significantly different frequencies,  $F_{(6, 2394)} = 60.4$ ,  $p < 0.0001$ . Tukey post-hoc analyses indicated the following differences in the occurrence of self-reports: focused > curious > (confused = frustrated = excited) > bored > anxious. While it is difficult to compare this pattern of states to other environments because of differences in collection procedures, there are interesting similarities and differences. Students interacting with both traditional and game-based environments tend to spend most of their time in a state of focus or engaged concentration (Baker et al., 2010). However, positive affective states such as delight, curiosity and excitement appear more prevalent in game systems, compared with confusion in traditional tutoring systems (Baker et al., 2010). These trends suggest that there may be affective benefits to game-based learning environments, although more controlled studies (such as (Hallinen et al., 2009)) are necessary to further understand the differences.

Additional analyses were conducted to examine the role of affect played on learning and motivational outcomes in CRYSTAL ISLAND. Using student's self-reported affective states, correlations were conducted to determine the relationship between the occurrences of

these states and learning outcomes. Results indicated that positive affect was strongly correlated with learning gains,  $r(398) = 0.16$ ,  $p = 0.001$ , while negative affect was negatively correlated with learning gains. Additionally, two negatively valenced emotions appeared to be particularly associated with reduced learning. Both confusion,  $r(398) = -0.11$ ,  $p = 0.027$  and boredom,  $r(398) = -0.15$ ,  $p = 0.035$  were negatively correlated with learning outcomes.

Further investigation sought to identify whether students' affective states corresponded with feelings of value, interest, and motivation towards the task (Table 2). Correlations were run between the occurrence of student emotion and five subscales of the Intrinsic Motivation Inventory (McAuley et al., 1989): Interest/Enjoyment, Perceived Competence, Effort/Importance, Pressure/Tension, and Value/Usefulness. Many affective states had significant correlations with each of these metrics. Additionally, there is strong evidence ( $p < 0.001$ ) that positive affect was associated with increased feelings of interest ( $r=0.49$ ), competence ( $r=0.35$ ), importance ( $r=0.33$ ) and value ( $r=0.35$ ). Understandably, positive affect was also associated with reduced feelings of tension ( $r = -0.20$ ). Together, these results correlating affective states with learning and motivation measures corroborate with many findings in the psychological community. Related research suggests that positive affect can lead to in-

**Table 2.** Affective states correlated with motivation outcome measures.  
Bold indicates  $p < 0.05$ , highlighted indicates  $p < 0.01$

|            | Interest /<br>Enjoyment | Perceived<br>Competence | Effort /<br>Importance | Pressure /<br>Tension | Value /<br>Usefulness |
|------------|-------------------------|-------------------------|------------------------|-----------------------|-----------------------|
| Anxious    | 0.07                    | 0.00                    | 0.05                   | 0.01                  | 0.09                  |
| Bored      | <b>-0.43</b>            | <b>-0.30</b>            | <b>-0.37</b>           | 0.04                  | <b>-0.38</b>          |
| Confused   | <b>-0.15</b>            | -0.11                   | -0.05                  | 0.10                  | -0.03                 |
| Curious    | <b>0.16</b>             | 0.11                    | <b>0.13</b>            | -0.03                 | 0.08                  |
| Excited    | <b>0.32</b>             | <b>0.22</b>             | <b>0.19</b>            | -0.08                 | <b>0.22</b>           |
| Focused    | <b>0.22</b>             | <b>0.15</b>             | <b>0.15</b>            | <b>-0.16</b>          | <b>0.19</b>           |
| Frustrated | <b>-0.18</b>            | -0.09                   | -0.10                  | <b>0.15</b>           | <b>-0.17</b>          |

creased learning (Bless et al., 1996; Kanfer & Ackerman, 1989; Pekrun et al., 2002), while negative emotions are believed to lead to decreased motivation and effort (Meyer & Turner, 2006; Pekrun et al., 2002; Ramirez & Dockweiler, 1987).

## 4.2 Engagement and Disengagement

Prior work also examined the occurrence of disengagement in the form of off-task behavior. Specifically, this involved students disengaging from the learning content and focusing instead on the game-based features of the environment. This is notably different from disengagement in traditional tutoring systems as students are still engaged by the system, though not with the learning task. Analyses conducted with the initial corpus of students indicate that on average, students spent approximate 4.58% ( $SD = 6.82$ ) of their time off-task (Figure 7). While this is significantly lower than reported by many other environments, it is not unex-

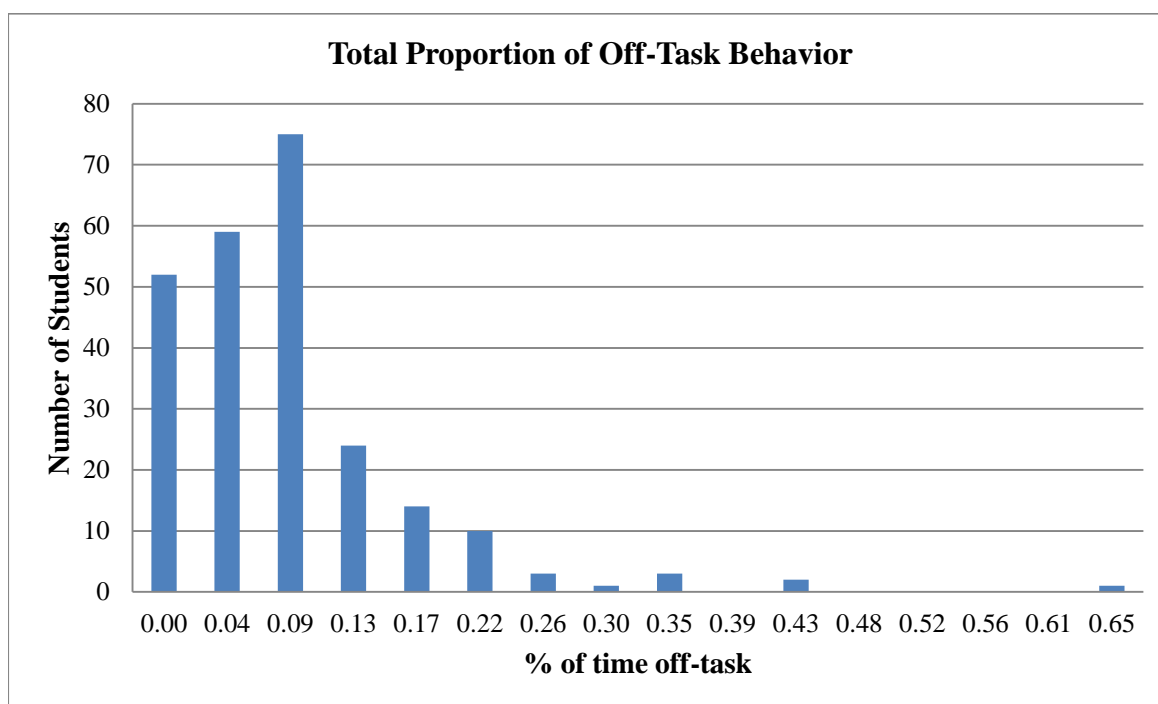


Figure 7. Histogram of off-task behavior

pected given the very different nature of the environment and unique definition of off-task behavior. There was also a wide range between students with approximately a third of students engaging in no off-task behavior, and at the maximum, one student spent 63.2% of his time off-task.

Further analyses were conducted comparing off-task behavior to student learning. Results resembled findings reported from other investigations of off-task behavior in alternate intelligent tutoring systems (Baker, Corbett, Koedinger, & Wagner, 2004; Cocea, Hershkovitz, & Baker, 2009; Gong, Beck, Heffernan, & Forbes-Summers, 2010). Off-task behavior was found to negatively correlate with students' normalized learning gains,  $r(258) = -0.18, p = 0.004$ . There was no evidence that low prior-knowledge students engaged in more off-task behavior, as the correlation between time off-task and pre-test score was not statistically significant,  $r(258) = -0.08, p = 0.21$ . This result contrasted with a previous investigation of off-task behavior using an earlier version of the CRYSTAL ISLAND learning environment (Rowe, McQuiggan, Robison, & Lester, 2009).

The results also highlighted evidence that off-task behavior may have a significant affective component. In particular, total time off-task was negatively correlated with curiosity  $r(258) = -0.12, p = 0.04$  and frustration,  $r(258) = -0.13, p = 0.04$ . This result was surprising given prior work that demonstrated frustration as a trigger for off-task behavior (Baker et al., 2010). The finding prompted an examination of whether off-task behavior helps alleviate frustration in the CRYSTAL ISLAND environment.

Analyses examining the affective state transitions of students based on their off-task behavior indicated that students may be using off-task behavior as a means to regulate negative affect. For example, students reporting frustration and subsequently going off-task were most likely to return to feeling focused, while students who remained engaged in the learning task reported higher levels of boredom. Further analyses indicated that not all students were able to appropriately use off-task behavior as a mechanism for regulating affect. In fact, for students who experienced affective benefits for going off-task, there was no negative correlation between duration of off-task behavior and learning gains. However, for students who did

not appear to have this skill the negative impacts of off-task behavior on learning were still observed (Sabourin et al., 2013).

Together these findings indicate a positive relationship between engagement and learning behaviors in the CRYSTAL ISLAND environment. They also indicate differences in student's abilities to regulate their behavior during interactions. Inspired by these findings, further investigations sought to examine the role of inquiry strategies on learning, engagement, and interest.

### 4.3 Inquiry Behaviors

Prior work examining students' inquiry behaviors in the game-based learning environment suggested that effective inquiry strategies (e.g., gathering background information prior to formulating and testing hypotheses in a virtual laboratory) were not necessarily associated with improved content learning gains (Sabourin, Rowe, et al., 2012b). However, effective inquiry strategies were associated with improved problem-solving outcomes. Conversely, students who did not use good strategies (e.g., gathering background information after formulating and testing hypotheses) were less effective at solving the overall task. These observations led to the hypothesis that individual differences in inquiry strategies may also be associated with differences in affective outcomes. In particular, emotions such as frustration and curiosity were anticipated to correlate with different inquiry behaviors.

For this analysis students were split into two groups based on whether they gathered more information prior to hypothesis testing or waited until they received failed results to gather background information. Overall it was found that more effective inquiry behaviors corresponded with better affective experiences (Sabourin, Rowe, et al., 2012a). Specifically, students who performed more information-gathering behaviors prior to hypothesis testing reported more positive emotions, such as curiosity,  $t(318) = 1.97, p = 0.05$ , and excitement,  $t(318) = 2.51, p = 0.01$ . It is possible that these states may have fueled students to learn more about the environment. Good inquiry strategies were also associated with fewer negative cognitive-affective states, such as frustration,  $t(122) = 2.09, p = 0.04$ , and confusion,  $t(122) =$

2.14,  $p = 0.03$ . This is likely due to the fact that these students were more efficient at solving the problems. These findings provide further evidence that there is utility in introducing supplemental guidance for some students in game-based learning environments. While some students are able to engage in effective inquiry behaviors and have positive learning and affective outcomes, others are not as successful. Going forward, it will be important to identify and support learners with less effective inquiry strategies to improve the overall experience for all learners.

## **CHAPTER 5**

### **Identifying Self-Regulated Learning**

In order to build predictive models of the cognitive, metacognitive and behavioral components of self-regulated learning, it was necessary to first devise measures and classifications for these components. The objective was to select processes and phenomena that were particularly relevant to the CRYSTAL ISLAND environment and were representative of the important processes. We were also interested in exploring different mechanisms for identifying behaviors and classifying students. For this reason, three different approaches were taken for each of the categories: machine learning, manual annotation, and validated survey measure. It is hoped that by addressing a broad spectrum of behaviors, processes, and techniques the underlying ideas and methods described will be more easily applied to measuring SRL in other learning environments.

#### **5.1 Cognitive/Behavioral Classifications**

To identify evidence of cognitive/behavioral strategy use we were interested in identifying the key differences in problem solving behaviors. CRYSTAL ISLAND is naturally open-ended environment and students have a great deal of flexibility in the strategies they use. Through many years of iterative refinement of CRYSTAL ISLAND, observational and experimental data have indicated that there appears to be a distinct dichotomy between successful students and less successful students (Rowe et al., 2010). However, few problem-solving metrics have directly predicted learning in explorations of CRYSTAL ISLAND. For example, in the study of inquiry behaviors discussed in Section 4.3, there was evidence that students who gathered data before hypothesis testing were more effective problem-solvers but this did not translate to differences in learning outcomes. Consequently, one of the objectives of identifying cognitive/behavioral classifications of students was to be able to identify problem-solving strate-

gies that were both more effective and tied to learning outcomes. The approach to this task was k-means clustering.

**K-Means Clustering.** K-means clustering is an unsupervised machine-learning technique that is designed to recognize groups of similar observations from a sample. The primary objective is to create  $k$  groupings from a set of observations while minimizing the distance of each observation from the mean of the cluster to which it belongs. Finding the optimal cluster is NP-hard, although iterative refinement algorithms support rapid convergence on possible clusters. In this process,  $k$  random means are selected at the start. Each observation is then assigned to a cluster based on the closest mean. The means are then updated to reflect the actual mean of the observations within the cluster. This process is repeated until the means (and assigned observations) no longer change.

One of the primary challenges in k-means clustering is specifying the number of clusters being searched for. Sometimes there is a natural number based on the types of patterns being searched for. In cases where there is no obvious choice it is common to try multiple values of  $k$  and select the option that yields the most meaningful clusters whether meaning is defined observationally or by the relative “tightness” of the clusters. Because k-means clustering is often used as an unsupervised exploration of the data it is common that researchers are unsure of what they are looking for until potential clusters have been approved. Another issue with k-means clustering is determining how to deal with cases of missing data. A common solution in many machine learning approaches is to replace missing data with average or set values of the feature. However, because of the nature of the k-means clustering algorithm this can inappropriately skew the clusters to focus on members with the same missing values. Finally, the random assignment of initial clusters may impact the clusters that are ultimately recognized. A variety of initialization approaches have been proposed, each with their own bias. The approach that is most appropriate typically depends on the data and algorithms being used.

Because of its usefulness in exploring data, k-means clustering has been applied to many educational applications (Baker & Yacef, 2009). For example, Amershi, and Conati have applied clustering techniques to physiological signals to identify common patterns of response to events in a game for teaching factorization (Amershi & Conati, 2006). It has also been used to group patterns of students in web-based learning environments for the purpose of improving recommendations (Tang & McCalla, 2004). One unique application of clustering used the output of learned Hidden Markov Models and clustered students based on the transition probabilities from the learned models. The clusters of these transitions indicated different patterns of engagement in how students interacted with a mathematics tutoring system (Beal, Mitra, & Cohen, 2007).

**Classification Procedure.** To identify clusters related to cognitive strategy use and problem solving, we focused on clustering using features directly related to these behaviors. Specifically, features were related to the testing behavior and completing the mystery along with the data-gathering behaviors associated with these tasks. Here, data-gathering behaviors refer to interactions with the primary information sources in CRYSTAL ISLAND including conversations with characters and reading books, posters, and the information in the microbiology app. In total, 10 features were selected, of these 6 related to testing and 4 to mystery completion. These features were selected to indicate problem-solving effectiveness as well as inquiry strategy use. Testing attributes included (1) the number of tests run before success, (2) the time that passed between the first test and a successful test, (3) the total number of tests run, (4) the number of data gathering behaviors before testing, (5) the number of data gathering behaviors before a successful test, and (6) the proportion of features (4) to (5). Similar features were used for the mystery completion which is dependent upon submission of the diagnosis worksheet. These features included (1) the total number of submissions, (2) the number of data gathering behaviors before the first submission, (3) the total number of investigative behaviors (4) the proportion of features (2) to (3). Here we did not include features related to number of attempts before a correct submission and time before a correct submis-

sion because only 33% of students successfully completed the mystery. Consequently, there were too many instances of missing data for these two variables.

Missing data was an issue for many of the other features as well. Of the 260 students in the original sample 13 (5%) never conducted a single test. These students were removed from *all* subsequent analyses since they represent significant outliers. Of the remaining 247 students, 13% of students did not achieve a successful test and 11% of students never attempted to submit their worksheet for verification. For students who did not achieve a successful test, we took data from fellow students who achieved a successful test after more than one attempt. These students took an average of 15.0 minutes ( $SD = 8.5$ ) and 7.3 ( $SD = 5.9$ ) tests before conducting a successful test, representing an average of 121 seconds per test. On average their proportion of investigative behaviors before the first test to the total at the end was 0.65 ( $SD = 0.24$ ). For students who did not achieve a successful test, the missing value of tests before success was set by adding 7.3 (average tests for successful students) to the number of tests they had completed so far. This approach was taken since we had no evidence of how close students were to reaching a successful test and to reflect the number of test they had already completed unsuccessfully. The time before successful tests was then set to 70 minutes because students were stopped from gameplay at 55 minutes, and we estimate that it would have taken 15 more minutes to conduct remaining tests until success. The proportion of investigative behaviors before and during testing was set to the average of 0.65 and for each student the number of investigative behaviors before success was adjusted to make the set proportion accurate.

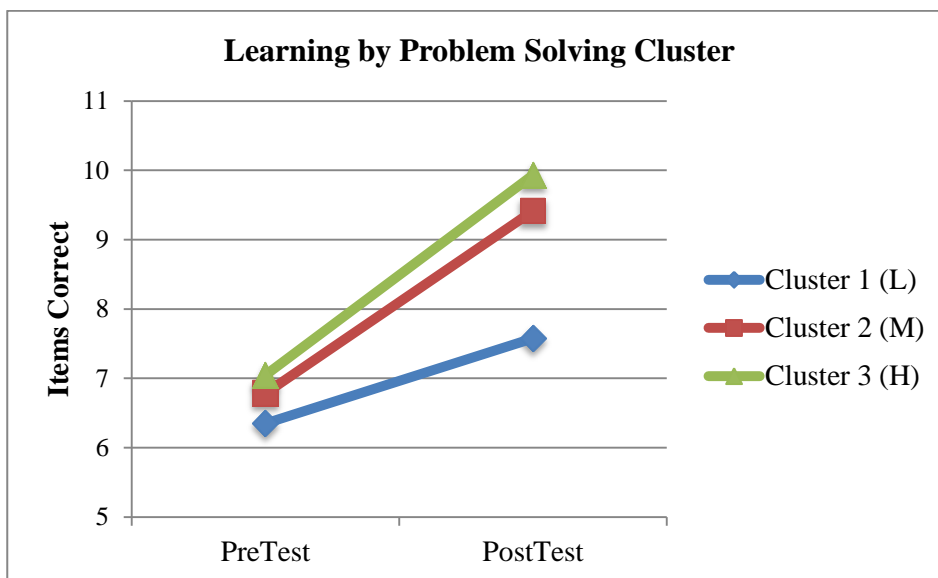
A similar approach was taken for students who did not have a successful diagnosis worksheet check. On average students took 2.5 ( $SD = 1.5$ ) attempts before solving the mystery with an average of 12.3 ( $SD = 10.1$ ) minutes between the first and successful check. The average proportion of investigative behaviors before the first check was 0.71 ( $SD = 0.23$ ). Furthermore, the average time between a successful test and completing the mystery was 16.0 ( $SD = 12.3$ ) minutes. For students who did not have a successful check we added 2.5 checks to the number they had completed so far. For estimating the time until they would

have completed the mystery we added 12.3 minutes to the game end time of 55 minutes (67.3 minutes) if they had at least one worksheet check and a successful test. Otherwise we added 16.0 minutes to the time of the successful test or estimate of successful test time if they had not yet had a successful test. Again we set the proportion of investigative behaviors to the average of 0.71 and the value of total investigative behaviors to a value to make the proportion accurate for the individual student.

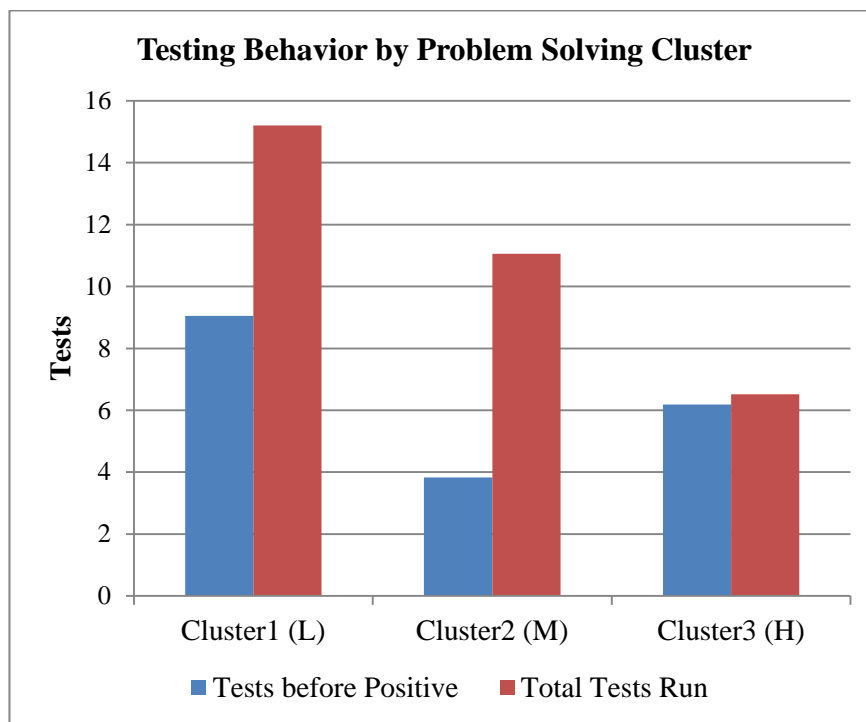
Once the missing values had been handled, k-means clustering was run using the WEKA machine learning toolkit (Hall et al., 2009). Clusters of  $k = 2, 3,$  and 4 were tried. To select the best clusters we chose to optimize based on the number of clusters that produced the most descriptive differences in learning gains. Based on this criterion, 3 clusters were chosen.

**Class Details.** The three chosen clusters were examined to determine evidence of problem solving strategies and learning gains. An ANOVA with Tukey post-hoc tests indicated clusters had significant learning differences between clusters  $F_{(2, 244)} = 7.03, p = 0.001$ . Specifically, students in Cluster 1 ( $N = 104$ ) had significantly lower learning gains than students in Clusters 2 ( $N = 73$ ) and 3 ( $N = 70$ ) (Figure 8). This is particularly interesting because none of the original features that went into making the clusters had any direct ties to learning gains. This suggests that the clusters may be capturing more meaningful groups of strategies than any individual metric. Further analyses indicated that there was no statistically significant difference between clusters and pre-test score,  $F_{(2, 244)} = 1.85, p = 0.16$ , indicating that problem solving strategies were likely not caused by prior knowledge differences.

Unsurprisingly, analyses indicate a significant difference for each of the features that was used to form the clusters (Table 3) However, there are some particularly interesting differences when looking at problem solving behaviors. Cluster 3 ( $N = 110$ ) represented efficient problem solving. These students conducted fewer tests and worksheet checks before arriving at a correct solution. They also moved on to the next problem-solving step after receiving positive feedback on a test and did not continue testing. (Figure 9) These students



**Figure 8.** Learning differences by problem-solving cluster



**Figure 9.** Testing behavior by problem-solving cluster

**Table 3.** Class differences on cluster features

| Feature                                | Cluster | 1      | 2      | 3      | F      | p       | Class Differences |
|--|---------|--------|--------|--------|--------|---------|-------------------|
|  | N       | 104    | 73     | 70     |        |         |                   |
|  | Class   | L      | M      | H      |        |         |                   |
| Total Tests Run                        | mean    | 15.4   | 10.9   | 6.9    | 15.26  | < .0001 | H < L, M          |
|  | std.    | 11.2   | 10.0   | 6.4    |        |         |                   |
| Time before Successful Test            | mean    | 2355.9 | 1416.5 | 2542.2 | 41.15  | <.0001  | M < L, H          |
|  | std.    | 1105.9 | 507.9  | 1056.0 |        |         |                   |
| Tests Run Before Success               | mean    | 8.4    | 4.4    | 6.2    | 22.8   | <.0001  | M < L < H         |
|  | std.    | 9.6    | 4.2    | 5.7    |        |         |                   |
| Data-Gathering Before First Test       | mean    | 10.8   | 8.7    | 21.5   | 99.97  | <.0001  | L, M < H          |
|  | std.    | 5.4    | 3.9    | 10.3   |        |         |                   |
| Data-Gathering Before Successful Test  | mean    | 14.0   | 10.2   | 28.5   | 115.14 | <.0001  | M < L < H         |
|  | std.    | 7.2    | 4.3    | 13.2   |        |         |                   |
| Data-Gathering Proportion over Testing | mean    | 0.811  | 0.849  | 0.767  | 8.06   | 0.0004  | H < L, M          |
|  | std.    | 0.218  | 0.219  | 0.219  |        |         |                   |
| Total Worksheet Checks                 | mean    | 4.6    | 3.0    | 2.4    | 43.6   | <.0001  | H < M < L         |
|  | std.    | 1.8    | 1.6    | 1.1    |        |         |                   |
| Data-Gathering before First Check      | mean    | 14.8   | 12.6   | 29.5   | 177.23 | <.0001  | L, M < H          |
|  | std.    | 6.7    | 4.6    | 11.0   |        |         |                   |
| Total Data-Gathering                   | mean    | 20.9   | 21.5   | 37.9   | 112.79 | <.0001  | L, M < H          |
|  | std.    | 9.4    | 6.5    | 13.6   |        |         |                   |
| Data-Gathering Proportion over Checks  | mean    | 0.710  | 0.619  | 0.780  | 21.31  | < .0001 | M < L, H          |
|  | std.    | 0.137  | 0.220  | 0.166  |        |         |                   |
| Pre-Test                               | mean    | 6.3    | 6.8    | 7.0    | 1.85   | 0.1592  | --                |
|  | std.    | 2.2    | 2.5    | 2.6    |        |         |                   |
| Post-Test                              | mean    | 7.6    | 9.4    | 9.9    | 12.85  | <.0001  | L < M, H          |
|  | std.    | 3.1    | 3.4    | 3.4    |        |         |                   |
| Learning Gains                         | mean    | 1.2    | 2.6    | 2.9    | 7.03   | 0.0011  | L < M, H          |
|  | std.    | 3.0    | 3.5    | 3.2    |        |         |                   |
| Normalized Learning Gains              | mean    | 0.087  | 0.193  | 0.237  | 6.96   | 0.0011  | L < M, H          |
|  | std.    | 0.246  | 0.300  | 0.292  |        |         |                   |

conduct the most investigative behaviors before problem solving which may explain their success. Students in Cluster 2 showed puzzling patterns of behavior. They reached positive solutions faster than students in Cluster 3, but would continue testing and gathering information even after they had the solution. This suggests these students may have been unclear of the next problem solving step, or may have had reached the correct test by chance. These students have the lowest level of investigative behaviors so they may not be particularly well informed. Finally, Cluster 1 represents inefficient problem solving. These students took far longer to reach a successful solution and also continued investigations after a positive test, similar to Cluster 2. This pattern of behavior may stem from poor understanding of the problem or a “guess and check” approach. These students also have low levels of investigative behaviors. Chi-squared analyses also indicated that students finished at significantly different rates based on the cluster ( $\chi^2(2, N=247) = 28.3, p < .0001$ ), with 16.3% of Cluster 1 students solving the mystery and 50.7% and 47.14% of Cluster 2 and 3 students finishing, respectively.

Based on the patterns of behavior demonstrated and the associated learning outcomes, Cluster 1, 2, and 3 were reclassified as Low, Medium, and High, respectively for the remainder of the analyses. These classifications are used as an example of cognitive/behavioral self-regulation. The differences suggest better problem-solving abilities and more targeted strategy use which is most closely related to this categorization of SRL.

## 5.2 Metacognitive Classifications

One of the primary challenges of open-ended learning environments is goal-setting and monitoring. In these environments student goals are not necessarily made explicit. Students are often free to set their own goals and may not have clear indicators of progress (Land, 2000). Goal-setting and monitoring is one of the key components of metacognitive processes of SRL (Winne & Hadwin, 1998; Zimmerman, 1990) and is related to success in many learning scenarios (B. Zimmerman, 2008). Furthermore, these behaviors are expected to be of increased importance in CRYSTAL ISLAND due to the open-ended nature of the environment and

time-constrained interaction. In this work goal-setting is identified through manually annotated statements that were collected as part of an experimental, in-game “social network”.

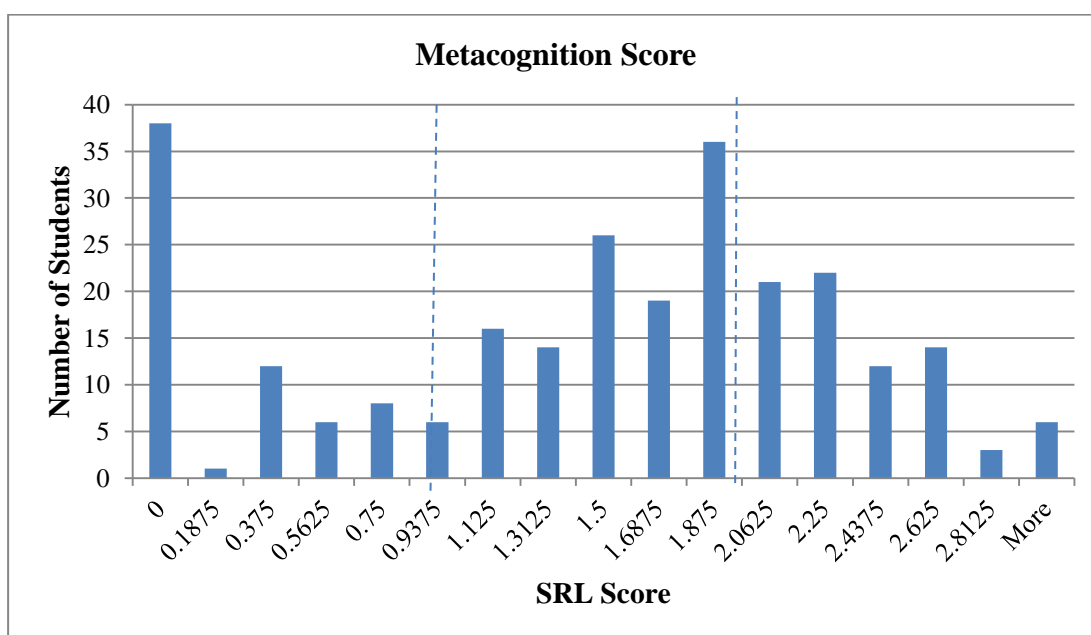
**Annotation Protocol.** During their interaction with CRYSTAL ISLAND, students were occasionally prompted to indicate their emotional state and to briefly type a few words about their current “status”, similarly to how they might update their status in an online social network. Though these prompts did not directly ask students to report on goal-setting or monitoring, many student responses contained evidence of these behaviors. Student status reports were tagged for metacognitive evidence use using the following four ranked classifications: (1) specific reflection, (2) general reflection, (3) non-reflective statement, or (4) unrelated (Table 4). This ranking was motivated by the observation that setting and reflecting upon goals is a hallmark of self-regulatory behavior and that specific goals are more beneficial than those that are more general (Zimmerman, 2008).

**Table 4.** Metacognition tagging scheme

| <b>SRL Category</b>        | <b>Description</b>  | <b>Examples</b>  |
|----------------------------|---|--|
| <i>Specific reflection</i> | Student evaluates progress towards a specific goal or area of knowledge   | “I am trying to find the food or drink that caused these people to get sick.”<br>“Well...the influenza is looking more and more right. I think I'll try testing for mutagens or pathogens – [I] ruled out carcinogens” |
| <i>General reflection</i>  | Student evaluates progress or knowledge but without referencing a particular goal   | “I think I’m getting it”<br>“I don’t know what to do”  |
| <i>Non-reflective</i>      | Student describes what they are doing or lists a fact without providing an evaluation   | “testing food”<br>“in the lab”   |
| <i>Unrelated</i>           | Any statement which did not fall into the above three categories is considered unrelated, including non-word or unidentifiable statements | “having fun”<br>“arghhh!”  |

From the remaining primary corpus of 247 students, a total of 1742 statements were collected, resulting in an average of 7.2 statements per student. All statements were tagged by one member of the research team with a second member of the research team tagging a randomly selected subset (10%) of the statements to assess the validity of the protocol. Interrater reliability was measured at  $\kappa = 0.77$ , which is an acceptable level of agreement. General reflective statements were the most common (37.2%), followed by unrelated (35.6%), specific reflections (18.3%) and finally non-reflective statements (9.0%).

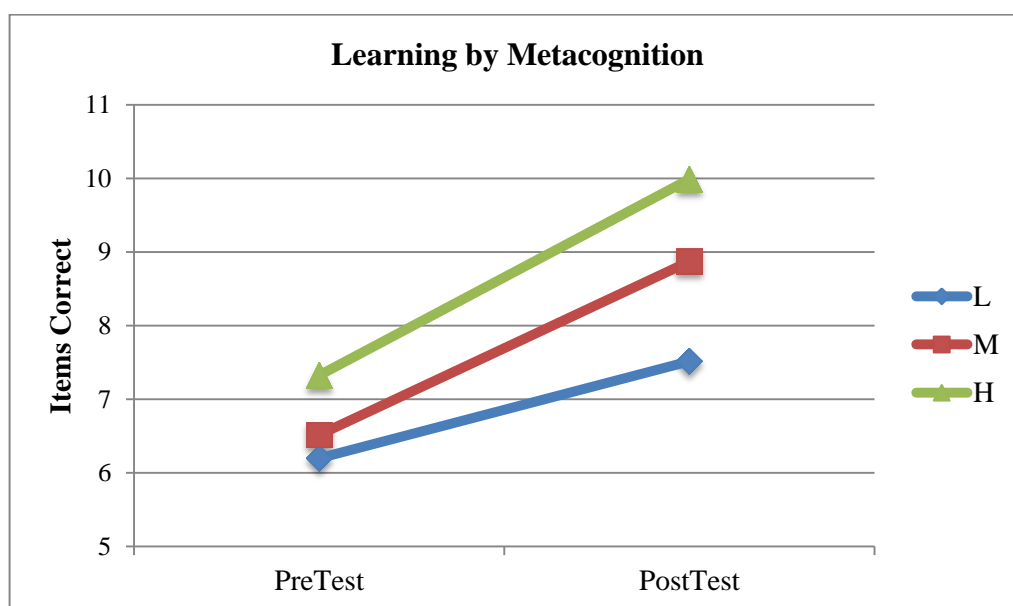
After tagging, students were given an overall metacognition score based on the average score of their metacognitive statements. Scores ranged from 0 (for unrelated) to 3 points (for specific reflection) per statement. The average was taken since students had different frequencies of reports based on when they solved the mystery. Because students could not enter statements at will we did not consider the count of statements as indicative of metacognition. The average metacognition score was 1.40 ( $SD = 0.84$ ) with a minimum and maximum score of 0 and 3, respectively. An even ternary split was then used to assign the students to a Low, Medium, and High metacognition category. This even split occurred at the



**Figure 10.** Histogram of metacognitive score

levels of 1.0 and 2.0 (Figure 10). This split implies that High metacognition students are making mostly specific reflections, while Low metacognition students are primarily un-reflective. Three groups were chosen to cover a larger gradient of metacognitive behaviors and to match the number of classifications in the cognition/behavior component of SRL.

**Class Details.** Just as analyses were conducted to explore the differences in behavior and learning between classes in the cognitive/behavioral component of SRL, similar analyses were conducted for the metacognitive classifications. An ANOVA indicated a significant difference in normalized learning gains between the groups ( $F_{(2, 244)} = 4.40, p = 0.01$ ). Tukey post-hoc comparisons indicated that both High and Medium metacognition students experienced significantly better learning gains than Low metacognition students at the  $\alpha = 0.05$  level. Analyses also indicated that there were significant differences on pre-test scores between groups ( $F_{(2, 244)} = 4.84, p < 0.01$ ) suggesting that students with high metacognitive tendencies may be better students or perhaps their increased prior knowledge helped them to identify and evaluate their goals more efficiently. Figure 11 shows the pre- and post- test scores

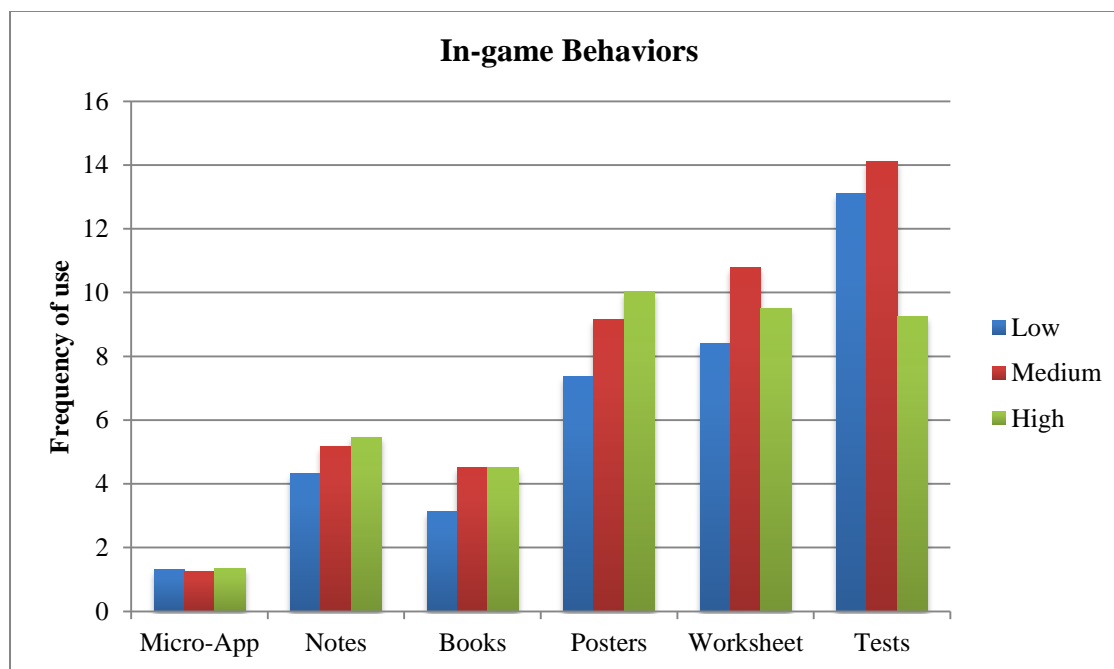


**Figure 11.** Learning differences by metacognitive class

across groups, highlighting both the differences in pre-knowledge and learning during interaction with CRYSTAL ISLAND.

The next set of analyses was conducted to investigate differences in student behavior based on their SRL tendencies. A chi-squared analysis indicated that the percentage of students who solved the mystery varied only moderately significantly on SRL group ( $\chi^2(2, N = 247) = 5.01, p = 0.08$ ). Low metacognition students had the lowest completion rate of 25.6% compared to 39.3% and 40.7% of Medium and High metacognition students, respectively. It is expected that better goal-setting may have aided students in completing the mystery in a more timely manner, though effective problem-solving behaviors would also be required to solve the mystery.

While a significant difference in students' abilities to solve the mystery was not found, there were differences in the in-game resources that students used (Figure 12). An ANOVA indicated a significant difference in data-gathering behavior,  $F(2,244) = 7.88, p < 0.01$ , with Tukey post-hoc tests indicating that High metacognition students interacted with



**Figure 12.** In-game behaviors by SRL classification

more resources than Medium or Low metacognition students. Further analyses were conducted to see if there were any differences on specific resources. ANOVAs for student use of each of these resources indicated a significant difference in student use of *posters* ( $F_{(2, 244)} = 5.11$ ,  $p < 0.01$ ), though none of the other resources had significant differences. Tukey post-hoc tests indicated that High metacognition students looked at more posters than Low metacognition students. There was also a significant difference in testing behavior,  $F_{(2, 244)} = 6.68$ ,  $p < 0.01$ , with High metacognition students conducting significantly fewer tests than Medium or Low metacognition students. As with the problem-solving analysis for the cognitive component of SRL, this is evidence of more effective resources use though the trends are less strong.

Further analyses indicated no significant difference in off-task behavior,  $F_{(2, 244)} = 1.82$ ,  $p = 0.16$ , suggesting that these behaviors do not have a clear relationship with metacognitive strategies. This is corroborated by additional findings that some students are capable of intelligently using off-task behavior to regulate negative affect strategy where other students are simply disengaging and see harmful impacts on learning. Consequently, off-task behavior is indicative of both effective and maladaptive strategies of self-regulation and without more detailed understanding it is difficult to determine which of the two is occurring for any individual student.

### 5.3 Motivation Classification

As a game-based learning environment, CRYSTAL ISLAND offers unique opportunities for fostering motivation and interest. It is possible that students who would otherwise be unmotivated by scientific learning opportunities would experience increased motivation with this environment and be more likely to self-regulate. This context-specific aspect of motivation and self-regulated learning supports the need for context-base measures and assessment.

**Measure of Motivation.** The Intrinsic Motivation Inventory (IMI) is designed to report on multiple dimensions of interest experience during a targeted activity. The six dimensions

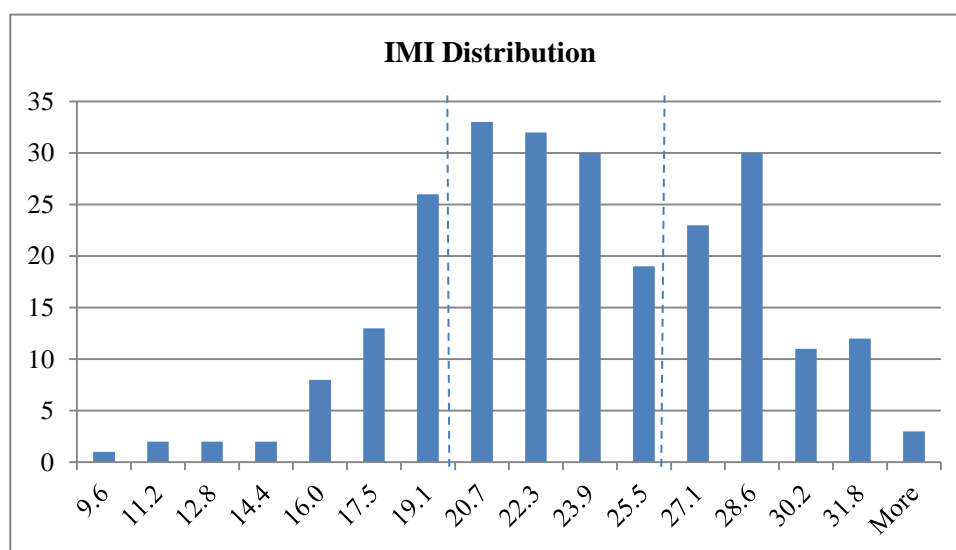
measured are Interest/Enjoyment, Perceived Competence, Effort/Importance, Pressure/Tension, and Value/Usefulness, Perceived Choice. The last of these was not included as part of the study because it was not an outcome measure of interest. Each subscale is measured by multiple Likert-scale questions where students report on a scale of 1-7 how true a statement is of themselves, such as “I put a lot of effort into this” and “I thought this was a boring activity!” Each subscale is scored based on the average of the items while inverting statements that are phrased in the negative.

The IMI has been used as a measure of motivation in a variety of studies related to self-regulation. For example, Ryan *et al.* have explored the relationship between subscales and persistence during problem solving (Ryan, Koestner, & Deci, 1991) and have additionally correlated affect and learning measures with subscales of the inventory in a non-directed text-based learning system (Ryan, Connell, & Plant, 1990). These findings mirror the correlations found within the CRYSTAL ISLAND environment as reported in Section 4.1.

**Results of IMI.** The range of possible scores on the IMI is 1-7 on each of the 5 included subscales, resulting in a total range possibility of 5-35 (Table 5). The average score on the IMI scale was 22.9 (SD=4.6), with a minimum score of 9.6 and a maximum of 33.4. In gen-

**Table 5.** IMI subscale differences by motivation class

| Subscale   | N Class      | 86 L       | 86 M       | 75 H       | F     | p       | Class Differences |
|------------|--------------|------------|------------|------------|-------|---------|-------------------|
| Interest   | mean<br>std. | 3.4<br>1.0 | 4.6<br>1.0 | 6.2<br>0.7 | 181.1 | < .0001 | L < M < H         |
| Competence | mean<br>std. | 3.4<br>1.4 | 4.4<br>1.0 | 5.6<br>1.1 | 68.29 | <.0001  | L < M < H         |
| Effort     | mean<br>std. | 4.0<br>1.0 | 4.9<br>1.0 | 6.0<br>0.8 | 101.9 | <.0001  | L < M < H         |
| Pressure   | mean<br>std. | 3.6<br>1.4 | 4.2<br>1.5 | 4.1<br>1.7 | 4.2   | 0.017   | L < M             |
| Value      | mean<br>std. | 3.6<br>1.3 | 4.9<br>1.1 | 6.4<br>0.7 | 133.6 | <.0001  | L < M < H         |

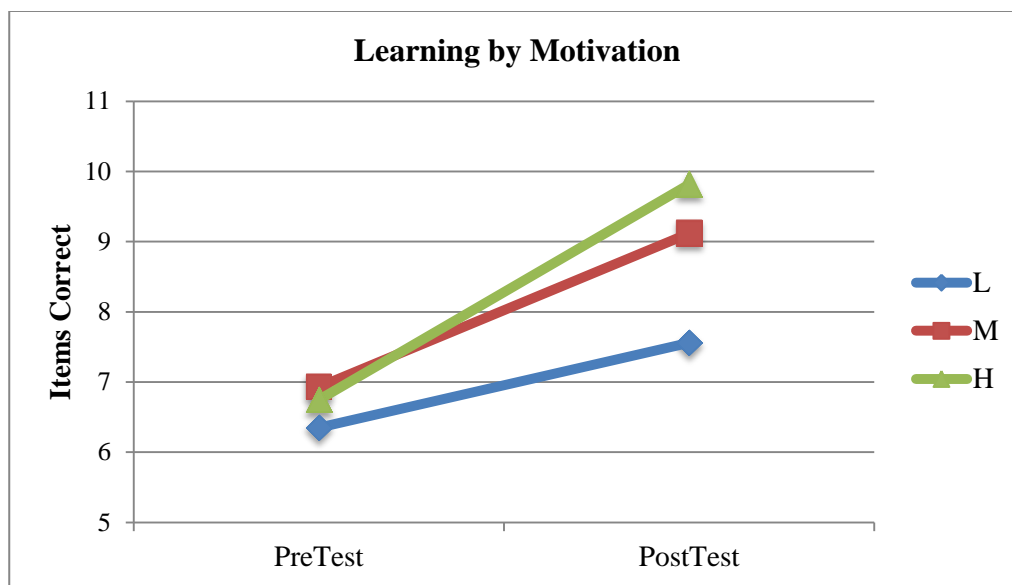


**Figure 13.** Histogram of IMI Score

eral the individual sub-score values were positive with an average score of 4.7 ( $SD$  1.7) on Interest, 4.9 ( $SD$  = 1.2) on Effort, and 4.9 ( $SD$  = 1.7) on Value. In general students reported feeling competent during their interaction ( $M$  = 4.4,  $SD$  = 1.5) with very moderate feelings of pressure ( $M$  = 4.0,  $SD$  = 1.5).

Using the total IMI score, students were divided into three even groups. Low motivation students had total scores less than 20.2, while High motivation students had a score of 25.9 or higher (Figure 13). Interestingly, the Low/Medium cut-off is approximately the point where students report less interested than the median value on the Likert-scale. These students therefore were on average uninterested in the interaction. Medium students had positive interest levels but not at the same levels as some of the highly motivated students interacting with CRYSTAL ISLAND.

**Class Details.** When comparing classes of students by motivation, analyses again indicated significant differences in learning gains,  $F_{(2,244)} = 7.06$ ,  $p < 0.01$ . Tukey post-hoc tests indicated that High motivation students had significantly greater learning gains than Medium or Low motivation students (Figure 14). Further analyses indicated no difference in pre-test



**Figure 14.** Learning differences by motivation class

score  $F_{(2,244)} = 1.15$ ,  $p = 0.32$ . This suggests that prior knowledge was not responsible for differences in motivation.

Further analyses examined differences in game behaviors. A chi-squared analysis indicated that students completed the mystery at different levels depending on their motivation, group ( $\chi^2(2, N = 247) = 10.3$ ,  $p < 0.01$ ). 50% of highly motivated students completed the mystery, while only 33% of Medium and 26% of Low motivation students successfully finished the game. Additional analyses indicated no significant difference in overall data investigation behaviors  $F_{(2,244)} = 0.94$ ,  $p = 0.39$ , though there was a significant difference in the number of books read by students,  $F_{(2,244)} = 4.79$ ,  $p = 0.01$ , with High motivation students reading significantly more books than Low motivation students. Perhaps this difference relates to the information density in the books, which are the richest and possibly most challenging resources in the environment. It is possible that only highly motivated students were willing to engage with these materials. There were also significant differences in testing behavior,  $F_{(2,244)} = 3.52$ ,  $p = 0.03$ , with Low motivation students conducting more tests than High motivation students. This is again evidence of unguided, careless behavior and possibly

attempting to game the system. Analyses of off-task behavior suggested significant differences in off-task behavior,  $F_{(2,244)} = 3.01$ ,  $p = 0.05$  though Tukey post-hoc tests did not identify any significant differences. However, it is interesting to note that all five of the outliers exhibiting more than 20% off-task behavior fell in the Low motivation class.

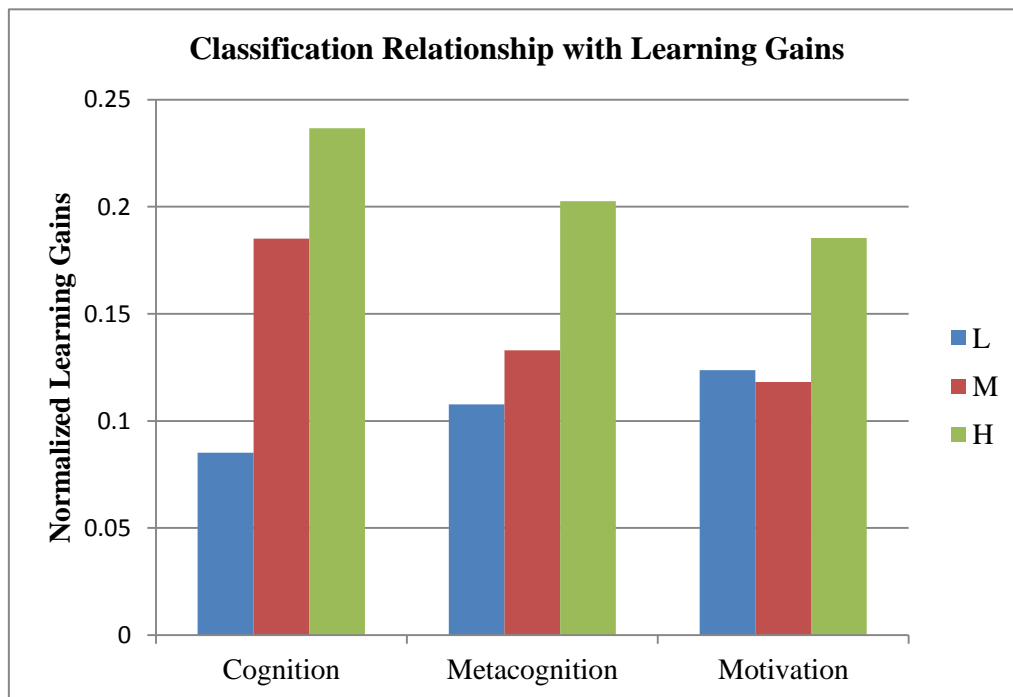
It is hypothesized that game-based environments foster interest and motivation though there is often concern that this may disproportionately motivate students who are already inclined to use digital games and may ostracize other students. For this reason analyses were conducted to identify differences in motivation based on self-reported gaming behavior. Analyses indicated no significant difference in self-reported gaming frequency (not at all, not often, sometimes, regularly, all the time) and motivation classification, ( $\chi^2(8, N = 247) = 10.5$ ,  $p = 0.22$ ). Similarly there was no difference in the number of hours spent weekly on digital games and motivation classification ( $\chi^2(8, N = 247) = 12.3$ ,  $p = 0.15$ ). This suggests that gaming experience was not a significant influence on motivation outcomes.

## 5.4 Discussion

In order to explore patterns of student behavior and build predictive models it was necessary to establish discrete classes of students. Clustering of problem solving strategy was used to identify classes of cognitive SRL. Annotation of student goal setting and monitoring was used to create classes of metacognitive SRL. An externally validated measure was used to identify classes of motivational SRL. We have explored how these classifications relate to measures of interest (Table 6). Overall it was found that all proposed classifications were able to differentiate successful students based on learning gains. Students scoring high on the cognitive, metacognitive, and motivational measures of interest showed significantly better learning outcomes than students who did not (Figure 15). Classifications were also tied to further measures of interest such as in-game behavior and engagement. Overall, this provides evidence that the proposed classifications offer meaningful distinctions of students and are a good categorization of students' self-regulated learning.

**Table 6.** Classification details

|                |   | N   | Pretest |      | Posttest |      | NLG  |      |
|----------------|---|-----|---------|------|----------|------|------|------|
|                |   |     | Mean    | SD   | Mean     | SD   | Mean | SD   |
| Cognition      | L | 104 | 6.35    | 2.17 | 7.58     | 3.04 | 0.09 | 0.24 |
|                | M | 73  | 6.78    | 2.57 | 9.41     | 3.33 | 0.19 | 0.32 |
|                | H | 70  | 7.04    | 2.57 | 9.93     | 3.47 | 0.24 | 0.27 |
| Meta-cognition | L | 82  | 6.20    | 2.24 | 7.51     | 3.05 | 0.09 | 0.25 |
|                | M | 84  | 6.51    | 2.31 | 8.87     | 3.72 | 0.18 | 0.29 |
|                | H | 81  | 7.32    | 2.59 | 9.99     | 2.96 | 0.21 | 0.28 |
| Motiva-tion    | L | 86  | 6.35    | 2.29 | 7.56     | 3.38 | 0.08 | 0.30 |
|                | M | 86  | 6.93    | 2.54 | 9.12     | 3.33 | 0.17 | 0.27 |
|                | H | 75  | 6.75    | 2.42 | 9.81     | 3.11 | 0.24 | 0.24 |

**Figure 15.** Learning gains by classification

While the objective was to represent phenomena from each of the three components of SRL, it is important to note that the classes presented do not encompass all of the features of each component. For example, we consider problem solving strategies as evidence of cognitive SRL, though there are a variety of other processes involved in the cognitive components of self-regulated learning. This was done because problem-solving strategy was expected to be one of the cognitive components of SRL that would be most relevant to success in CRYSTAL ISLAND. The types of cognitive, metacognitive, and motivational features that are of most interest to other researchers will likely vary depending on the exact details of their environment. It is expected that no two environments will have exactly replicable measures of SRL. Instead, we aim to provide a framework for identifying processes of interest and developing classifications accordingly. This is the motivation behind using a data-mining, annotation, and validated measure for creating the different classes. We hope this shows that classes can focus on any meaningful set of behaviors and can be identified using a variety of processes. There are many ways to represent self-regulated learning measures of interest, and we believe that representing components from cognitive, metacognitive, and motivational processes will yield the most robust models of SRL.

## **CHAPTER 6**

### **Feature Selection and Creation**

After specifying classifications of interest, the next step is to identify the features that will be used in building the predictive run-time models. Here, it is important to identify features that are both predictive and meaningful. Furthermore, since we are modeling three distinct components of SRL, each model will have its own set of features. Consequently, data mining techniques such as feature selection and differential sequence mining are used to identify and create features of interest for modeling each component of SRL. Specifically, feature selection is used to identify personal attribute and occurrence event features that are most predictive of each component of SRL. Differential sequence mining is used to identify predictive patterns of behavior, which are then used to generate contingency and patterned contingency event features. In this way, each feature type described by Winne (2010) is represented and has been empirically learned from the corpus.

#### **6.1 Feature Selection**

As discussed in Chapter 2, a common distinction in SRL measurement is whether it is a stable student attribute, or whether it is an event occurring in response to some stimuli (Winne & Perry, 2000). In general, it is accepted that both views are necessary to represent a full picture of SRL. In a similar vein, both attribute and event information is important in providing evidence of SRL. Highly self-regulated students have general tendencies to be more driven, persistent and cognizant of their learning. However, the exact application of these tendencies varies with environment (Schraw, 2010) so it is important to consider actions and events as further evidence of SRL. In order to identify the attribute and event features that are most predictive of SRL in CRYSTAL ISLAND, stepwise logistic regression was applied to a large feature set.

**Stepwise Regression.** The objective of feature selection is to take a large, complete set of features and create a subset of features that offer the most predictive power on the outcome variable. Stepwise regression is a common tool for this approach. Stepwise logistic regression involves iteratively adding and removing features to a predictive logistic regression model based on whether the inclusion of the feature significantly improves the model's predictive capabilities. A chi-squared test is used to measure how well each logistic regression model fits the data. There are a variety of approaches for choosing which features are included. In forward selection, the model starts with no information and features are iteratively added based on which offers the greatest reduction in predicted error. This process is repeated until there is no longer any improvement, or if the improvement is below a certain threshold. Backward elimination works in the opposite manner, with all features initially included in the model. Features are then removed based on if their removal offers an improvement in the model. There are also many hybrid approaches such as stepwise selection, where the model is iteratively built up as in forward selection; however, at each step it is tested to determine if any features can be eliminated using the same technique as backward elimination.

**Procedure.** The stepwise logistic regression was run using the SAS® 9.3 statistical modeling package for each component of SRL. A significance level of  $\alpha < 0.05$  was required for a feature to be added to or remain in the selected model. Forward selection, backward elimination, and stepwise selection were explored for each component. In total 24 personal attributes and 31 event features were considered in the models. Of the 55 possible features, forward selection and stepwise selection recommended the same set of features across the three components: 15 for cognition, 22 for metacognition and 20 for motivation. This was lower than backward elimination, which recommended 18 features for cognition, 25 for metacognition and 20 for motivation. In each case the features recommended by forward and step-wise selection represented a subset of the other two approaches.

For this reason the features recommended by forward selection were used for subsequent analyses. A smaller subset of features is expected to be more likely to generalize to unseen population. Backward elimination often yields better training accuracy but is more

prone to overfitting (Alpaydin, 2004). There is additional concern with the Bayesian techniques used for predictive modeling that too many features can cause sparsity and complexity problems.

**Results.** Closer examination of the features selected for each component of SRL yields interesting patterns. Of the 15 features (Table 7) for the cognitive component related to problem – solving strategy, only 7 are in-game features. This is surprising given that the classes were generated from in-game features alone. Most of the features selected related to resource use and goal progression. The other 8 features focused mostly on emotion regulation and personality, with only one goal orientation feature selected.

Of the 22 features (Table 8) selected for the metacognitive component of goal setting,

**Table 7.** Feature selection for cognition

| <b>Feature</b>                  | <b>Step Added</b> | <b>Chi-Square</b> | <b>Pr &gt; ChiSq</b> |
|---------------------------------|-------------------|-------------------|----------------------|
| Notes                           | 1                 | 230.5725          | <.0001               |
| Total Goals Completed           | 2                 | 81.6103           | <.0001               |
| Posters                         | 3                 | 117.0847          | <.0001               |
| Books                           | 4                 | 59.822            | <.0001               |
| Personality - Openness          | 5                 | 42.0558           | <.0001               |
| Personality - Extraversion      | 6                 | 34.8224           | <.0001               |
| Microbiology App                | 7                 | 24.6856           | <.0001               |
| Time Since Last Goal Completed  | 8                 | 28.4416           | <.0001               |
| Gaming Hours                    | 9                 | 15.3196           | <.0001               |
| Worksheet Checks                | 10                | 16.5646           | <.0001               |
| Goal Ori – Performance Approach | 11                | 15.423            | <.0001               |
| Pre Test                        | 12                | 18.3612           | <.0001               |
| Em Reg – Other Blame            | 13                | 23.5285           | <.0001               |
| Em Reg - Catastrophizing        | 14                | 8.6524            | 0.0033               |
| EmReg – Self Blame              | 15                | 10.9565           | 0.0009               |

**Table 8.** Feature selection for metacognition

| <b>Feature</b>                    | <b>Step Added</b> | <b>Chi-Square</b> | <b>Pr &gt; ChiSq</b> |
|-----------------------------------|-------------------|-------------------|----------------------|
| Average Status Length             | 1                 | 92.7707           | <.0001               |
| Em. Reg. – Catastrophizing        | 2                 | 36.987            | <.0001               |
| Personality – Agreeableness       | 3                 | 18.3774           | <.0001               |
| Personality - Conscientiousness   | 4                 | 16.8945           | <.0001               |
| Personality - Extraversion        | 5                 | 17.6251           | <.0001               |
| Tests Run                         | 6                 | 14.2446           | 0.0002               |
| Pre Test                          | 7                 | 10.9388           | 0.0009               |
| Goal Ori. – Mastery Avoidance     | 8                 | 11.0209           | 0.0009               |
| Em. Reg. – Perspective Taking     | 9                 | 8.8732            | 0.0029               |
| Total Goals Completed             | 10                | 9.2172            | 0.0024               |
| Worksheet Access                  | 11                | 7.578             | 0.0059               |
| Em. Reg. - Acceptance             | 12                | 7.6458            | 0.0057               |
| Em. Reg. – Positive Reappraisal   | 13                | 7.833             | 0.0051               |
| Em. Reg. – Planning               | 14                | 9.2928            | 0.0023               |
| Microbiology App                  | 15                | 6.6014            | 0.0102               |
| Gaming Frequency                  | 16                | 7.2205            | 0.0072               |
| Personality – Openness            | 17                | 6.8417            | 0.0089               |
| Off Task Percentage               | 18                | 6.2412            | 0.0125               |
| Personality – Neuroticism         | 19                | 5.516             | 0.0188               |
| Time Since Last Goal Completed    | 20                | 5.6885            | 0.0171               |
| Goal Ori. – Performance Avoidance | 21                | 5.2023            | 0.0226               |
| Goal Ori. – Performance Approach  | 22                | 4.5318            | 0.0333               |

7 were related to in-game behaviors. The behaviors cover a broad spectrum from goal completion, resources use, and off-task behavior. Unsurprisingly, 5 of the 15 personal features were related to emotion regulation. Metacognition involves monitoring goals and progress and using this information adaptively, which is very similar to the processes involved in cog-

**Table 9.** Feature selection for motivation

| <b>Feature</b>                   | <b>Step Added</b> | <b>Chi-Square</b> | <b>Pr &gt; ChiSq</b> |
|----------------------------------|-------------------|-------------------|----------------------|
| Goal Ori. – Mastery Approach     | 1                 | 60.0079           | <.0001               |
| Personality - Extraversion       | 2                 | 15.7138           | <.0001               |
| Em. Reg. – Positive Refocusing   | 3                 | 12.7388           | 0.0004               |
| Em. Reg. – Self Blame            | 4                 | 17.7955           | <.0001               |
| Personality – Openness           | 5                 | 13.6539           | 0.0002               |
| Pre Test                         | 6                 | 10.4515           | 0.0012               |
| Posters                          | 7                 | 8.7843            | 0.003                |
| Microbiology App                 | 8                 | 8.6356            | 0.0033               |
| Gaming Hours                     | 9                 | 10.1582           | 0.0014               |
| Personality – Agreeableness      | 10                | 9.702             | 0.0018               |
| Off Task Percentage              | 11                | 7.0603            | 0.0079               |
| Worksheet Right                  | 12                | 6.0429            | 0.014                |
| Worksheet Checks                 | 13                | 7.4744            | 0.0063               |
| Personality – Conscientiousness  | 14                | 5.0835            | 0.0242               |
| Worksheet Updates                | 15                | 3.8581            | 0.0495               |
| Worksheet Filled                 | 16                | 4.7212            | 0.0298               |
| Em. Reg. – Other Blame           | 17                | 4.7612            | 0.0291               |
| Em. Reg. - Catastrophizing       | 18                | 4.9318            | 0.0264               |
| Em. Reg. – Positive Reappraisal  | 19                | 5.2857            | 0.0215               |
| Goal Ori. – Performance Approach | 20                | 4.0115            | 0.0452               |

nitive emotion regulation. Additionally, all five of the personality features and three of the four goal-orientation behaviors were selected. These features also relate to how students approach learning tasks and monitor their progress.

Of the 20 features (Table 9) selected for motivation, seven were related to in-game behaviors. These focused primarily on worksheet use, but also involved off-task behavior. The remaining 13 personal attributes were primarily dominated by attributes associated with

positivity. For example, the two goal orientation attributes selected were for approach characteristics, which is associated with more positive affect. The personality attributes of openness and agreeableness are also associated with a positive outlook.

Overall, attribute features were selected more frequently than in-game behaviors. This is particularly interesting when considering how predictive models and their features may be generalized to other learning environments. It is significantly easier for other researchers to apply the same widely available metrics than to identify similar patterns and events in their environment. The focus on attribute features suggests that the models may be easily adapted to other environments. However, event attributes do provide additional information and are practically more feasible to collect in a run-time system.

## 6.2 Feature Creation

Though the event features were selected less frequently than the attribute features, these events only considered occurrence-level information. They measured how often resources had been used, how many goals had been completed, and how often information was tracked. Winne (2010) argues that when utilizing event features it is important to consider contingency information. He suggests that SRL be considered as a set of IF-THEN relationships. In this paradigm, occurrence data focuses only on THENs without considering the IF that preceded it. *Contingent* events, on the other hand, specify an IF-THEN relationship. For example, IF Sarah reads a book, THEN she will take a note. This provides more information than either event alone and instead indicates patterns of processing information and responding to specific stimuli. *Patterned contingency* takes the IF-THEN relationship a step further, by representing the proportion of time the THEN follows the IF. For example, IF Sarah reads a book, THEN she will take a note 30% of the time. This provides more information about how frequently and consistently a strategy is applied. Including contingent features in the modeling process involves first identifying sequences of behavior that are indicative of SRL. To accomplish this we utilize a differential sequence mining approach.

**Differential Sequence Mining.** Differential sequence mining is a technique that has recently been adapted from the genomics field and applied to educational data (Kinnebrew & Biswas, 2012; Kinnebrew, Loretz, & Biswas, 2013). The objective is to identify patterns of sequences that occur at statistically different frequencies across pre-determined groups of patterns. This approach identifies two metrics for representing the frequency of a pattern in different groups. The sequence support (*s-support*) metric refers to the percentage of sequences the pattern occurs in, regardless of frequency. Alternatively, the instance support (*i-support*) metric represents the average number of times the pattern occurs per sequence. The algorithm can be simplified into three steps:

1. *Action abstraction.* Traditional logging often does not represent actions in the most meaningful way. For example, CRYSTAL ISLAND logs every movement and action in the environment at a rate of multiple logs per second. Individually, these items are not meaningful. Therefore, the first step involves identifying the meaningful components of a sequence and creating a sequence of abstracted information
2. *Identify the most common patterns.* To ensure the analyses are meaningful, only patterns that occur at some level of regularity across the population are considered. For this, an *s-support* threshold is set. This means that at least some number of students have to have demonstrated the pattern at least once in order to be considered for analysis. An appropriate threshold is determined by the sparsity of the patterns. Common thresholds include 20% for sparse data and 50% for more richly populated data.
3. *Compare differences in frequency.* Next the *i-support* metric is calculated for each student and each common pattern. This value is then differentially compared across groups to identify significant differences in frequency.

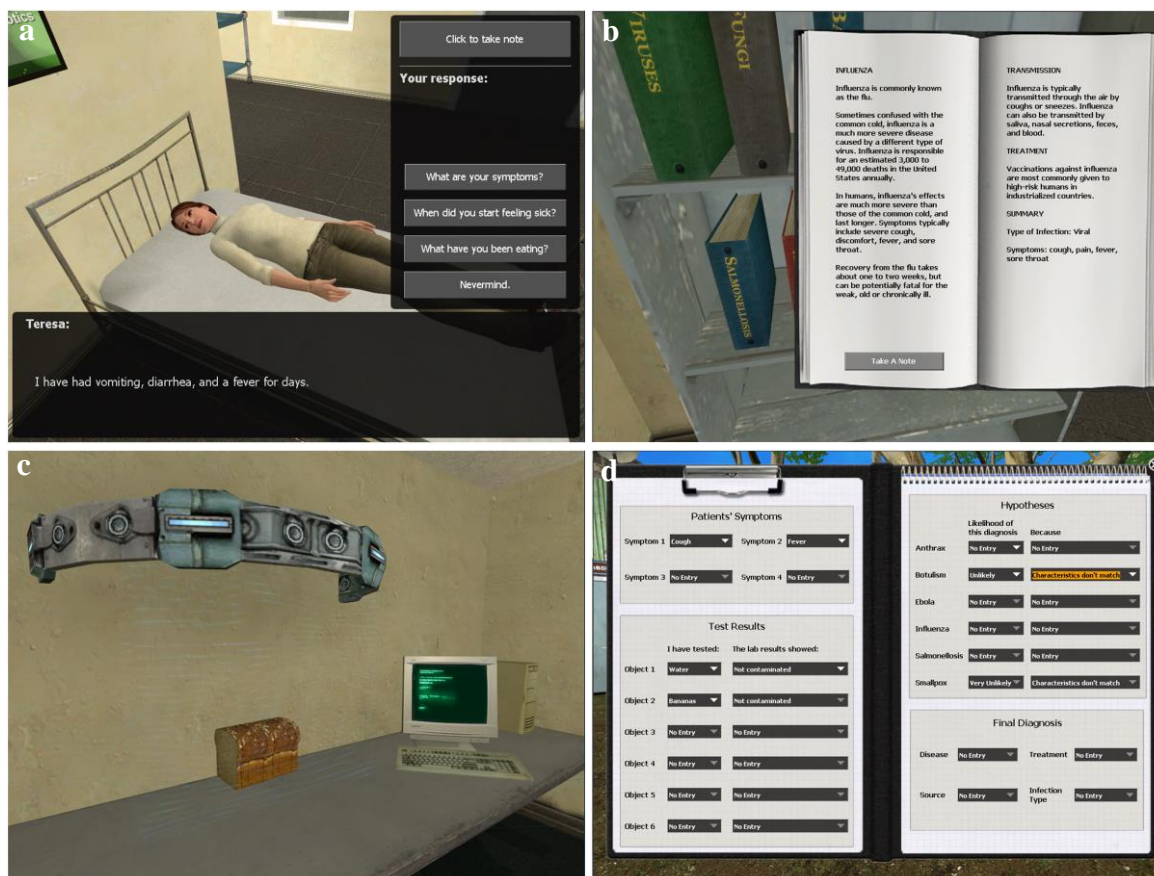
Differential sequence mining is relatively novel in the area of educational data mining, but has seen great popularity since its introduction. It was originally applied to a teachable agent environment to explore differences in how successful students make connections to concept maps and utilize resources. It has since been applied to handwritten physics homework assignments as well as collaborative tabletop interactions (Herold, Zundel, &

Stahovich, 2013; Martinez-Maldonado, Yacef, & Kay, 2013). Recent work has also examined adding temporal context to identify how frequencies of patterns change over time (Kinnebrew, Mack, & Biswas, 2013). This approach is extremely versatile and provides easily interpretable and actionable results unlike many other data mining techniques. This likely explains the growing interest in applying this technique to a broad range of domains.

**Procedure.** The first step in applying the differential sequence mining technique to identify meaningful contingent behavior patterns was to transform the highly detailed trace logs from interactions with CRYSTAL ISLAND into a more abstract representation of the overall behaviors being performed. This involved removing irrelevant or uninteresting actions (e.g., entering buildings, or manipulating individual objects) and grouping together instances of similar behaviors (e.g., reading a book on influenza and then a book on ebola).

In total, four general action types were identified as important distinguishing behaviors: TALK, READ, TEST, and WORKSHEET (Figure 16). The first two actions represent the primary source of gathering data in the environment, while the second two represent the primary problem-solving and hypothesis testing tasks. These behaviors are central to the inquiry-based problem solving in CRYSTAL ISLAND. Additional details were also considered for each action and are described below:

- **TALK:** One of the primary ways students gather information is through talking with in-game characters. Students may talk with patients to learn about the symptoms of their illness (TALK<sub>SYM</sub>). There are also experts on pathogens, bacteria, and viruses that students may talk to (TALK<sub>PATH</sub>, TALK<sub>BAC</sub>, TALK<sub>VIR</sub>). Finally, some of the characters also describe the nature of the illness and how it spread to students and provide details about the specific problem-solving task (TALK<sub>PROB</sub>).
- **READ:** There are several books and posters scattered around the environment that students may use for additional information. Many of these resources cover the same topics as conversations with experts on the island (READ<sub>PATH</sub>, READ<sub>BAC</sub>, READ<sub>VIR</sub>). There is also a variety of books and posters that describe specific diseases (READ<sub>DIS</sub>).



**Figure 16.** Targeted behaviors (a) TALK, (b) READ, (c) TEST, (d) WORKSHEET

- TEST:** To identify contaminated items students must run tests on individual food items. They must also specify whether they are testing the item for a pathogen, mutagen or carcinogen. Based on the nature of the illness, students should rule out mutagen or carcinogen as possible sources and testing for this is considered irrelevant ( $TEST_{IRR}$ ). Tests for pathogens are identified as correct ( $TEST_{CORR}$ ) if the proper food item was selected and incorrect ( $TEST_{INC}$ ) otherwise.
- WORKSHEET (WS):** The diagnosis worksheet is where students keep notes about their findings and hypotheses. There are several sections of information that can be filled out. They can record symptoms of patients ( $WS_{SYM}$ ) and the results of their tests ( $WS_{TEST}$ ). They can also keep track of hypotheses ( $WS_{HYP}$ ) about individual dis-

eases and their reasoning. The final section of the worksheet ( $WS_{REP}$ ) is used to report their final conclusions to the nurse in order to complete the mystery.

Student logs contained an average of 58.7 ( $SD = 17.2$ ) actions or about 1 action per minute. From this patterns were created with 2-5 actions with a possible omission of one action to be included in the same pattern. Because of the sparsity of the data set an *s-support* threshold of 20% was selected. This resulted in 214 common patterns across all students.

The *i-support* metric was then calculated for each student. For each of the three components of SRL, the average frequency of each pattern was compared across High, Medium, and Low students. Specifically, t-tests with a Bonferroni correction were conducted to compare the *i-support* metrics across each pair of classifications within each component. The Bonferroni correction was conducted for 95% confidence across the three pairwise tests but did not account for the multiple comparisons across patterns. This approach was employed because the primary purpose of differential sequence mining is to identify meaningful patterns, not to prove statistical differences between populations, and most researchers do not use any such correction (Kinnebrew, Loretz, et al., 2013).

**Table 10.** Differential patterns for cognition

| Sample Sequences   |  | s-support |      |      | i-support |      |      |
|--|--|-----------|------|------|-----------|------|------|
|  |  | L         | M    | H    | L         | M    | H    |
| High SRL Students  | <b>P1: Reading about diseases and updating hypotheses in WS</b>                            |           |      |      |           |      |      |
|  | READ <sub>DIS</sub> -WS <sub>HYP</sub> -READ <sub>DIS</sub> -WS <sub>HYP</sub>             | 0.13      | 0.29 | 0.29 | 0.93      | 2.99 | 2.91 |
|  | WS <sub>HYP</sub> -READ <sub>DIS</sub> -WS <sub>HYP</sub> -READ <sub>DIS</sub>             | 0.10      | 0.28 | 0.31 | 0.70      | 2.51 | 2.45 |
|  | <b>P3: Learn about pathogens and problem then start investigating relevant information</b> |           |      |      |           |      |      |
|  | TALK <sub>PATH</sub> -TALK <sub>PROB</sub> -TALK <sub>SYM</sub> -TALK <sub>BAC</sub>       | 0.20      | 0.22 | 0.50 | 0.49      | 0.47 | 1.29 |
| TALK <sub>PATH</sub> -TALK <sub>PROB</sub> -TALK <sub>SYM</sub> -READ <sub>DIS</sub> | 0.20   | 0.27      | 0.46 | 0.33 | 0.49      | 0.67 |      |
| Low SRL Students   | <b>P4: Alternating incorrect and irrelevant tests</b>                                      |           |      |      |           |      |      |
|  | TEST <sub>IRR</sub> -TEST <sub>INC</sub> -TEST <sub>IRR</sub> -TEST <sub>INC</sub>         | 0.32      | 0.11 | 0.10 | 0.65      | 0.23 | 0.17 |
|  | TEST <sub>INC</sub> -TEST <sub>IRR</sub> -TEST <sub>IRR</sub>                              | 0.34      | 0.30 | 0.13 | 1.20      | 1.33 | 0.31 |
|  | <b>P8: Fill out diagnosis worksheet and check with nurse</b>                               |           |      |      |           |      |      |
|  | WS <sub>REP</sub> -TALK <sub>PROB</sub> -WS <sub>REP</sub>                                 | 0.19      | 0.07 | 0.03 | 0.33      | 0.07 | 0.03 |
| TALK <sub>PROB</sub> -WS <sub>REP</sub> -TALK <sub>PROB</sub>                        | 0.18   | 0.05      | 0.03 | 0.30 | 0.08      | 0.03 |      |

**Table 11.** Differential patterns for metacognition

| Sample Sequences   |  | s-support |      |      | i-support |      |      |
|--|--|-----------|------|------|-----------|------|------|
|  |  | L         | M    | H    | L         | M    | H    |
| High SRL Students  | <b>P1: Reading about diseases and updating hypotheses in WS</b>  |           |      |      |           |      |      |
|  | READ <sub>DIS</sub> -WS <sub>HYP</sub> -READ <sub>DIS</sub> -WS <sub>HYP</sub> -READ <sub>DIS</sub>      | 0.10      | 0.19 | 0.26 | 0.28      | 0.53 | 0.74 |
|  | WS <sub>HYP</sub> -READ <sub>DIS</sub> -WS <sub>HYP</sub> -READ <sub>DIS</sub> -WS <sub>HYP</sub>        | 0.11      | 0.18 | 0.24 | 0.31      | 0.54 | 0.70 |
|  | <b>P2: Talk about symptoms and update symptoms in WS</b>   |           |      |      |           |      |      |
|  | TALK <sub>SYM</sub> -WS <sub>SYM</sub>   | 0.42      | 0.61 | 0.61 | 0.74      | 1.16 | 1.13 |
|  | TALK <sub>SYM</sub> -TALK <sub>PROB</sub> -WS <sub>SYM</sub>   | 0.03      | 0.08 | 0.12 | 0.03      | 0.08 | 0.13 |
| Low SRL Students   | <b>P4: Alternating incorrect and irrelevant tests</b>  |           |      |      |           |      |      |
|  | TEST <sub>IRR</sub> -TEST <sub>INC</sub> -TEST <sub>IRR</sub>  | 0.61      | 0.55 | 0.47 | 1.79      | 1.55 | 1.01 |
|  | TEST <sub>INC</sub> -TEST <sub>IRR</sub>   | 0.71      | 0.71 | 0.66 | 2.27      | 2.04 | 1.50 |
|  | <b>P5: Read about diseases and ask about symptoms</b>  |           |      |      |           |      |      |
|  | TALK <sub>SYM</sub> -READ <sub>DIS</sub> -TALK <sub>SYM</sub> -READ <sub>DIS</sub>                       | 0.39      | 0.26 | 0.23 | 0.39      | 0.31 | 0.25 |
|  | READ <sub>DIS</sub> -TALK <sub>SYM</sub> -READ <sub>DIS</sub> -TALK <sub>SYM</sub> -TALK <sub>PROB</sub> | 0.35      | 0.19 | 0.18 | 0.35      | 0.22 | 0.19 |
|  | <b>P6: Learn relevant things and run irrelevant tests</b>  |           |      |      |           |      |      |
|  | TALK <sub>BAC</sub> -READ <sub>BAC</sub> -TEST <sub>IRR</sub>  | 0.14      | 0.04 | 0.05 | 0.20      | 0.15 | 0.10 |
| READ <sub>VIR</sub> -READ <sub>PATH</sub> -TEST <sub>IRR</sub> | 0.20   | 0.20      | 0.07 | 0.55 | 0.43      | 0.21 |      |

**Table 12.** Differential patterns for motivation

| Sample Sequences  |  | s-support |      |      | i-support |      |      |
|-------------------|--|-----------|------|------|-----------|------|------|
|                   |  | L         | M    | H    | L         | M    | H    |
| High SRL Students | <b>P1: Reading about diseases and updating hypotheses in WS</b>                |           |      |      |           |      |      |
|                   | READ <sub>DIS</sub> -WS <sub>HYP</sub> -READ <sub>DIS</sub> -WS <sub>HYP</sub> | 0.13      | 0.29 | 0.29 | 0.93      | 2.99 | 2.91 |
|                   | WS <sub>HYP</sub> -READ <sub>DIS</sub> -WS <sub>HYP</sub> -READ <sub>DIS</sub> | 0.10      | 0.28 | 0.31 | 0.70      | 2.51 | 2.45 |
| Low SRL Students  | <b>P6: Learn relevant things and run irrelevant tests</b>                      |           |      |      |           |      |      |
|                   | READ <sub>BAC</sub> -TALK <sub>PROB</sub> -TEST <sub>IRR</sub>                 | 0.42      | 0.29 | 0.25 | 1.14      | 0.62 | 0.51 |
|                   | TALK <sub>SYM</sub> -TALK <sub>PROB</sub> -TEST <sub>IRR</sub>                 | 0.23      | 0.17 | 0.08 | 0.48      | 0.33 | 0.13 |
|                   | <b>P7: Running irrelevant test and then going to learn</b>                     |           |      |      |           |      |      |
|                   | TEST <sub>IRR</sub> -TALK <sub>BAC</sub>                                       | 0.17      | 0.10 | 0.07 | 0.34      | 0.17 | 0.07 |
|                   | TEST <sub>IRR</sub> -TALK <sub>VIR</sub>                                       | 0.20      | 0.12 | 0.08 | 0.51      | 0.23 | 0.16 |

**Results.** Of the 214 most common sequence patterns, 39 occurred differentially across cognition class, 29 differed across metacognition class and 25 differed across motivation class. Further interpretation of these sequences suggested general behavior patterns that occurred at different frequencies between the groups (Table 10-12). Of these, three patterns were more frequently displayed by High SRL students, while the remaining five patterns were more frequent among Low SRL students. Some patterns were only recognized as differing significantly based on classification for certain categories. These general patterns of behavior provide important insight into how students differentially interact with the environment given their level of each component of SRL.

For instance, patterns **P1** and **P2** both highlight High SRL students' usage of the diagnosis worksheet. **P1** was demonstrated differentially across all three components of SRL, while **P2** was only differential across the metacognitive component. These patterns indicate that High SRL students in these categories are more likely to keep track of information as they receive it. Both the hypothesis and symptoms area of the diagnosis worksheet are optional, suggesting that High SRL students are choosing to use the resource to help themselves keep track of their ideas. Additionally, while the symptoms section of the worksheet involves simple recording of facts, the hypothesis area requires students to synthesize what they know and make inferences about the likelihood of different hypotheses, indicating strong inquiry skills. Together these patterns indicate that High SRL students are utilizing resources to keep track of what they know and are actively reflecting on the inquiry process.

In contrast, pattern **P5**, which is demonstrated more frequently by Low SRL students in the metacognition component, indicates poor planning and inquiry skills. This pattern involves students reading about diseases, then visiting patients to ask about their symptoms, and repeating this process. This pattern suggests that Low metacognition students are gathering data "just in time." They are repeatedly checking the information from patients against the information in books and posters to arrive at a hypothesis. These students are not keeping track of this information in their diagnosis worksheet and consequently are going back and forth between the books and posters on diseases to the infirmary with the patients. This rep-

resents a much less effective approach to problem solving when compared with the High metacognition students. Additionally, these students are likely experiencing an increased cognitive load as they are trying to recall all the details they have gathered without the aid of the in-game resources. These patterns indicate that Low metacognition students may need scaffolding for effective organization of knowledge and use of external cognitive tools, which is an important component of self-regulated learning (Land, 2000; Zimmerman, 1990).

Another important distinction concerns students making connections about the type of illness affecting the patients. Specifically, students learn that the illness was spread through food that the camp members ate (TALKPROB). Students should also learn (through TALKPATH or READPATH) that a pathogen is a type of illness that can be spread through food or contact, whereas mutagens and carcinogens are not spread from person to person. Students should consequently conclude that the illness is a form of pathogen. This may be what is occurring in pattern **P3** demonstrated by High cognitive SRL students. These students ask about the nature of the illness and about pathogens and are going on to further investigation behaviors. They are likely using this information to draw the conclusion that the illness is a form of pathogen and moving on to find out more about other important information such as symptoms or possible diseases. It is important to note that making this connection is critical for successful testing behavior.

When running tests in the lab, students select whether they are testing for pathogens, mutagens or carcinogens. Knowledge of the pathogens and the nature of the illness should preclude students from running tests on carcinogens or mutagens (TESTIRR); however, pattern **P6** indicates that Low SRL students in the metacognitive and motivational categories are not making this connection or choose to ignore it. Since we see this pattern applying to these groups and not differentiated by the cognitive component it suggests that it may be poor regulation of knowledge and application of what has already be learned, or it may also simply be evidence of carelessness or disinterest. Furthermore, **P4** suggests that Low cognitive and metacognitive SRL students may not be carefully selecting their testing strategy based on prior knowledge and may be trying any form of test to get a positive result. This suggests that

Low SRL students may need more guidance in making the connection between the nature of the problem and type of illness. Additionally, they should be encouraged to identify whether the source is a pathogen, mutagen, or carcinogen before beginning to test. Finally, pattern **P7** indicates that Low motivation SRL students are getting negative tests results after testing something other pathogens and then go to learn about the information necessary to correctly get a result. Perhaps these students hoped they could get by without gathering the necessary information first and have found that they are unable to just “game the system.” They then proceed to learn the material they should have tackled before running any tests.

In some ways, pattern **P8** also indicates poor problem solving behaviors and is more commonly demonstrated by Low cognitive SRL students. Specifically this pattern consists of students checking their worksheet, finding something is wrong with it, making a change and checking again. They are not doing any investigation behaviors in between such as re-reading about diseases or double-checking the symptoms. This may indicate that they are trying to game the system by repeated guess and check. This is consistent with many of the traits that were used to create the initial classification of cognitive SRL.

**Feature Creation.** The objective of the differential sequence mining was two-fold, first to understand patterns of behavior that are indicative of self-regulated learning, but also lead to informed creation of features that represent the contingency and patterned contingency event evidence of SRL. These features are expected to increase the predictive accuracy of the runtime SRL assessment models. Using the paradigm put forth by Winne (2010) we first specified the IF and THEN actions and sequences that exemplified the pattern (Table 13). For example, pattern **P3** is described as learning about the problem and pathogens then starting to investigate relevant information. The IF for this pattern is an action of TALKPATH and TALKPROB occurring back to back in either order. The THEN for this pattern are either of the two actions commonly occurring in this pattern READDIS or TALKSYM. Consequently, the pattern is only exhibited in cases where the IF is immediately followed by the THEN.

**Table 13.** Identified patterns of behavior

|                   | Pattern |   | Models |      |      | Feature Creation   |  |
|-------------------|---------|---|--------|------|------|--|--|
|                   | ID      | Description   | COG    | META | MOTI | IF   | THEN   |
| High SRL Students | P1      | Reading about diseases and updating hypotheses in worksheet                         | X      | X    | X    | WS <sub>HYP</sub> , READ <sub>DIS</sub>  | WS <sub>HYP</sub> , READ <sub>DIS</sub>  |
|                   | P2      | Reading about symptoms and updating symptoms in worksheet                           |        | X    |      | WS <sub>SYM</sub> , TALK <sub>SYM</sub>  | WS <sub>SYM</sub> , TALK <sub>SYM</sub>  |
|                   | P3      | Learn about the problem and pathogens then start investigating relevant information | X      |      |      | [TALK <sub>PATH</sub> , TALK <sub>PROB</sub> ]*2   | READ <sub>DIS</sub> , TALK <sub>SYM</sub>  |
| Low SRL Students  | P4      | Alternating between incorrect and irrelevant tests                                  | X      | X    |      | TEST <sub>IRR</sub> , TEST <sub>INC</sub>  | TEST <sub>IRR</sub> , TEST <sub>INC</sub>  |
|                   | P5      | Reading about diseases and asking about symptoms                                    |        | X    |      | READ <sub>DIS</sub> , TALK <sub>SYM</sub>  | READ <sub>DIS</sub> , TALK <sub>SYM</sub>  |
|                   | P6      | Learning relevant things and running irrelevant tests                               |        | X    | X    | TALK <sub>PROB</sub> , TALK <sub>BAC</sub> , TALK <sub>VIR</sub> , READ <sub>BAC</sub> , READ <sub>VIR</sub> | TEST <sub>IRR</sub>  |
|                   | P7      | Running irrelevant tests and going to learn   |        |      | X    | TEST <sub>IRR</sub>  | TALK <sub>PROB</sub> , TALK <sub>BAC</sub> , TALK <sub>VIR</sub> , READ <sub>BAC</sub> , READ <sub>VIR</sub> |
|                   | P8      | Filling out diagnosis worksheet and going to check it with the nurse                | X      |      |      | WS <sub>REP</sub> , TALK <sub>PROB</sub>   | WS <sub>REP</sub> , TALK <sub>PROB</sub>   |

Three metrics were then calculated for each student for each possible pattern: the frequency of the IF sequence, the frequency of the IF-THEN sequence, the proportion of IF-THEN sequences to IF sequences. To illustrate, consider a student who exhibited the IF pattern of **P3** (talking about pathogens and the problem) a total of three distinct times during his interaction with CRYSTAL ISLAND. He would have an IF frequency of 3. Suppose that he exhibited the entire IF-THEN sequence (talking about pathogens and the problem, then immediately reading about diseases) twice during the interaction. His IF-THEN frequency would be 2. Consequently, he demonstrated the THEN behavior in response to the IF a total of 66% of the time, which is calculated by the frequency of IF sequences over the frequency of IF-THEN sequences.

Using these calculations we then represent the contingency and patterned contingency event types. Contingency, which is described as a student completing an action in direct response to relevant stimulus is quantified by the IF-THEN frequency. Patterned contingency which is described as the proportion of times this response followed the stimulus is quantified by the proportion of IF-THEN over IF frequencies. Together with the occurrence level features identified during the feature selection process all three event-based evidences described by Winne (2010) are represented.

### **6.3 Discussion**

To aid in the development of real-time assessment models it was necessary to establish a set of features that were both meaningful and predictive of SRL. This work used the framework described by Winne (2010) to guide feature selection and creation. Stepwise logistic regression was used to identify attribute and occurrence event features that offer the most power in predicting each category of SRL. Differential sequence mining was used to identify patterns of behavior that are more commonly exhibited by students of higher or lower SRL skill. These results were then used to guide the creation of contingency and patterned contingency features for the predictive models. This represents a novel approach for feature creation and

highlights the ways in which theoretical background knowledge can guide the use of empirical machine learning techniques.

These analyses have produced a feature set that is expected to improve real-time predictive models but also provides some insight into self-regulated learning in CRYSTAL ISLAND. Specifically, the patterns of behavior that were identified using the differential sequence mining approach show different strategies of resources use and highlight some differences in how students draw connections between different sets of materials. While this has been used for feature creation in this work, these patterns could also be used to guide further development in CRYSTAL ISLAND. Specifically, these patterns could inform adaptive scaffolding to be given in response when the predictive models identify students with low SRL skills.

## CHAPTER 7

### Predictive Models of Self-Regulated Learning

The ultimate goal of the line of investigation reported on in this dissertation is to build real-time models for assessing self-regulated learning. The primary approach has been to incorporate theoretical background knowledge at every step to improve the predictive power of these machines. Here we aim to test the hypotheses that machine learning approaches guided by theoretical information will improve real-time predictive models of self-regulated learning. The first hypothesis (**H1-b**) is that the features created through differential sequence mining will improve the predictive accuracy of the models. This will be tested by comparing models trained on data sets with and without these features. The second hypothesis (**H2**) is that representing the key processes of SRL and the cyclic relationship between them will improve predictive accuracy of the models. This will be tested by comparing (a) naïve Bayesian models with no process representation, (b) static Bayesian models with process representation but no cyclic structure, and (c) dynamic Bayesian models that represent the processes and cyclic relationship.

#### 7.1 Bayesian Approach

A Bayesian modeling approach was selected for the real-time predictive modeling of SRL primarily because Bayesian networks can accommodate both empirical and theoretical knowledge (Conati & Maclaren, 2009; Sabourin, Mott, & Lester, 2011). Bayesian networks operate by representing the relationship between variables in terms of a probability distribution. Bayesian networks involve two main components, (1) a network structure, which describes which variables are related to others, and (2) a set of conditional dependencies which provide the exact specifications for these relationships. Both the structure and the conditional dependencies can be learned using a variety of possible algorithms (Alpaydin, 2004) or specified by hand.

Bayesian methodologies have been used to represent a wide variety of phenomena in intelligent tutoring systems including models of learning (Baker, Corbett, & Aleven, 2008; Corbett & Anderson, 1994), affect (Conati & Maclaren, 2009; Sabourin, Mott, et al., 2011), and hinting (Gertner, Conati, & VanLehn, 1998). They have also been successful in stealth assessment in game-based environments (Shute, 2011).

As part of the proposed framework, we explore how Bayesian approaches can be used to represent important processes of SRL. For this reason we explore naïve Bayesian networks which offer no representation of these phenomena, static Bayesian networks which represent the components of the process but in a static way, and dynamic Bayesian networks which incorporate time to represent the cyclical nature of SRL processes. For each model we also explore whether the addition of event features created through differential sequence mining offers additional predictive power.

**Naïve Approach.** A naïve Bayesian network operates under the “naïve” assumption that all variables are directly related to the outcome variable but are conditionally independent of each other (Alpaydin, 2004). There are no hidden variables or detailed relationships to represent complex phenomena or incorporate theoretical grounding. In this way, naïve Bayesian networks represent a baseline measure to compare future models against in order to identify the benefit of theoretical grounding of models.

For this modeling task we consider three feature sets: **All Features**, **Feature Selection**, and **Feature Selection + Event Feature Creation**. The **All Features** set consists of all 55 features before feature selection was applied. The **Feature Selection** set uses only the attribute and occurrence features identified during feature selection. Finally, the **Feature Selection + Event Feature Creation** set adds the contingency and patterned contingency features that were created through the differential sequence mining approach. Here we consider the **All Features** set to identify any possible gain or loss in predictive power by applying feature selection. Separate models are learned for the cognitive, metacognitive and motivational components of SRL using the features that were selected and created specifically for that

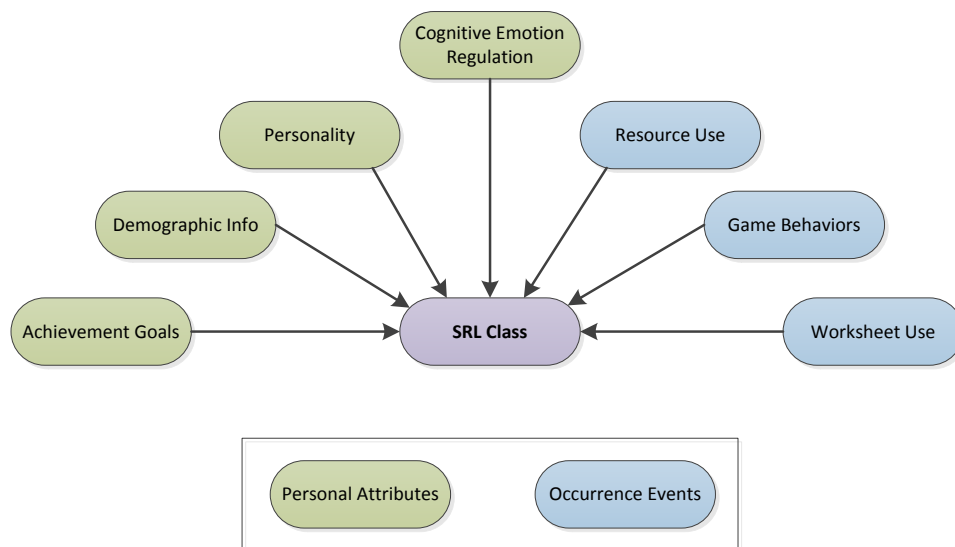
component. All naïve models were trained and evaluated in the WEKA machine learning toolkit using 10-fold cross validation.

Following the assumptions of naïve Bayesian networks, all features relate directly to the outcome variable. Therefore, for models learned with all features and selected features, each of the personal attributes and event features are directly used to predict the SRL class of interest (Figure 17). Similarly, the naïve models learned with the addition of the created contingency features have the same structure with the addition of the extra features (Figure18).

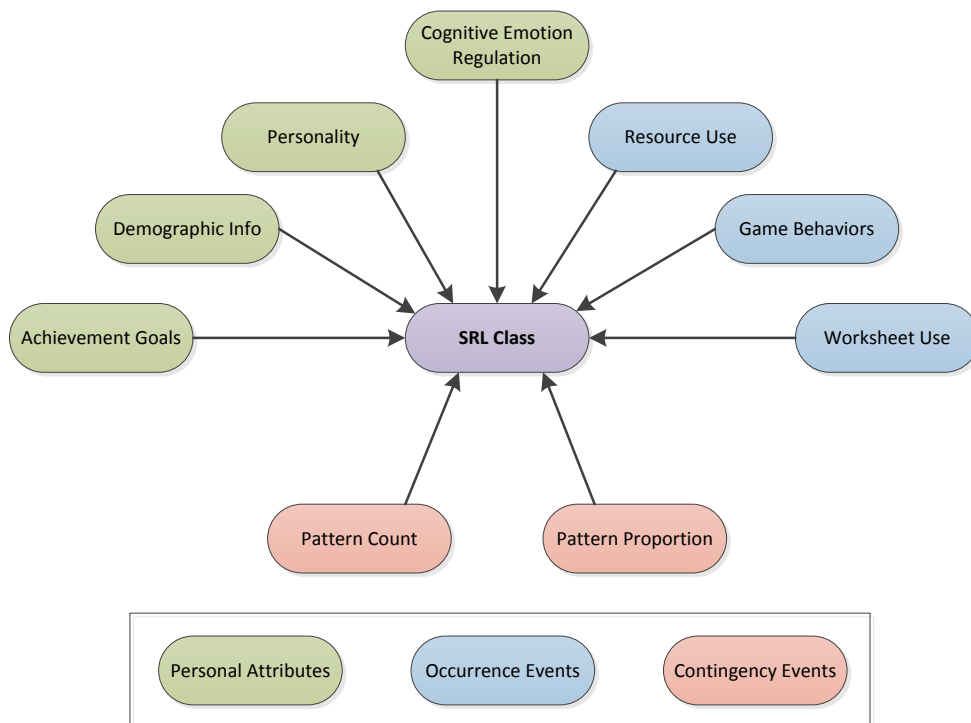
**Static Approach.** The static Bayesian network offers advantages over the naïve Bayesian network by allowing representation of conditional dependencies between variables and through the introduction of hidden variables that can represent more complex phenomena than each feature independently. These relationships are specified through the structure of the network which can be learned or hand-crafted. In this work the structure is hand-crafted to allow for the three key theoretical processes of self-regulated learning: forethought, performance, and reflection.

For identifying the structure of the static Bayesian network it was important to utilize a procedure that could be applied equally across each of the three components of SRL and could be similarly adopted by other researchers. This is critical in ensuring that the approach is not overly biased by the data and can be validated in other domains. Specifically, we wished to avoid the bit-by-bit tweaking of models that often offers gains in predictive accuracy but is more prone to overfitting.

Following this approach we have the three hidden states related to the processes of self-regulated learning: forethought, performance, and reflection. Only these hidden variables have a direct conditional relationship with the outcome class of interest. Each of the observable features is then tied to one or more of these processes. For the static Bayesian modeling we consider only the **Feature Selection** and **Feature Selection + Event Feature Creation** sets. The **All Features** set is no longer considered because of issues of model complexity and sparsity.



**Figure 17.** Structure of naïve Bayesian network using *All Features* and *Feature Selection*



**Figure 18.** Structure of naïve Bayesian network using *Feature Selection + Event Feature Creation*

**Table 14.** Personal features to SRL processes

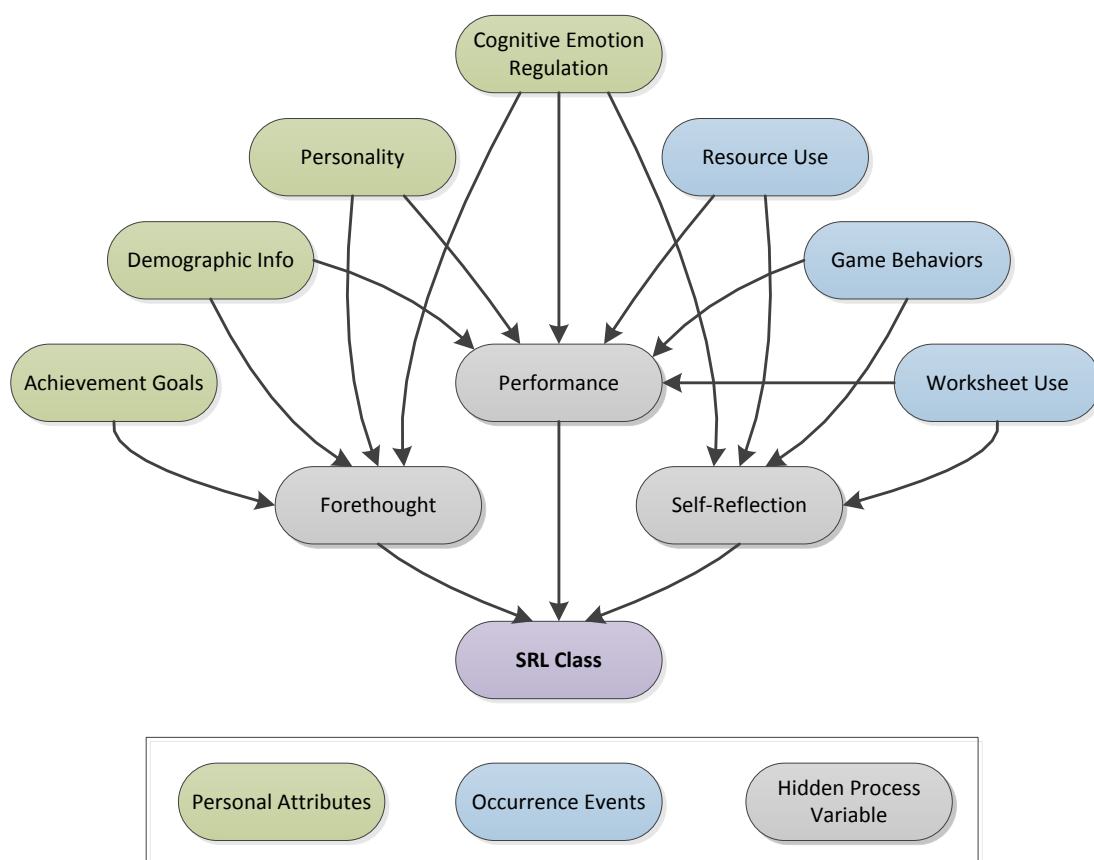
|   |                       | COG | META | MOTI | Fore-<br>thought | Perfor-<br>mance | Self<br>Reflection |
|---|-----------------------|-----|------|------|------------------|------------------|--------------------|
| Achievement Goals<br>Questionnaire              | Mastery Avoidance     |     | X    |      | X                |                  |                    |
|   | Mastery Approach      |     |      | X    | X                |                  |                    |
|   | Performance Avoidance |     | X    |      | X                |                  |                    |
|   | Performance Approach  | X   | X    | X    | X                |                  |                    |
| Big Five In-<br>ventory                         | Openness              | X   | X    |      | X                |                  |                    |
|   | Conscientiousness     |     | X    | X    | X                | X                |                    |
|   | Extraversion          | X   | X    | X    | X                |                  |                    |
|   | Agreeableness         |     | X    | X    | X                |                  |                    |
|   | Neuroticism           |     | X    |      | X                | X                |                    |
| Cognitive Emotion Reg-<br>ulation Questionnaire | Self-Blame            | X   |      | X    |                  |                  | X                  |
|   | Acceptance            |     | X    |      |                  |                  | X                  |
|   | Planning              |     | X    |      | X                |                  | X                  |
|   | Positive Refocusing   |     |      | X    |                  | X                | X                  |
|   | Positive Reappraisal  |     | X    | X    |                  |                  | X                  |
|   | Perspective           |     | X    |      |                  |                  | X                  |
|   | Catastrophizing       | X   | X    | X    |                  |                  | X                  |
|   | Other-Blame           | X   |      | X    |                  |                  | X                  |
| Demo-<br>graphic                                | Pre Test              | X   | X    | X    |                  | X                |                    |
|   | Gaming Frequency      |     | X    |      | X                | X                |                    |
|   | Gaming Hours          |     |      | X    | X                | X                |                    |

For the models created using the **Feature Selection** set we first had to identify the relationship between each feature and the three processes of forethought, performance, and reflection. This was done primarily through the explanations of the sub-processes and related phenomena described by Zimmerman (2000). In general, the static personal attributes were mostly related to the processes of forethought and self-reflection (Table 14). Specifically, goal orientation and personality were expected to be most related to goal-setting, self-efficacy,

**Table 15.** Occurrence features to SRL processes

|                |                                 | COG | META | MOTI | Fore-thought | Perfor-mance | Self Reflection |
|----------------|---------------------------------|-----|------|------|--------------|--------------|-----------------|
| Resource Use   | Posters                         | X   |      | X    |              | X            |                 |
|                | Books                           | X   |      |      |              | X            |                 |
|                | MicroApp                        | X   | X    | X    |              | X            |                 |
|                | Notes                           | X   |      |      |              | X            | X               |
|                | Tests Run                       |     | X    |      |              | X            | X               |
| Worksheet Use  | Worksheet Updates               |     |      | X    |              | X            | X               |
|                | Worksheet Access                |     | X    |      |              | X            | X               |
|                | Worksheet Filled                |     |      | X    |              | X            | X               |
|                | Worksheet Right                 |     |      | X    |              | X            | X               |
|                | Worksheet Checks                | X   |      | X    |              |              | X               |
| Game Behaviors | Average Status Length           |     | X    |      |              | X            | X               |
|                | Off Task Behavior               |     | X    | X    |              | X            |                 |
|                | Total Goals Completed           | X   | X    |      |              | X            |                 |
|                | Time Since Last Goal Completion | X   | X    |      |              | X            |                 |

and general expectations. Emotion regulation features on the other hand were expected to be more related to self-reflection processes such as causal attribution, satisfaction, and adaptation. The event features (Table 15) were all considered to be tied to performance as these are direct measures of students' actions, strategies, and attention. Furthermore, some resource use is tied to self-reflection because it represents opportunities for thinking about current progress and knowledge. Based on these classifications, the static Bayesian model was constructed for the **Feature Selection** set. Each feature is related to the specified hidden processes, which then in turn predict the SRL classification (Figure 19). Again, separate models were constructed for the cognitive, metacognitive and motivational components of SRL using only the features identified during the feature selection process. However, the overall structure (Figure 19) was constant across the three models.



**Figure 19.** Structure of static Bayesian network using *Feature Selection*

For the models created using the and **Feature Selection + Event Feature Creation** set, the same feature-process relationships were kept constant from the **Feature Selection** models. The addition here is in the two features for each pattern identified during the differential sequence mining approach. For each pattern an additional hidden variable is created to represent the total prevalence of that pattern with only the two contingency and patterned contingency features directly relating to it (Figure 20). This pattern is then attributed to one or more of the processes of SRL using a similar approach to that used with the attribute and occurrence event features (Table 16). The patterns were mostly related to performance and reflection like the occurrence event features, but there were some patterns that were also tied forethought.



**Figure 20.** Structure of static Bayesian network using *Feature Selection + Event Feature Creation*

**Table 16.** Created contingency event features to SRL processes

|                            |                                      | COG | META | MOTI | Fore-<br>thought | Perfor-<br>mance | Self<br>Reflection |
|----------------------------|--------------------------------------|-----|------|------|------------------|------------------|--------------------|
| Contingent Event Variables | P1: Use of work-<br>sheet hypotheses | X   | X    | X    |                  | X                | X                  |
|                            | P2: Use of work-<br>sheet symptoms   | X   |      |      |                  | X                | X                  |
|                            | P3: Immediate inves-<br>tigation     |     | X    |      | X                | X                |                    |
|                            | P4: Bad testing be-<br>haviors       | X   | X    |      |                  | X                | X                  |
|                            | P5: Connections<br>without tools     | X   |      |      |                  | X                | X                  |
|                            | P6: Failure to make<br>connections   | X   |      | X    |                  | X                |                    |
|                            | P7: Learning after<br>failed testing |     |      | X    | X                |                  | X                  |
|                            | P8: Repeated work-<br>sheet checks   |     | X    |      |                  |                  | X                  |

The structure for each network was specified using the GeNIe modeling environment developed by the Decision Systems Laboratory of the University of Pittsburgh (<http://dsl.sis.pitt.edu>) following the approach described above. This represents the theoretical grounding of the models. The exact values of the conditional dependencies are then learned using the Expectation-Maximization (EM) algorithm. Models were evaluated using 10-fold cross validation.

**Dynamic Approach.** Dynamic Bayesian Networks (DBNs) extend static Bayesian networks by accounting for temporal relationships between variables. Specifically this is done by introducing conditional dependencies between time-slices with observations at time  $t_n$  informing observations at time  $t_{n+1}$ . This temporal relationship is expected to offer significant benefit in modeling SRL behaviors, which are very dynamic and have an important temporal component. The cyclical relationship of the forethought, performance, and self-reflection processes is a key underpinning of most theoretical models of SRL. Specifically, the out-

comes of a student's forethought impact their performance, which in turn impacts the outcomes of self-reflection. The cycle continues with self-reflection guiding future forethought.

To represent this cycle, three temporal relationships were added to the static Bayesian networks described above:

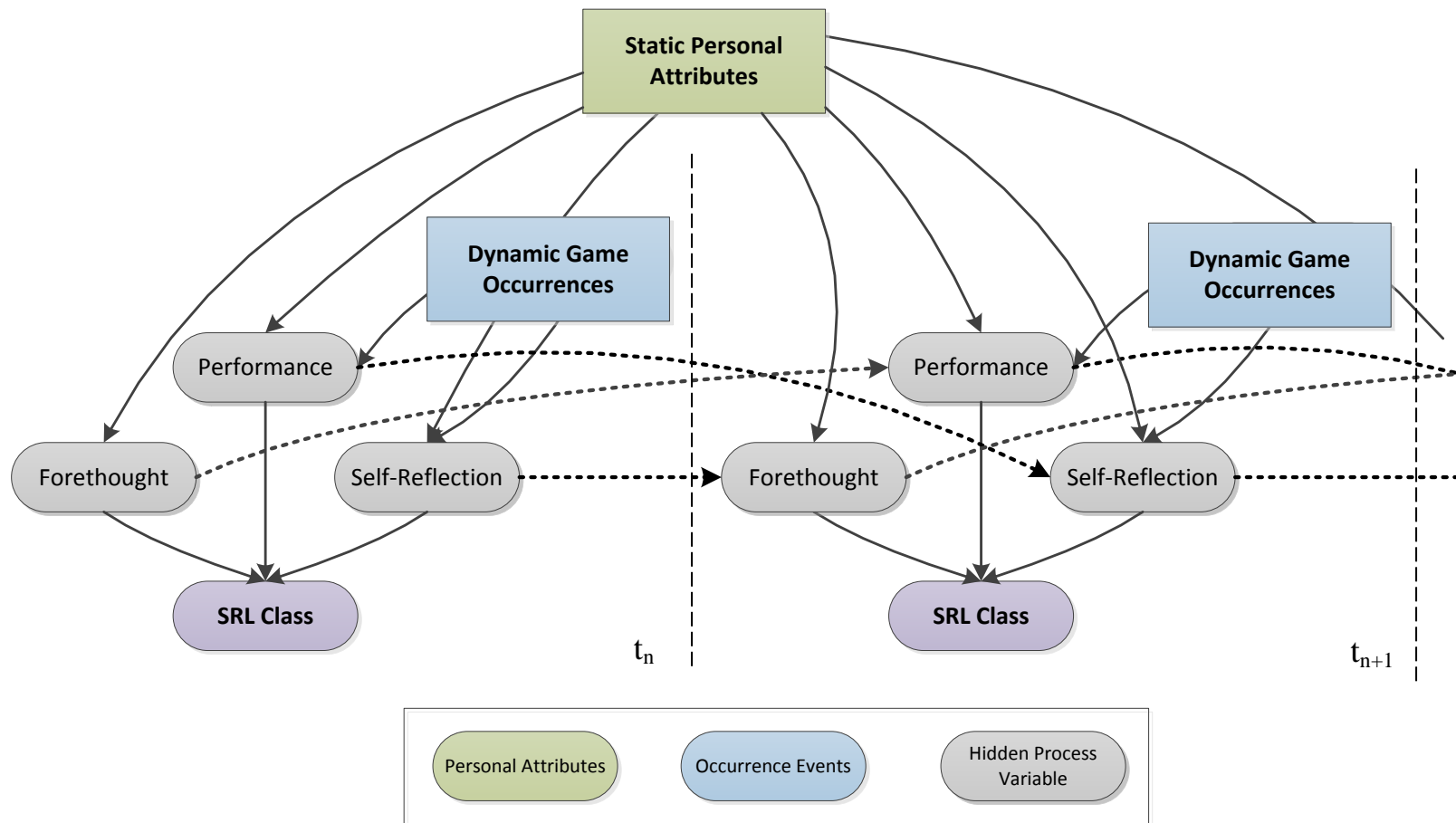
- $\text{Forethought}_n \rightarrow \text{Performance}_{n+1}$
- $\text{Performance}_n \rightarrow \text{Self-Reflection}_{n+1}$
- $\text{Self-Reflection}_n \rightarrow \text{Forethought}_{n+1}$

These temporal relationships were added to extend the Bayesian models built using both the **Feature Selection** (Figure 21) and **Feature Selection + Event Feature Creation** (Figure 22) data sets. As with the static models, the structure was specified in the GeNIe modeling environment. Conditional dependencies were learned using the EM algorithm and models were evaluated using 10-fold cross validation.

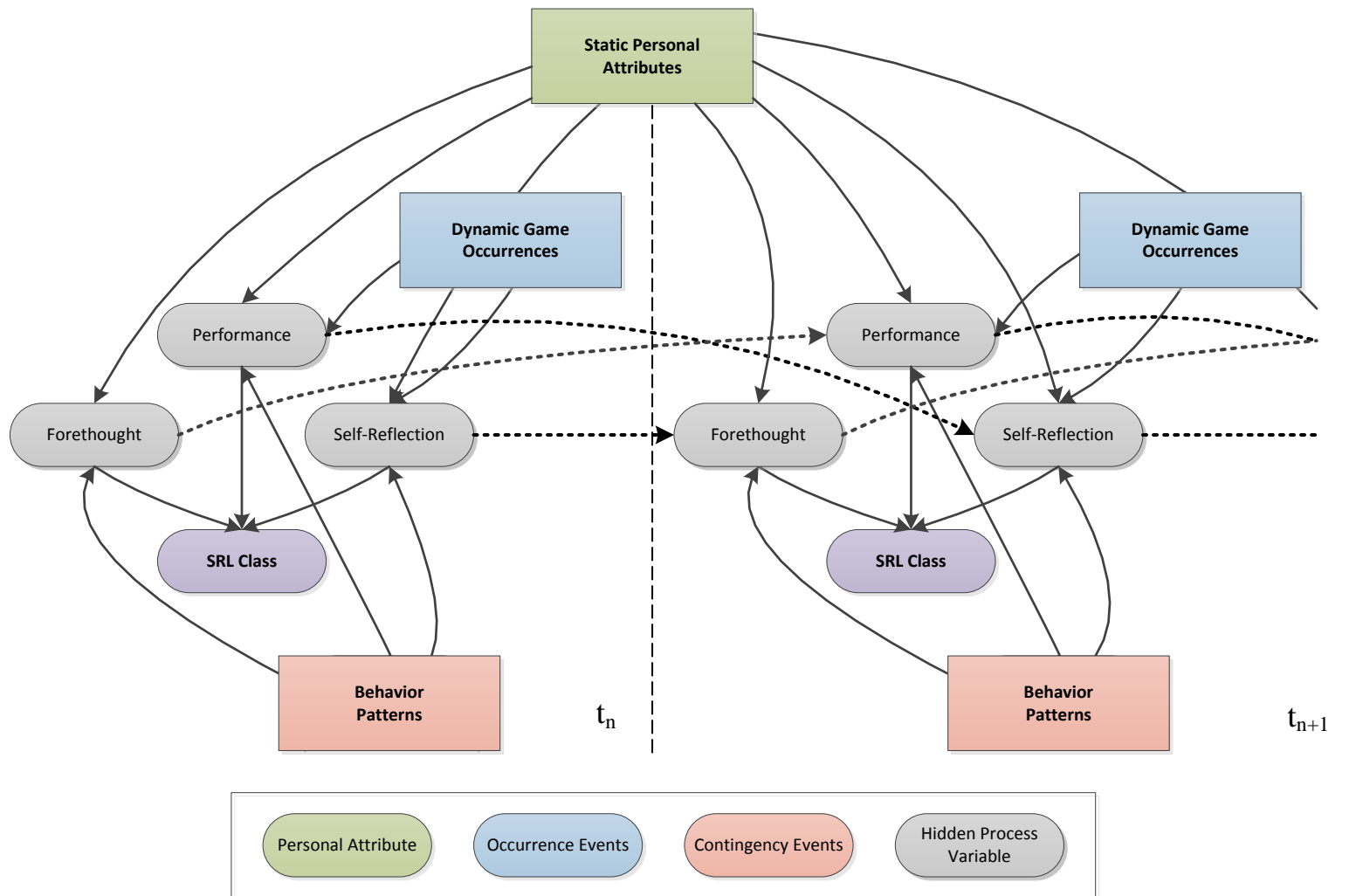
## 7.2 Results

One of the objectives of real-time assessment of SRL is to identify students' SRL tendencies early enough into the interaction to provide adaptive scaffolding. For this reason the predictive accuracy of each of the learned models was evaluated at different points in time throughout the interaction. Specifically, each model was evaluated at the time of the self-reports, or every seven minutes. Additionally, models were evaluated at a time before the interaction began using only personal features. This results in nine total time points for evaluation, Initial and  $T_{1-8}$

To evaluate the predictive power of each learned model we use 10-fold cross validation and report three metrics: accuracy, kappa, and weighted kappa. The predictive accuracy tells us the percent of students who were correctly classified. This measure is traditionally compared to a baseline measure of most frequent class to identify how much more predictive power is provided by the mode. The kappa metric uses agreement levels to indicate how much better than chance the model is performing. Using this metric, positive values ap-



**Figure 21.** Structure of dynamic Bayesian network using *Feature Selection*



**Figure 22.** Structure of dynamic Bayesian network using *Feature Selection + Event Creation*

proaching one indicate performance that is greater than chance, while values less than zero indicate performance that is worse than chance levels. Weighted kappa measures extend this metric by differentially weighting the cost of incorrect classifications. This approach is commonly used when dealing with ordered classifications, as is the case in this work. A distance from diagonal approach is used to apply more cost to increasingly dissimilar classifications. For example, classifying a Low SRL student as High SRL is more costly than a misclassification of Medium SRL.

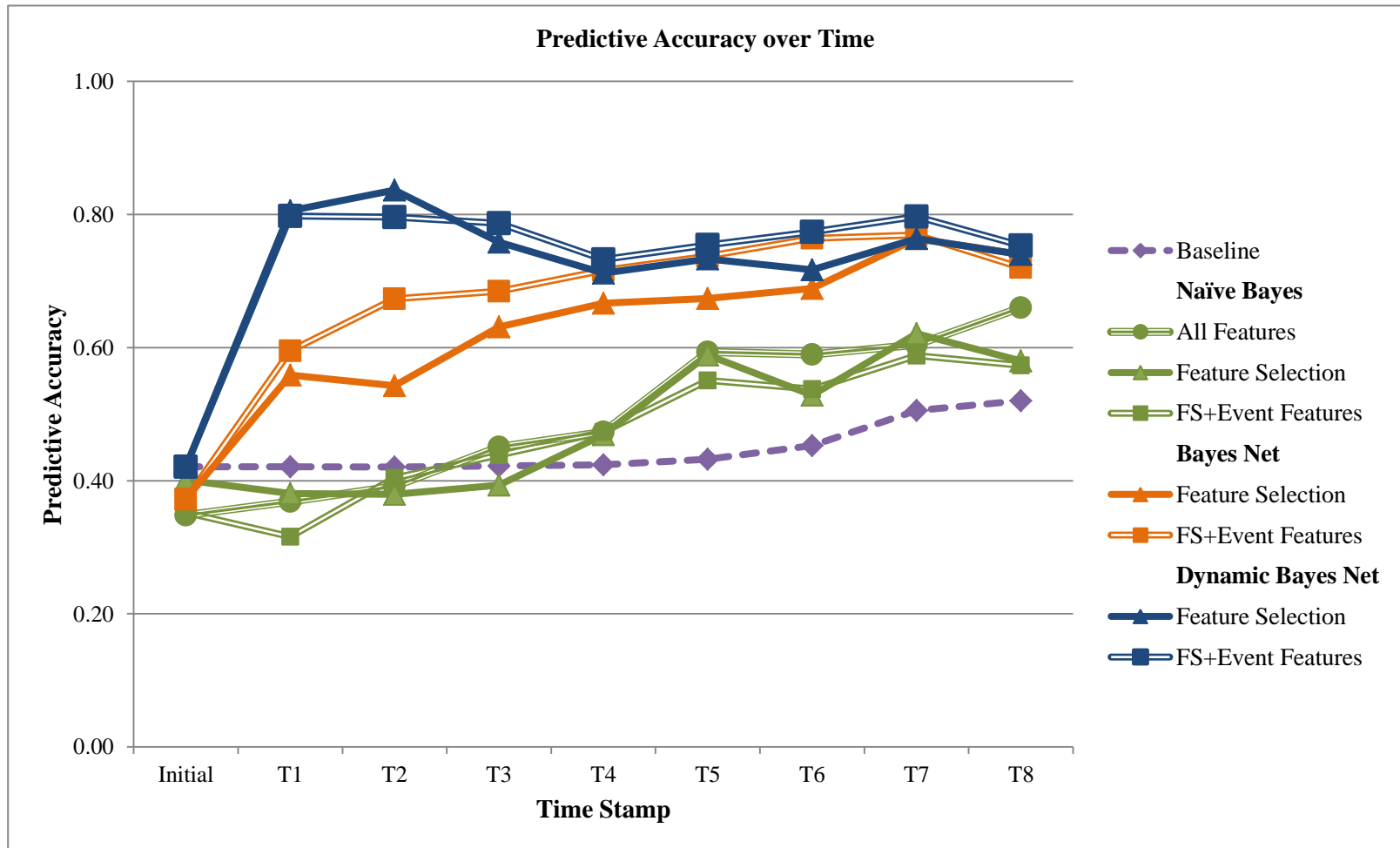
**Assessing Cognitive SRL.** For the cognitive component of SRL the baseline of most frequent class (Low) starts at 42% at the start of the interaction and progresses to 52% by the T<sub>8</sub>. This is because the higher students are finishing at a higher rate through the interaction so the proportion of Low cognitive SRL students increases. Averaged across all points in time, the baseline measure was 45%. Table 17 lists the performance measures for each trained model, while Figure 23 charts the predictive accuracy across time.

Overall, the best performing naïve model was trained on the **All Features** dataset, though this model did not offer significant ( $p = 0.41$ ) performance over the other models. This model had an overall predictive accuracy of 50%, with a standard kappa and weighted kappa of 0.22 indicating a slight improvement over baseline. Each of the models started with performance below baseline at the Initial time slice, but improved steadily with time, exceeding the baseline at TS<sub>4</sub>. ANOVAs indicated that in general, predictive accuracy continued steadily over time with statistically significant ( $F_{(8,261)} = 29.91$ ,  $p < 0.001$ ) differences between various points in time (Initial, T<sub>1</sub> < T<sub>2-3</sub> < T<sub>4</sub> < T<sub>5</sub> < T<sub>6-8</sub>). Overall the naïve models offered slight improvement over baseline with steadily increasing performance over time. There was no difference in performance based on data set.

When evaluating the static Bayesian networks, the model trained on the **Feature Selection + Event Feature Creation** data set offered the best performance (acc = 67%,  $\kappa = 0.48$ , weighted-  $\kappa = 0.49$ ) with significantly greater accuracy than the model trained with the

**Table 17.** Predictive model performance for cognition

|                   | Initial                                    | T <sub>1</sub> | T <sub>2</sub> | T <sub>3</sub> | T <sub>4</sub> | T <sub>5</sub> | T <sub>6</sub> | T <sub>7</sub> | T <sub>8</sub> | Overall |      |
|-------------------|--|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|---------|------|
| Baseline          | 0.42                                       | 0.42           | 0.42           | 0.42           | 0.42           | 0.43           | 0.45           | 0.51           | 0.52           | 0.45    |      |
| Naïve Bayes       | All Features                               |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.35           | 0.37           | 0.39           | 0.45           | 0.47           | 0.59           | 0.59           | 0.60           | 0.66    | 0.50 |
|                   | Kappa                                      | 0.01           | 0.04           | 0.06           | 0.16           | 0.19           | 0.38           | 0.37           | 0.36           | 0.44    | 0.22 |
|                   | W. K.                                      | 0.02           | 0.05           | 0.07           | 0.17           | 0.20           | 0.33           | 0.36           | 0.32           | 0.43    | 0.22 |
|                   | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.40           | 0.38           | 0.38           | 0.39           | 0.47           | 0.59           | 0.53           | 0.62           | 0.58    | 0.48 |
|                   | Kappa                                      | 0.07           | 0.04           | 0.05           | 0.07           | 0.19           | 0.37           | 0.27           | 0.38           | 0.30    | 0.19 |
|                   | W. K.                                      | 0.09           | 0.07           | 0.03           | 0.07           | 0.21           | 0.39           | 0.29           | 0.39           | 0.33    | 0.21 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.35           | 0.32           | 0.40           | 0.44           | 0.47           | 0.55           | 0.54           | 0.59           | 0.57    | 0.47 |
| Kappa             | 0.01                                       | -0.06          | 0.09           | 0.14           | 0.18           | 0.31           | 0.28           | 0.33           | 0.28           | 0.17    |      |
| W. K.             | 0.01                                       | -0.02          | 0.10           | 0.13           | 0.21           | 0.31           | 0.29           | 0.35           | 0.31           | 0.19    |      |
| Bayes Net         | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.37           | 0.56           | 0.54           | 0.63           | 0.67           | 0.67           | 0.69           | 0.76           | 0.74    | 0.63 |
|                   | Kappa                                      | 0.04           | 0.31           | 0.28           | 0.42           | 0.47           | 0.48           | 0.49           | 0.60           | 0.56    | 0.41 |
|                   | W. K.                                      | 0.01           | 0.32           | 0.30           | 0.45           | 0.51           | 0.54           | 0.54           | 0.62           | 0.57    | 0.43 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.37           | 0.60           | 0.67           | 0.68           | 0.72           | 0.74           | 0.76           | 0.77           | 0.72    | 0.67 |
| Kappa             | 0.04                                       | 0.37           | 0.50           | 0.51           | 0.56           | 0.59           | 0.63           | 0.62           | 0.53           | 0.48    |      |
| W. K.             | 0.01                                       | 0.32           | 0.53           | 0.55           | 0.60           | 0.62           | 0.66           | 0.60           | 0.53           | 0.49    |      |
| Dynamic Bayes Net | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.42           | 0.81           | 0.84           | 0.76           | 0.71           | 0.73           | 0.72           | 0.76           | 0.74    | 0.72 |
|                   | Kappa                                      | 0.00           | 0.70           | 0.74           | 0.61           | 0.53           | 0.56           | 0.52           | 0.58           | 0.52    | 0.53 |
|                   | W. K.                                      | 0.00           | 0.68           | 0.74           | 0.63           | 0.55           | 0.58           | 0.53           | 0.56           | 0.53    | 0.53 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.42           | 0.80           | 0.80           | 0.79           | 0.73           | 0.75           | 0.77           | 0.80           | 0.75    | 0.73 |
| Kappa             | 0.00                                       | 0.69           | 0.68           | 0.66           | 0.57           | 0.60           | 0.62           | 0.64           | 0.55           | 0.56    |      |
| W. K.             | 0.00                                       | 0.67           | 0.68           | 0.67           | 0.59           | 0.60           | 0.64           | 0.63           | 0.56           | 0.56    |      |



**Figure 23.** Predictive accuracy of cognition over time

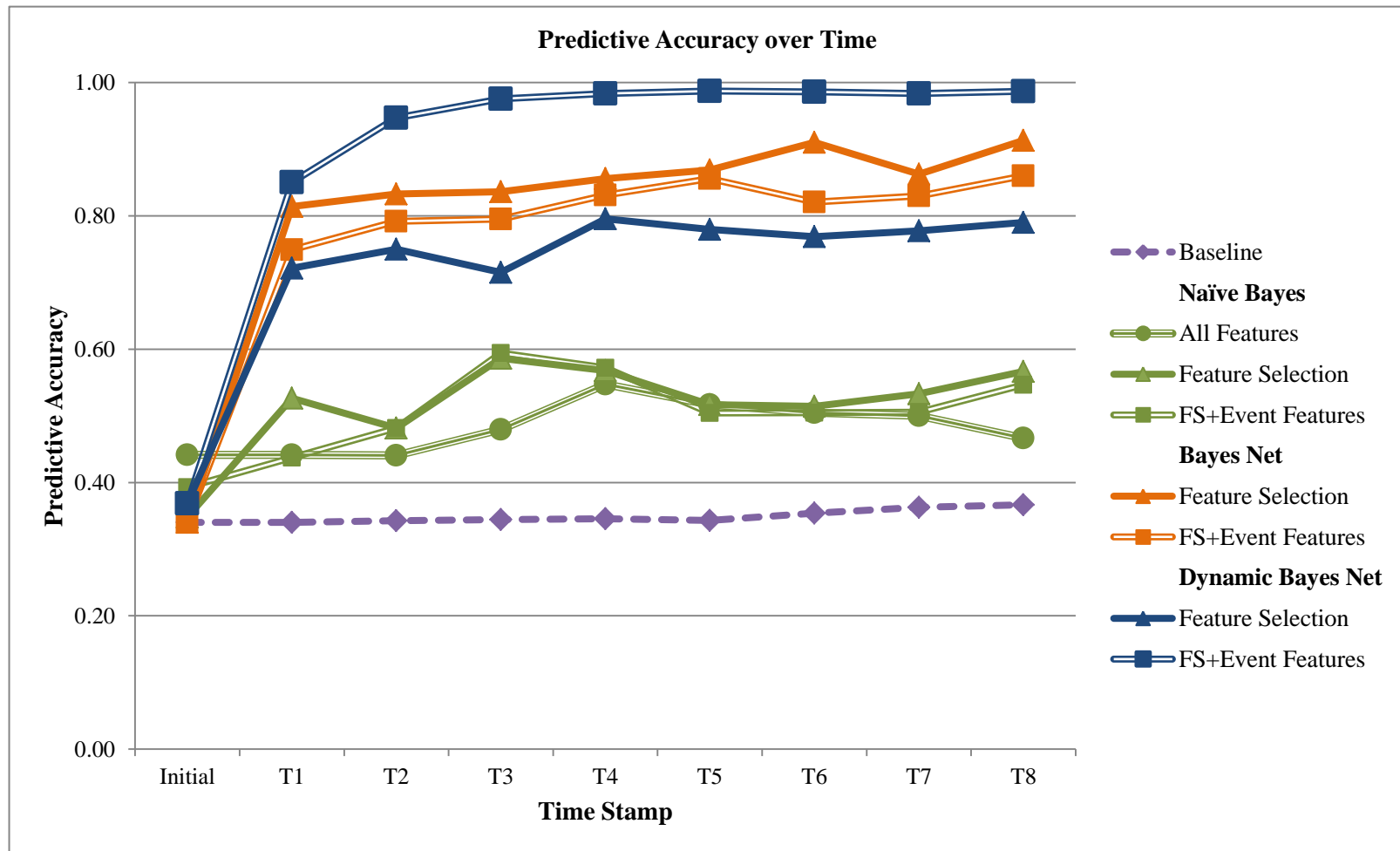
**Feature Selection** set alone (acc = 63%,  $\kappa = 0.41$ , weighted-  $\kappa = 0.43$ ),  $t(89) = 2.01$ ,  $p = 0.05$ . Both models represented a significant improvement over baseline as well as over the performance of the naïve models. These models showed a similar trend of steady improvement over time (Initial < T<sub>1-2</sub> < T<sub>3-8</sub>). These models show the improvement of theoretically grounded structure and suggest a predictive accuracy boost with the created contingency features.

For the dynamic Bayesian networks, the model trained on the **Feature Selection + Event Feature Creation** data set offered the best performance (acc = 73%,  $\kappa = 0.56$ , weighted-  $\kappa = 0.56$ ) though this was not statistically significant from the **Feature Selection** model (acc = 72%,  $\kappa = 0.53$ , weighted-  $\kappa = 0.53$ ),  $t(89) = 0.49$ ,  $p = 0.63$ . Like the prior naïve Bayes and static Bayes nets these models performed at levels close to baseline in the Initial time point with a significant boost in performance after interaction starts. Further analyses indicate no significant improvement across time for the dynamic Bayesian networks, which were relatively steady from T<sub>1</sub> on. Specifically, the best performing model is able to predict students' cognitive SRL class with approximately 80% accuracy starting at the first time stamp. These results are promising for early prediction models. Overall the dynamic models also offer a significant boost in performance over the static Bayesian networks highlighting the performance improvement offered by incorporating dynamic relationships between the important processes.

**Assessing Metacognitive SRL.** For the metacognitive component of SRL the baseline of most frequent class starts at 34% and increases slightly to 37% by T<sub>8</sub>. Averaged across all time points, the overall baseline is 35%. Table 18 lists the performance measures for each trained model, while Figure 24 charts the predictive accuracy across time. All of the naïve models of metacognitive SRL offered significant improvement over this baseline. The model trained on the **Feature Selection** set offered the best performance (acc = 50%,  $\kappa = 0.27$ , weighted-  $\kappa = 0.33$ ) though it was not statistically different ( $p = 0.20$ ) from the other models. Again, model performance was near baseline at the Initial time slice but improved over time

**Table 18.** Predictive model performance for metacognition

|                   | Initial                                    | T <sub>1</sub> | T <sub>2</sub> | T <sub>3</sub> | T <sub>4</sub> | T <sub>5</sub> | T <sub>6</sub> | T <sub>7</sub> | T <sub>8</sub> | Overall |      |
|-------------------|--|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|---------|------|
| Baseline          | 0.34                                       | 0.34           | 0.34           | 0.34           | 0.35           | 0.34           | 0.35           | 0.36           | 0.37           | 0.35    |      |
| Naïve Bayes       | All Features                               |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.44           | 0.44           | 0.44           | 0.48           | 0.55           | 0.52           | 0.50           | 0.50           | 0.47    | 0.48 |
|                   | Kappa                                      | 0.16           | 0.16           | 0.16           | 0.22           | 0.32           | 0.28           | 0.26           | 0.25           | 0.20    | 0.22 |
|                   | W. K.                                      | 0.19           | 0.23           | 0.22           | 0.31           | 0.38           | 0.38           | 0.32           | 0.33           | 0.29    | 0.29 |
|                   | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.35           | 0.53           | 0.48           | 0.59           | 0.57           | 0.52           | 0.51           | 0.53           | 0.57    | 0.52 |
|                   | Kappa                                      | 0.02           | 0.29           | 0.22           | 0.38           | 0.35           | 0.28           | 0.27           | 0.30           | 0.35    | 0.27 |
|                   | W. K.                                      | 0.03           | 0.35           | 0.28           | 0.40           | 0.40           | 0.37           | 0.35           | 0.36           | 0.42    | 0.33 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.39           | 0.44           | 0.48           | 0.59           | 0.57           | 0.50           | 0.50           | 0.51           | 0.55    | 0.50 |
| Kappa             | 0.09                                       | 0.16           | 0.22           | 0.39           | 0.36           | 0.26           | 0.26           | 0.25           | 0.32           | 0.26    |      |
| W. K.             | 0.09                                       | 0.24           | 0.27           | 0.40           | 0.39           | 0.32           | 0.36           | 0.35           | 0.42           | 0.32    |      |
| Bayes Net         | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.34           | 0.81           | 0.83           | 0.84           | 0.86           | 0.87           | 0.91           | 0.86           | 0.91    | 0.80 |
|                   | Kappa                                      | 0.01           | 0.72           | 0.75           | 0.75           | 0.78           | 0.80           | 0.87           | 0.79           | 0.87    | 0.70 |
|                   | W. K.                                      | 0.03           | 0.69           | 0.74           | 0.74           | 0.77           | 0.79           | 0.84           | 0.78           | 0.83    | 0.69 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | 0.63                                       | 0.34           | 0.75           | 0.79           | 0.80           | 0.83           | 0.86           | 0.82           | 0.83           | 0.86    | 0.76 |
|                   | 0.44                                       | 0.01           | 0.62           | 0.69           | 0.69           | 0.75           | 0.78           | 0.73           | 0.74           | 0.79    | 0.64 |
| 0.44              | 0.03                                       | 0.65           | 0.69           | 0.68           | 0.75           | 0.78           | 0.74           | 0.77           | 0.78           | 0.65    |      |
| Dynamic Bayes Net | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.37           | 0.72           | 0.75           | 0.72           | 0.80           | 0.78           | 0.77           | 0.78           | 0.79    | 0.72 |
|                   | Kappa                                      | 0.05           | 0.46           | 0.50           | 0.45           | 0.57           | 0.54           | 0.52           | 0.52           | 0.55    | 0.46 |
|                   | W. K.                                      | -0.07          | 0.57           | 0.61           | 0.56           | 0.68           | 0.65           | 0.60           | 0.63           | 0.63    | 0.54 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.37           | 0.85           | 0.95           | 0.98           | 0.98           | 0.99           | 0.99           | 0.98           | 0.99    | 0.90 |
|                   | Kappa                                      | 0.05           | 0.78           | 0.92           | 0.96           | 0.98           | 0.98           | 0.98           | 0.98           | 0.98    | 0.84 |
| W. K.             | -0.07                                      | 0.78           | 0.93           | 0.96           | 0.98           | 0.98           | 0.98           | 0.98           | 0.99           | 0.83    |      |



**Figure 24.** Predictive accuracy of metacognition over time

(Initial,  $T_{1-2} < T_{3-8}$ ). In general, this is very similar to the patterns indicated by the cognitive models. The naïve models show steady improvement over baseline, but no difference across data set.

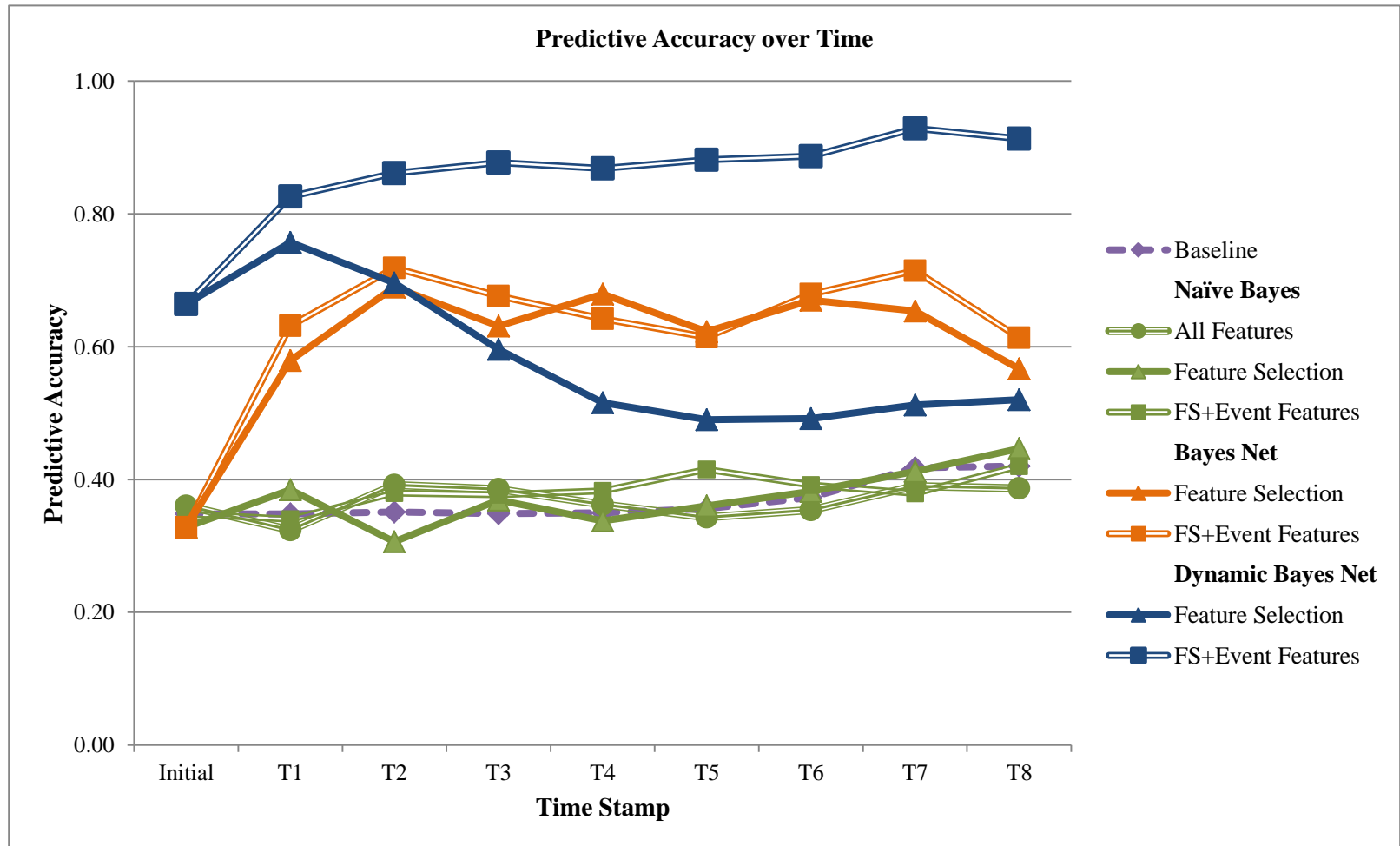
For the static Bayesian network the best performing model was again using the **Feature Selection** model (acc = 80%,  $\kappa = 0.70$ , weighted-  $\kappa = 0.69$ ), though there was no significant difference in performance when compared to the **Feature Selection + Event Feature Creation** model (acc = 76%,  $\kappa = 0.64$ , weighted-  $\kappa = 0.65$ ),  $t(89) = 1.44$ ,  $p = 0.15$ . Both models start with predictive accuracies near baseline at the Initial time slice and improve significantly once the interaction starts though there is no further significant increase in performance.

Of the dynamic Bayesian networks, the best performing model was created using the **Feature Selection + Event Feature** set (acc = 90%,  $\kappa = 0.84$ , weighted-  $\kappa = 0.83$ ). This model performed statistically significantly better than the **Feature Selection** model ( $t(89) = 15.96$ ,  $p < 0.001$ ) which actually performed slightly below the static Bayesian network (acc = 72%,  $\kappa = 0.46$ , weighted-  $\kappa = 0.54$ ). This suggests that adding the dynamic cycle information had differential impacts on the models from the two data sets. Both models demonstrated the common behavior of an accuracy near baseline at the Initial time stamp with significant improvements after the start of interaction. The best model is able to accurately predict 95% of student classifications by the second time stamp, which is an exceptionally high accuracy early into the interaction. Overall the learned models for metacognitive SRL models indicate that the dynamic cycle information may be beneficial in some cases, but there is unclear evidence in relation to data sets.

**Assessing Motivational SRL.** For the motivational classifications of SRL, the most common class baseline starts at 35% and increases until it reaches 42% at the final time point. Averaged across the entire time series, the total baseline accuracy is 37%. Table 19 lists the performance measures for each trained model, while Figure 25 charts the predictive accuracy across time. Overall, the naïve Bayesian models did not offer any improvement over the baseline. The best performing model in this case was trained on the **Feature Selection +**

**Table 19.** Predictive model performance for motivation

|                   | Initial                                    | T <sub>1</sub> | T <sub>2</sub> | T <sub>3</sub> | T <sub>4</sub> | T <sub>5</sub> | T <sub>6</sub> | T <sub>7</sub> | T <sub>8</sub> | Overall |      |
|-------------------|--|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|---------|------|
| Baseline          | 0.35                                       | 0.35           | 0.35           | 0.35           | 0.35           | 0.36           | 0.37           | 0.42           | 0.42           | 0.37    |      |
| Naïve Bayes       | All Features                               |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.36           | 0.32           | 0.39           | 0.39           | 0.36           | 0.34           | 0.35           | 0.39           | 0.39    | 0.37 |
|                   | Kappa                                      | 0.04           | -0.02          | 0.08           | 0.07           | 0.04           | 0.01           | 0.02           | 0.08           | 0.06    | 0.04 |
|                   | W. K.                                      | 0.01           | 0.00           | 0.16           | 0.11           | 0.05           | 0.03           | 0.06           | 0.14           | 0.09    | 0.07 |
|                   | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.33           | 0.38           | 0.31           | 0.37           | 0.34           | 0.36           | 0.38           | 0.41           | 0.45    | 0.37 |
|                   | Kappa                                      | -0.01          | 0.08           | -0.04          | 0.05           | 0.00           | 0.04           | 0.06           | 0.10           | 0.15    | 0.05 |
|                   | W. K.                                      | -0.01          | 0.08           | -0.06          | 0.11           | 0.01           | 0.06           | 0.05           | 0.11           | 0.15    | 0.05 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.35           | 0.34           | 0.38           | 0.38           | 0.38           | 0.42           | 0.39           | 0.38           | 0.42    | 0.38 |
| Kappa             | 0.02                                       | 0.01           | 0.07           | 0.06           | 0.07           | 0.12           | 0.08           | 0.05           | 0.11           | 0.07    |      |
| W. K.             | 0.04                                       | -0.01          | 0.08           | 0.05           | 0.07           | 0.13           | 0.10           | 0.10           | 0.10           | 0.07    |      |
| Bayes Net         | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.33           | 0.58           | 0.69           | 0.63           | 0.68           | 0.62           | 0.67           | 0.65           | 0.57    | 0.60 |
|                   | Kappa                                      | -0.01          | 0.37           | 0.53           | 0.45           | 0.52           | 0.43           | 0.50           | 0.48           | 0.36    | 0.40 |
|                   | W. K.                                      | 0.01           | 0.30           | 0.49           | 0.40           | 0.53           | 0.41           | 0.51           | 0.47           | 0.34    | 0.38 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | 0.63                                       | 0.72           | 0.68           | 0.64           | 0.61           | 0.68           | 0.71           | 0.61           | 0.62           | 0.33    | 0.63 |
|                   | 0.44                                       | 0.57           | 0.51           | 0.46           | 0.41           | 0.51           | 0.56           | 0.40           | 0.43           | -0.01   | 0.44 |
| 0.44              | 0.61                                       | 0.53           | 0.48           | 0.46           | 0.52           | 0.58           | 0.42           | 0.45           | 0.01           | 0.44    |      |
| Dynamic Bayes Net | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.66           | 0.76           | 0.70           | 0.60           | 0.52           | 0.49           | 0.49           | 0.51           | 0.52    | 0.58 |
|                   | Kappa                                      | 0.49           | 0.63           | 0.48           | 0.33           | 0.20           | 0.16           | 0.13           | 0.11           | 0.12    | 0.29 |
|                   | W. K.                                      | 0.36           | 0.70           | 0.51           | 0.33           | 0.20           | 0.14           | 0.13           | 0.11           | 0.12    | 0.29 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.66           | 0.83           | 0.86           | 0.88           | 0.87           | 0.88           | 0.89           | 0.93           | 0.91    | 0.86 |
|                   | Kappa                                      | 0.49           | 0.74           | 0.79           | 0.82           | 0.80           | 0.82           | 0.83           | 0.89           | 0.87    | 0.78 |
| W. K.             | 0.36                                       | 0.80           | 0.84           | 0.86           | 0.85           | 0.86           | 0.87           | 0.92           | 0.90           | 0.81    |      |



**Figure 25.** Predictive accuracy of motivation over time

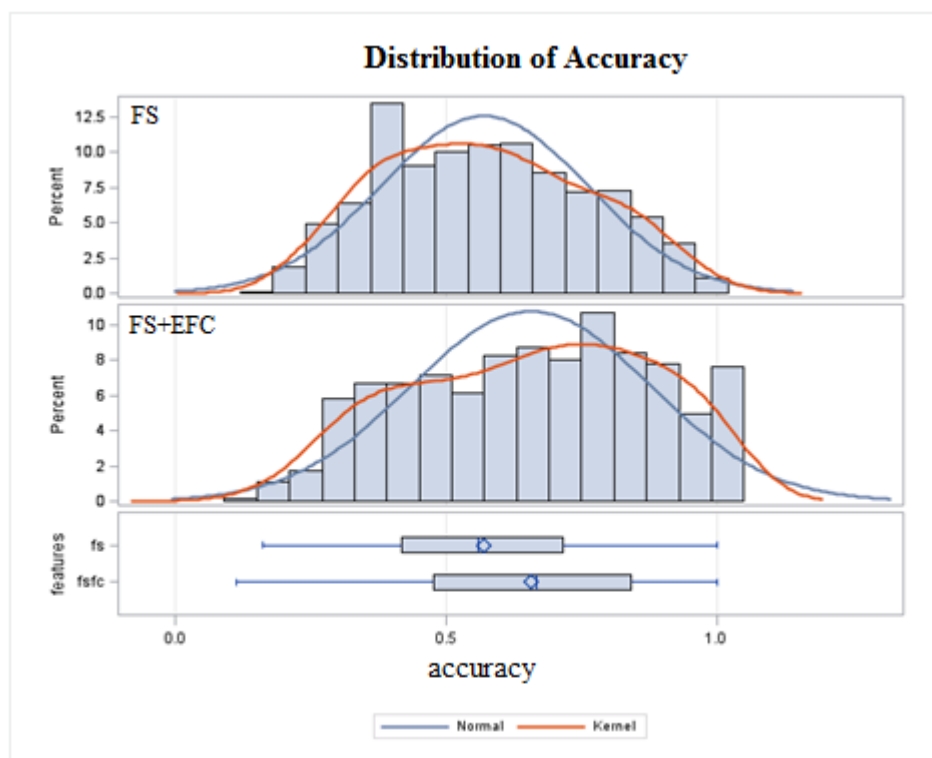
**Event Feature Creation** set (acc = 38%,  $\kappa = 0.07$ , weighted-  $\kappa = 0.07$ ). This did not differ significantly from baseline or from the other models. Furthermore, there is no improvement in performance across time as was seen in the other modeling approaches. Overall, the naïve modeling approach showed little promise in predicting motivational SRL.

The static Bayesian networks however did offer significant improvement over the baseline and naïve models. The highest performing model in this case was trained on the **Feature Selection +Event Feature Creation** set (acc = 62%,  $\kappa = 0.42$ , weighted-  $\kappa = 0.45$ ), though this was not significantly different from the **Feature Selection** model (acc = 60%,  $\kappa = 0.40$  weighted-  $\kappa = 0.38$ ),  $t(89) = 1.06$ ,  $p = 0.30$ . These models did demonstrate the common pattern of accuracies starting near baseline and increasing rapidly at the start of the interaction. Here the patterns appear more similar to the metacognitive modeling, with little accuracy gain after the start of the interaction.

Of the dynamic Bayesian networks, the best performing model was created using the **Feature Selection + Event Feature** set (acc = 86%,  $\kappa = 0.78$ , weighted-  $\kappa = 0.81$ ). This model performed statistically significantly better than the **Feature Selection** model ( $t(89) = 17.49$ ,  $p < 0.001$ ) which, like the metacognitive model of the same type, performed slightly below the static Bayesian network (acc = 58%,  $\kappa = 0.29$  weighted-  $\kappa = 0.29$ ). Here, there is an interesting pattern comparing predictive accuracy over time. Both models start with a predictive accuracy above baseline which is uncommon among the rest of the models. The **Feature Selection + Event Feature Creation** then sees the common increase in performance after the interaction starts and a steady growth in performance across time. The **Feature Selection** model on the other hand shows some initial benefit, but then accuracy steadily declines until  $T_4$  when it levels off again. This suggests there may be something different about how motivation is represented over time.

**Further Analyses.** Across all of the modeling approaches some patterns emerged. The **Feature Selection + Event Feature Creation** models were more often the best performing model. Similarly the dynamic Bayesian networks were typically better than static Bayesian

networks which in turn were better than the static Bayesian networks. However, to concretely describe the performance benefits of the theoretical decisions behind these modeling approaches it is important to compare the modeling approaches across the entire set. For this purpose a two-way ANOVA was conducted using the feature set and model type. All effects were statistically significant at  $p < 0.001$ . There was a main effect for feature set  $F_{(1, 1619)} = 113.55$ ,  $p < 0.001$ . Post-hoc t-tests indicated that the **Feature Selection + Event Feature Creation** models did significantly outperform the models trained on the **Feature Selection** set alone,  $t(1618) = 8.40$ ,  $p < 0.001$  (Figure 26). There was also a main effect for model type,  $F_{(2, 1618)} = 386.60$ ,  $p < 0.0001$ . Tukey post-hoc tests indicated that dynamic Bayes offered significantly better performance than static Bayesian networks, which in turn outperformed the naïve Bayesian networks (Figure 27). There was also a significant interaction effect,



**Figure 26.** Distributions of accuracy by feature set

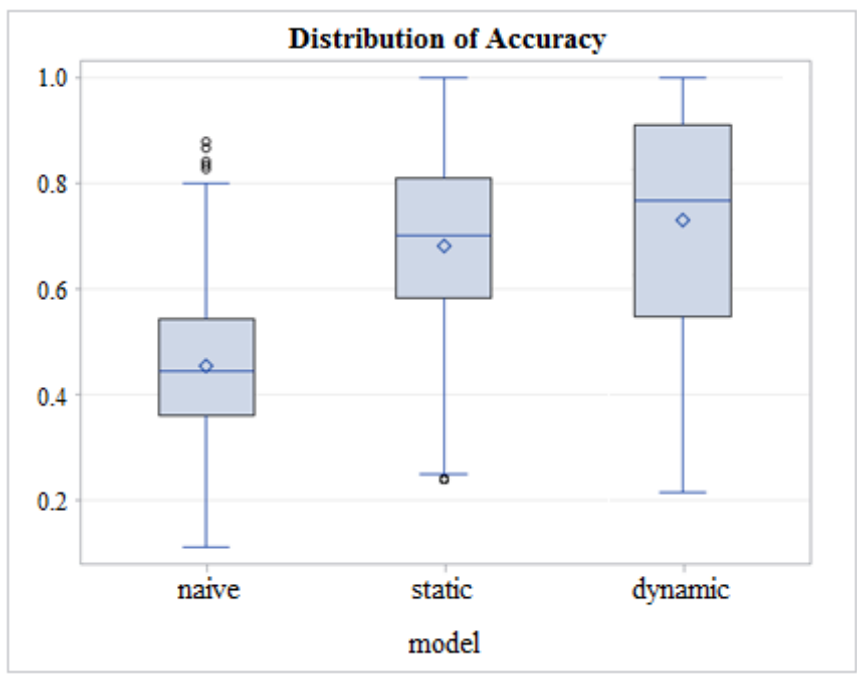


Figure 27. Distributions of accuracy by model type

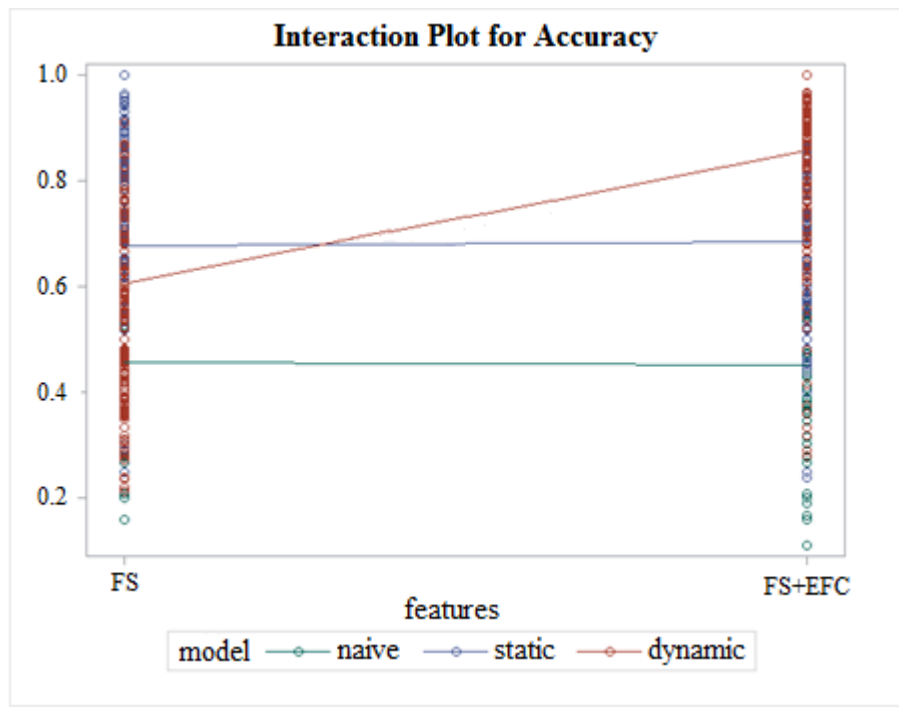


Figure 28. Interaction of feature set and model type

$F_{(2, 1618)} = 107.55, p < 0.001$ . This is caused by the improvement in performance for dynamic Bayesian networks trained on the **Feature Selection + Event Feature Creation** set, while there is reduced performance for the **Feature Selection** models (Figure 28).

### 7.3 Discussion

Overall, the predictive models were able to predict each component of self-regulated learning early into the interaction. It was common for accuracy to start near baseline and increase significantly after interaction information was added, surpassing baseline measures rapidly. This common pattern highlights the role that context and event data adds in improving these predictive models. Student attribute features alone did not offer enough predictive power to recognize self-regulated learning. Furthermore, comparisons of the models trained on the **Feature Selection** set and the **Feature Selection + Event Feature Creation** sets suggested that adding additional contingency and patterned contingency event data significantly improved the predictive power of the models. This was especially true for the dynamic Bayesian models. Overall, this suggests that each of the attribute and event features suggested by Winne (2010) were useful in the predictive models.

The next set of analyses compared the performance of the naïve, static, and dynamic Bayesian models. Overall, the results show that incorporating the important cyclic processes of SRL into the structure of the model improved performance. Models containing static representation of forethought, performance, and self-reflection outperformed those without this representations. Similarly, the dynamic Bayesian networks, which included the cyclic relationship of these processes across time had the best performance.

Overall these results support the two hypotheses being tested. The theoretical information led to predictive models that were capable of assessing self-regulated learning skill early into the interaction. This held true for the cognitive, metacognitive, and motivational models.

## CHAPTER 8

### Generalization

A key problem in machine learning is generalization. The primary assumption is that if the models are developed properly they should be able to recognize phenomena of interest in unseen populations. One of the primary ways people choose to estimate a model's generalizability is through cross-fold validation, where the model is trained on one dataset and tested on unseen data from the same population. However, this is not a guarantee that models will generalize to new populations (Alpaydin, 2004; Baker, Ocumpaugh, Gowda, & Heffernan, 2013). There are still many opportunities for overfitting, where the model is trained to specifically on the original population. There is also concern that different populations behave in fundamentally unique ways that would reduce any models' predictive power.

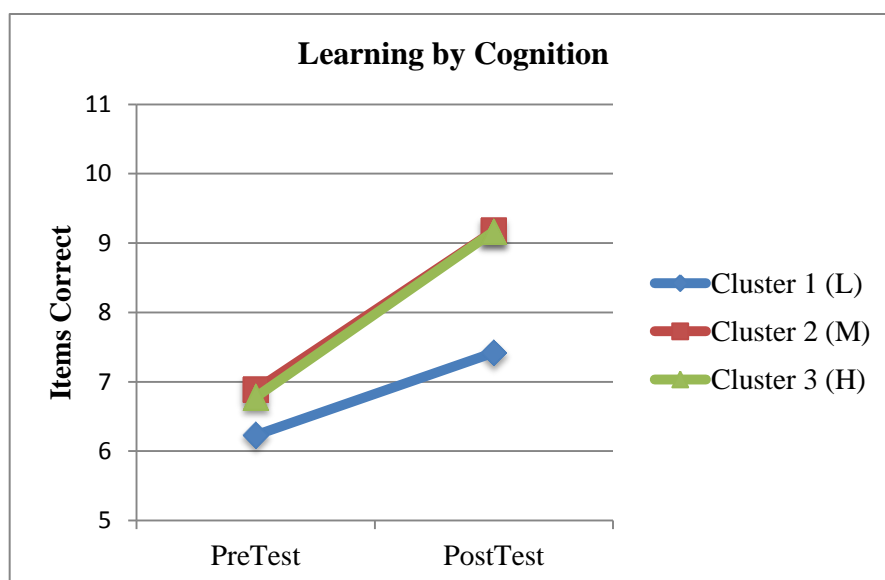
For this reason, the final step in this work is to explore how well the predictive SRL assessment models generalize to an unseen population. Here a separate testing corpus taken from a different school and population from the first is used to verify the generalization of the models. This will help identify if any of the techniques presented above are particularly prone to overfitting and further provide insight into whether theoretical grounding actually improves the predictive accuracy and generalizability of the models.

### 8.1 Population

Validation was conducted using the second corpus (Table 20) which included data from 140 students from the second North Carolina middle school mentioned in Section 3.2. The same classification procedures described in Chapter 5 were applied to this corpus of students, using the same splitting thresholds instead of forcing an even split ternary on the second population (Table 20).

**Table 20.** Secondary corpus classification details

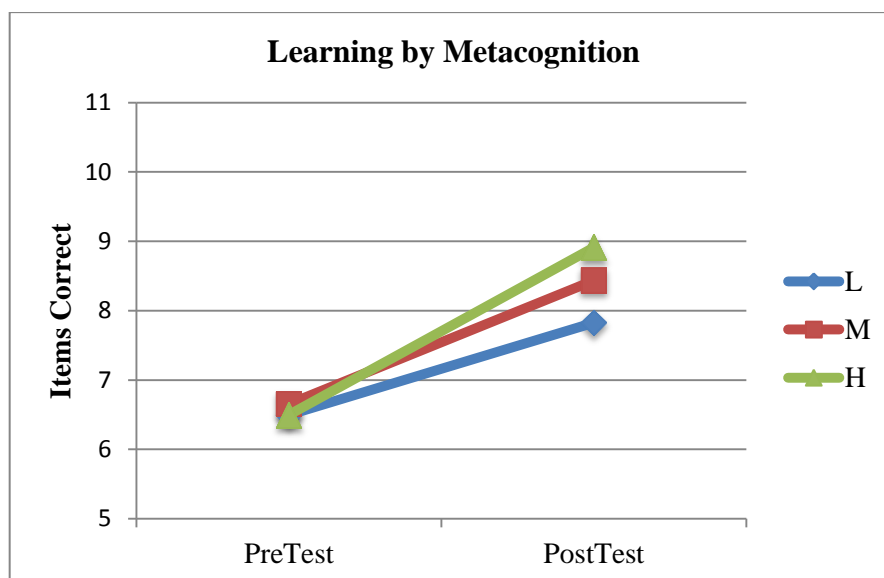
|               |   | N  | Pretest |     | Posttest |     | NLG  |      |
|---------------|---|----|---------|-----|----------|-----|------|------|
|               |   |    | Mean    | SD  | Mean     | SD  | Mean | SD   |
| Cognition     | L | 69 | 6.2     | 2.2 | 7.4      | 3.0 | 0.09 | 0.23 |
|               | M | 35 | 6.9     | 2.2 | 9.2      | 3.7 | 0.19 | 0.28 |
|               | H | 36 | 6.8     | 2.3 | 9.2      | 3.5 | 0.24 | 0.27 |
| Metacognition | L | 58 | 6.5     | 2.3 | 7.8      | 3.3 | 0.11 | 0.22 |
|               | M | 35 | 6.7     | 2.3 | 8.4      | 2.8 | 0.13 | 0.25 |
|               | H | 47 | 6.5     | 2.1 | 8.9      | 4.0 | 0.20 | 0.25 |
| Motivation    | L | 39 | 5.7     | 2.5 | 7.5      | 3.3 | 0.12 | 0.24 |
|               | M | 58 | 6.8     | 1.9 | 8.2      | 3.2 | 0.12 | 0.23 |
|               | H | 43 | 7.0     | 2.2 | 9.3      | 3.7 | 0.19 | 0.30 |

**Figure 29.** Secondary corpus learning differences by cognition

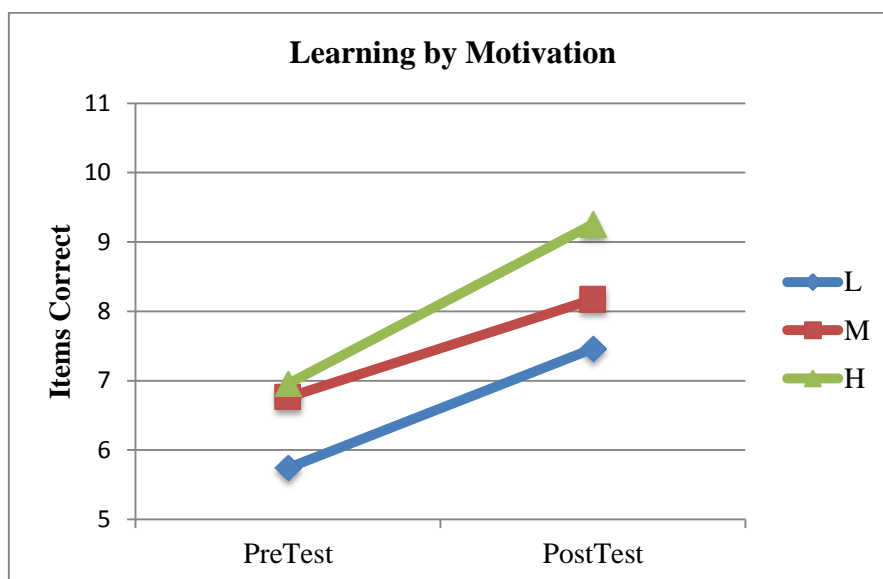
**Cognition.** The secondary corpus was fit to the clusters created from the primary corpus. This resulted in nearly half ( $N= 69$ , 41%) of the students being classified as Low cognitive SRL. Of the remaining students, there were 35 Medium (25%), and 36 High (26%) cognitive SRL students. As with the initial corpus, results indicate that there are no significant differences in pre-test ( $F_{(2, 137)} = 1.3$ ,  $p = 0.27$ ). Similarly, there were significant differences in both post-test ( $F_{(2, 137)} = 4.82$ ,  $p < 0.01$ ) and in normalized learning gains ( $F_{(2, 137)} = 3.35$ ,  $p = 0.04$ ). Tukey post-hoc tests indicated that High and Medium cognitive SRL students had significantly higher post-tests scores than Low cognitive SRL students (Figure 29). High cognitive SRL students also had higher normalized learning gains than Low cognitive SRL students. These patterns of learning closely resemble that of the first corpus.

**Metacognition.** Splitting on student metacognition scores resulted in an uneven distribution of 58 Low (41%), 35 Medium (25%), and 47 High (34%) students. Results indicate that there are no significant differences in either pre-test ( $F_{(2, 137)} = 0.06$ ,  $p = 0.94$ ), post-test ( $F_{(2, 137)} = 1.34$ ,  $p = 0.27$ ), or learning gains ( $F_{(2, 137)} = 1.64$ ,  $p = 0.19$ ). However, closer examination of the average test scores and learning gains between groups suggest that the same trend is demonstrated (Figure 30), though the population may not be large enough for statistical significance to be observed. Interestingly, while in the initial corpus Medium metacognitive SRL students performed closer to High metacognitive SRL students in terms of learning gains, they appear to perform more similarly to Low metacognitive SRL students in the validation corpus.

**Motivation.** Splitting on students' IMI scores at the same break points also resulted in an uneven distribution of 39 Low (28%), 58 Medium (41%), and 43 High (31%) students. Analyses indicate statistically significant differences in students' pre-test scores ( $F_{(2, 137)} = 3.67$ ,  $p = 0.02$ ) with Tukey post-hoc tests indicating that Low motivational SRL students had significantly less prior knowledge than Medium or High students (Figure 31). ANOVAs also indicated statistically significant differences on post-test ( $F_{(2, 137)} = 2.99$ ,  $p = 0.05$ ), though



**Figure 30.** Secondary corpus learning differences by metacognition



**Figure 31.** Secondary corpus learning differences by motivation

Tukey post-hoc tests did not yield any pairwise differences. Further analyses indicated no significant differences in learning gains ( $F_{(2, 137)} = 0.98$   $p = 0.38$ ). Similarly to the metacognition classification, it appears that Medium and Low motivation SRL students perform more similarly to each other than they did in the original corpus. These differences may be due to the nature of the even ternary split in the first corpus but the forced constraint in the second.

## 8.2 Predictive Models

To see how well the models extended to an unseen population, each model type was trained on the entire original corpus and then tested on the corpus from the second school. This provides us with a measure of how well the models perform on a completely unseen population.

**Cognition.** The performance measures for each generalized model of cognition are given in Table 21, while Figure 32 charts the generalized predictive accuracy across time. Of the models trained to recognize cognitive SRL, only the dynamic Bayesian network had any performance over a most common baseline (52%). The **Feature Selection** model correctly identified 61% of students and the **Feature Selection + Event Feature Creation** model recognized 63%. While the static Bayesian network did outperform the naïve Bayesian network, neither had a performance above baseline.

**Metacognition.** The performance measures for each generalized model of metacognition are given in Table 22, while Figure 33 charts the generalized predictive accuracy across time. Unlike the cognitive models, all the trained metacognitive models generalized to the unseen corpus and were able to outperform baseline measures (41%). The dynamic Bayesian networks performed the best with the **Feature Selection + Event Feature Creation** model correctly recognizing 64% of students, and the **Feature Selection** model recognizing 62% of students, though this is not a significant difference. While the dynamic Bayesian networks significantly outperformed the naïve and static Bayesian networks, there was no difference between the latter two. The static Bayesian network trained on **Feature Selection** only

**Table 21.** Generalized model performance for cognition

|                   | Initial                                    | T <sub>1</sub> | T <sub>2</sub> | T <sub>3</sub> | T <sub>4</sub> | T <sub>5</sub> | T <sub>6</sub> | T <sub>7</sub> | T <sub>8</sub> | Overall |       |
|-------------------|--|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|---------|-------|
| Baseline          | 0.49                                       | 0.49           | 0.49           | 0.50           | 0.50           | 0.50           | 0.51           | 0.55           | 0.59           | 0.52    |       |
| Naïve Bayes       | All Features                               |                |                |                |                |                |                |                |                |         |       |
|                   | Acc.                                       | 0.34           | 0.30           | 0.22           | 0.25           | 0.16           | 0.18           | 0.15           | 0.19           | 0.20    | 0.22  |
|                   | Kappa                                      | 0.03           | -0.03          | -0.13          | -0.09          | -0.23          | -0.19          | -0.21          | -0.11          | -0.13   | -0.12 |
|                   | W. K.                                      | 0.02           | -0.05          | -0.13          | -0.09          | -0.24          | -0.18          | -0.19          | -0.10          | -0.12   | -0.12 |
|                   | Feature Selection                          |                |                |                |                |                |                |                |                |         |       |
|                   | Acc.                                       | 0.28           | 0.35           | 0.28           | 0.25           | 0.21           | 0.19           | 0.25           | 0.25           | 0.17    | 0.25  |
|                   | Kappa                                      | -0.05          | 0.04           | -0.05          | -0.07          | -0.16          | -0.17          | -0.08          | -0.06          | -0.15   | -0.08 |
|                   | W. K.                                      | -0.02          | 0.04           | -0.03          | -0.06          | -0.16          | -0.15          | -0.07          | -0.06          | -0.17   | -0.08 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |       |
|                   | Acc.                                       | 0.28           | 0.29           | 0.32           | 0.27           | 0.24           | 0.25           | 0.28           | 0.26           | 0.25    | 0.27  |
| Kappa             | -0.05                                      | -0.02          | 0.01           | -0.08          | -0.13          | -0.11          | -0.05          | -0.05          | -0.11          | -0.07   |       |
| W. K.             | -0.02                                      | -0.01          | 0.05           | -0.05          | -0.12          | -0.13          | -0.07          | -0.05          | -0.11          | -0.06   |       |
| Bayes Net         | Feature Selection                          |                |                |                |                |                |                |                |                |         |       |
|                   | Acc.                                       | 0.41           | 0.42           | 0.44           | 0.45           | 0.43           | 0.43           | 0.43           | 0.46           | 0.46    | 0.44  |
|                   | Kappa                                      | 0.02           | 0.03           | 0.07           | 0.07           | 0.02           | 0.04           | 0.02           | 0.05           | 0.01    | 0.04  |
|                   | W. K.                                      | -0.01          | 0.06           | 0.07           | 0.08           | 0.04           | 0.05           | 0.05           | 0.07           | -0.01   | 0.05  |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |       |
|                   | Acc.                                       | 0.41           | 0.39           | 0.45           | 0.40           | 0.40           | 0.42           | 0.39           | 0.44           | 0.46    | 0.42  |
| Kappa             | 0.02                                       | 0.00           | 0.10           | 0.03           | 0.04           | 0.09           | 0.04           | 0.12           | 0.17           | 0.07    |       |
| W. K.             | -0.01                                      | 0.03           | 0.12           | 0.04           | 0.06           | 0.11           | 0.09           | 0.15           | 0.18           | 0.09    |       |
| Dynamic Bayes Net | Feature Selection                          |                |                |                |                |                |                |                |                |         |       |
|                   | Acc.                                       | 0.49           | 0.75           | 0.60           | 0.51           | 0.66           | 0.63           | 0.65           | 0.57           | 0.61    | 0.61  |
|                   | Kappa                                      | 0.02           | 0.59           | 0.27           | 0.19           | 0.32           | 0.29           | 0.31           | 0.05           | 0.04    | 0.23  |
|                   | W. K.                                      | 0.02           | 0.57           | 0.28           | 0.20           | 0.32           | 0.29           | 0.30           | 0.06           | 0.06    | 0.23  |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |       |
|                   | Acc.                                       | 0.49           | 0.67           | 0.75           | 0.72           | 0.64           | 0.64           | 0.55           | 0.61           | 0.59    | 0.63  |
| Kappa             | 0.02                                       | 0.46           | 0.58           | 0.50           | 0.32           | 0.35           | 0.10           | 0.17           | 0.03           | 0.28    |       |
| W. K.             | 0.02                                       | 0.49           | 0.59           | 0.47           | 0.32           | 0.22           | 0.09           | 0.15           | 0.04           | 0.27    |       |

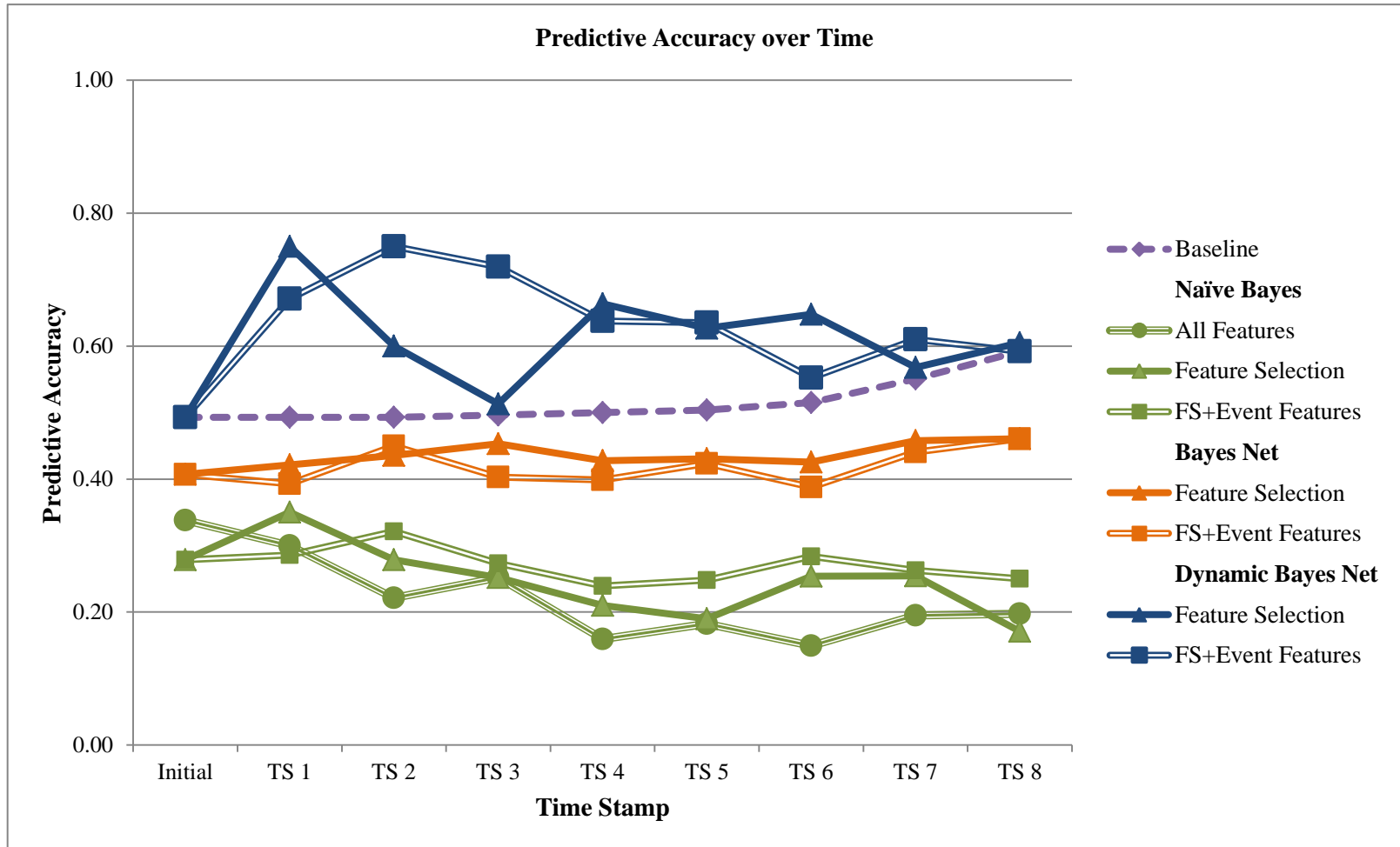
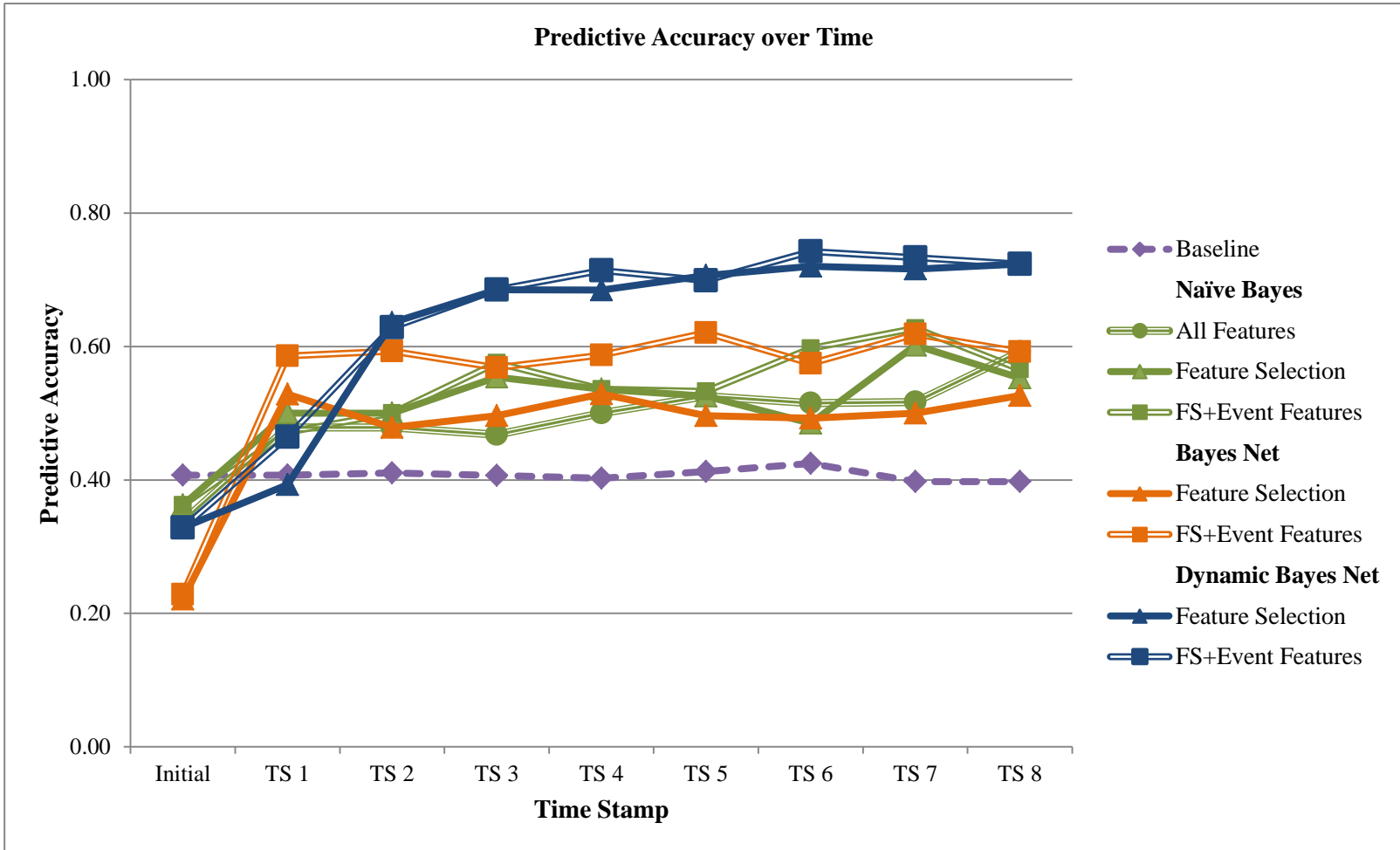


Figure 32. Generalized model predictive accuracy of cognition over time

**Table 22.** Generalized model performance for metacognition

|                   | Initial                                    | T <sub>1</sub> | T <sub>2</sub> | T <sub>3</sub> | T <sub>4</sub> | T <sub>5</sub> | T <sub>6</sub> | T <sub>7</sub> | T <sub>8</sub> | Overall |      |
|-------------------|--|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|---------|------|
| Baseline          | 0.41                                       | 0.41           | 0.41           | 0.41           | 0.40           | 0.41           | 0.42           | 0.40           | 0.40           | 0.41    |      |
| Naive Bayes       | All Features                               |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.34           | 0.48           | 0.48           | 0.47           | 0.50           | 0.53           | 0.51           | 0.52           | 0.59    | 0.49 |
|                   | Kappa                                      | -0.01          | 0.21           | 0.19           | 0.17           | 0.26           | 0.27           | 0.25           | 0.28           | 0.39    | 0.22 |
|                   | W. K.                                      | -0.02          | 0.25           | 0.27           | 0.20           | 0.26           | 0.32           | 0.30           | 0.30           | 0.45    | 0.26 |
|                   | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.36           | 0.50           | 0.50           | 0.55           | 0.54           | 0.53           | 0.49           | 0.60           | 0.55    | 0.51 |
|                   | Kappa                                      | 0.06           | 0.23           | 0.24           | 0.31           | 0.31           | 0.26           | 0.20           | 0.39           | 0.32    | 0.26 |
|                   | W. K.                                      | 0.04           | 0.26           | 0.27           | 0.35           | 0.31           | 0.33           | 0.28           | 0.43           | 0.35    | 0.29 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.36           | 0.47           | 0.50           | 0.58           | 0.54           | 0.53           | 0.60           | 0.63           | 0.57    | 0.53 |
| Kappa             | 0.06                                       | 0.20           | 0.24           | 0.35           | 0.29           | 0.28           | 0.38           | 0.42           | 0.34           | 0.28    |      |
| W. K.             | 0.04                                       | 0.22           | 0.30           | 0.40           | 0.29           | 0.32           | 0.40           | 0.49           | 0.39           | 0.32    |      |
| Bayes Net         | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.22           | 0.53           | 0.48           | 0.50           | 0.53           | 0.50           | 0.49           | 0.50           | 0.53    | 0.47 |
|                   | Kappa                                      | -0.02          | 0.28           | 0.21           | 0.24           | 0.30           | 0.25           | 0.24           | 0.24           | 0.29    | 0.23 |
|                   | W. K.                                      | -0.04          | 0.38           | 0.26           | 0.31           | 0.39           | 0.33           | 0.31           | 0.33           | 0.35    | 0.29 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.23           | 0.59           | 0.59           | 0.57           | 0.59           | 0.62           | 0.57           | 0.62           | 0.59    | 0.55 |
| Kappa             | -0.04                                      | 0.38           | 0.40           | 0.36           | 0.40           | 0.45           | 0.38           | 0.44           | 0.41           | 0.35    |      |
| W. K.             | -0.05                                      | 0.50           | 0.53           | 0.51           | 0.53           | 0.57           | 0.53           | 0.57           | 0.53           | 0.47    |      |
| Dynamic Bayes Net | Feature Selection                          |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.33           | 0.39           | 0.64           | 0.69           | 0.68           | 0.71           | 0.72           | 0.72           | 0.72    | 0.62 |
|                   | Kappa                                      | 0.05           | 0.30           | 0.55           | 0.60           | 0.60           | 0.62           | 0.64           | 0.63           | 0.64    | 0.52 |
|                   | W. K.                                      | 0.04           | 0.31           | 0.56           | 0.60           | 0.61           | 0.61           | 0.64           | 0.63           | 0.65    | 0.52 |
|                   | Feature Selection + Event Feature Creation |                |                |                |                |                |                |                |                |         |      |
|                   | Acc.                                       | 0.33           | 0.46           | 0.63           | 0.69           | 0.71           | 0.70           | 0.74           | 0.73           | 0.72    | 0.64 |
| Kappa             | 0.05                                       | 0.37           | 0.54           | 0.60           | 0.63           | 0.61           | 0.66           | 0.65           | 0.64           | 0.53    |      |
| W. K.             | 0.04                                       | 0.28           | 0.56           | 0.59           | 0.63           | 0.61           | 0.66           | 0.65           | 0.65           | 0.52    |      |



**Figure 33.** Generalized model predictive accuracy of metacognition over time

**Table 23.** Generalized model performance for motivation

|                   |  | Initial | T <sub>1</sub> | T <sub>2</sub> | T <sub>3</sub> | T <sub>4</sub> | T <sub>5</sub> | T <sub>6</sub> | T <sub>7</sub> | T <sub>8</sub> | Overall |
|-------------------|--|---------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|---------|
|                   | Baseline                                   | 0.40    | 0.40           | 0.40           | 0.40           | 0.39           | 0.40           | 0.37           | 0.40           | 0.46           | 0.40    |
| Naïve Bayes       | All Features                               |         |                |                |                |                |                |                |                |                |         |
|                   | Acc.                                       | 0.31    | 0.29           | 0.29           | 0.29           | 0.27           | 0.28           | 0.27           | 0.26           | 0.24           | 0.28    |
|                   | Kappa                                      | 0.03    | -0.03          | -0.04          | -0.04          | -0.06          | -0.04          | -0.04          | -0.04          | 0.00           | -0.03   |
|                   | W. K.                                      | 0.01    | -0.03          | -0.02          | -0.02          | -0.04          | -0.03          | -0.03          | -0.04          | 0.01           | -0.02   |
|                   | Feature Selection                          |         |                |                |                |                |                |                |                |                |         |
|                   | 0.44                                       | 0.29    | 0.31           | 0.36           | 0.31           | 0.30           | 0.36           | 0.49           | 0.36           | 0.40           | 0.44    |
|                   | 0.15                                       | -0.05   | -0.03          | 0.03           | -0.03          | -0.06          | 0.05           | 0.18           | 0.04           | 0.11           | 0.15    |
|                   | 0.17                                       | -0.02   | 0.02           | 0.09           | -0.03          | -0.03          | 0.01           | 0.26           | 0.07           | 0.14           | 0.17    |
|                   | Feature Selection + Event Feature Creation |         |                |                |                |                |                |                |                |                |         |
|                   | Acc.                                       | 0.31    | 0.31           | 0.29           | 0.29           | 0.29           | 0.28           | 0.29           | 0.31           | 0.36           | 0.30    |
| Kappa             | 0.03                                       | 0.04    | 0.00           | 0.00           | 0.00           | 0.00           | 0.00           | 0.00           | 0.00           | 0.01           |         |
| W. K.             | 0.01                                       | 0.03    | 0.01           | 0.01           | 0.00           | -0.01          | 0.00           | -0.01          | 0.00           | 0.00           |         |
| Bayes Net         | Feature Selection                          |         |                |                |                |                |                |                |                |                |         |
|                   | Acc.                                       | 0.36    | 0.45           | 0.31           | 0.30           | 0.33           | 0.36           | 0.32           | 0.42           | 0.36           | 0.36    |
|                   | Kappa                                      | 0.03    | 0.16           | -0.04          | -0.05          | 0.01           | 0.04           | -0.02          | 0.13           | 0.02           | 0.03    |
|                   | W. K.                                      | 0.01    | 0.16           | -0.07          | -0.05          | 0.04           | 0.08           | -0.02          | 0.20           | 0.08           | 0.05    |
|                   | Feature Selection + Event Feature Creation |         |                |                |                |                |                |                |                |                |         |
|                   | Acc.                                       | 0.40    | 0.44           | 0.29           | 0.31           | 0.36           | 0.31           | 0.30           | 0.36           | 0.49           | 0.36    |
| Kappa             | 0.11                                       | 0.15    | -0.05          | -0.03          | 0.03           | -0.03          | -0.06          | 0.05           | 0.18           | 0.04           |         |
| W. K.             | 0.14                                       | 0.17    | -0.02          | 0.02           | 0.09           | -0.03          | -0.03          | 0.01           | 0.26           | 0.07           |         |
| Dynamic Bayes Net | Feature Selection                          |         |                |                |                |                |                |                |                |                |         |
|                   | Acc.                                       | 0.33    | 0.54           | 0.54           | 0.59           | 0.58           | 0.61           | 0.62           | 0.62           | 0.62           | 0.56    |
|                   | Kappa                                      | 0.05    | 0.35           | 0.31           | 0.35           | 0.35           | 0.37           | 0.38           | 0.37           | 0.38           | 0.32    |
|                   | W. K.                                      | 0.04    | 0.37           | 0.32           | 0.35           | 0.36           | 0.36           | 0.38           | 0.37           | 0.38           | 0.33    |
|                   | Feature Selection + Event Feature Creation |         |                |                |                |                |                |                |                |                |         |
|                   | Acc.                                       | 0.33    | 0.61           | 0.53           | 0.59           | 0.61           | 0.60           | 0.64           | 0.63           | 0.62           | 0.57    |
| Kappa             | 0.05                                       | 0.46    | 0.31           | 0.35           | 0.37           | 0.36           | 0.39           | 0.39           | 0.38           | 0.34           |         |
| W. K.             | 0.04                                       | 0.32    | 0.32           | 0.34           | 0.37           | 0.36           | 0.40           | 0.39           | 0.38           | 0.33           |         |

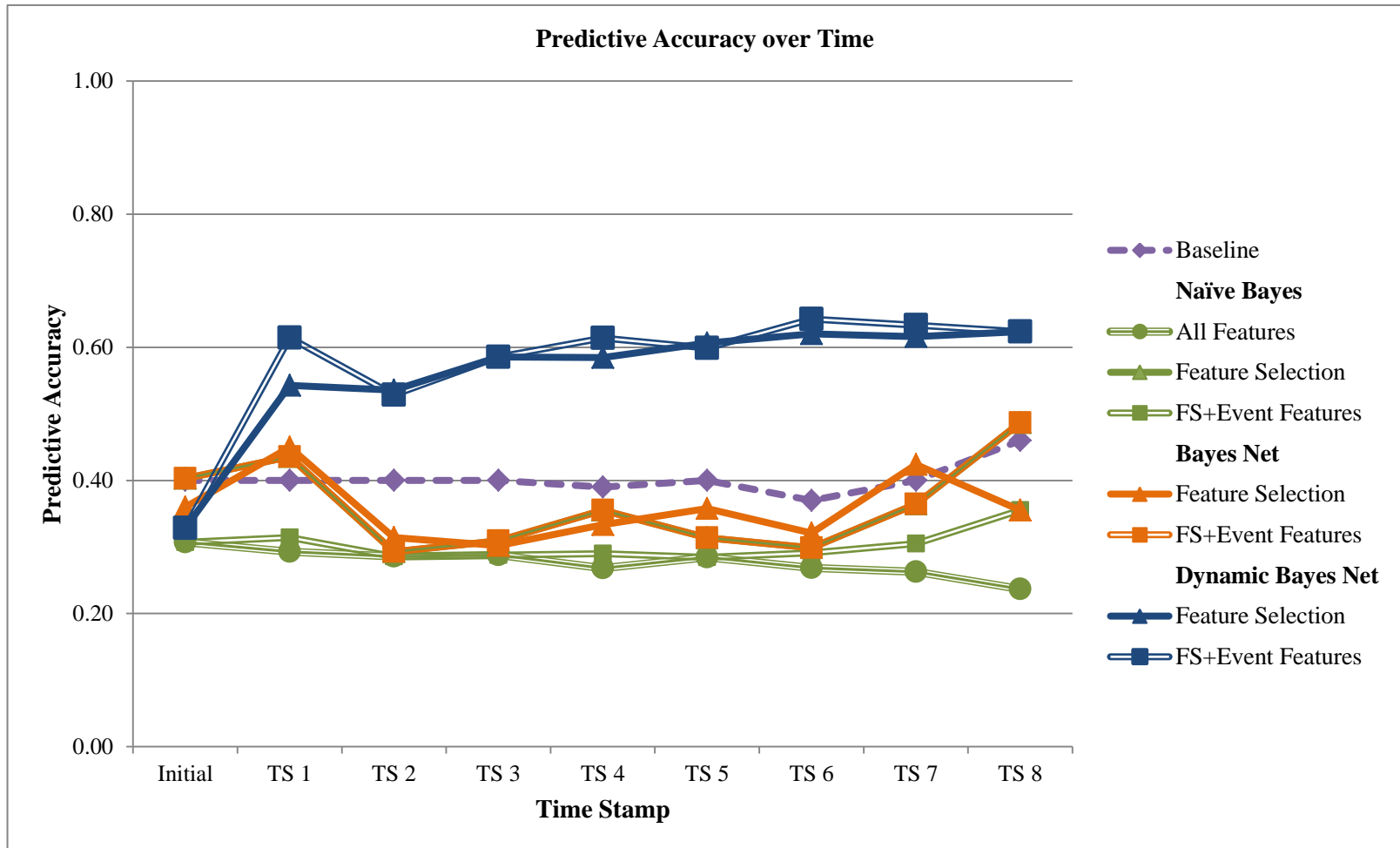


Figure 34. Generalized model predictive accuracy of motivation over time

recognized 47% of students, while the **Feature Selection + Event Feature Creation** model recognized 55% representing a significant difference in performance.

**Motivation.** The performance measures for each generalized model of motivation are given in Table 23, while Figure 34 charts the generalized predictive accuracy across time. Similar to the cognitive models, only dynamic Bayesian network significantly outperformed the most frequent baseline (40%). The dynamic Bayesian networks performed the best with the **Feature Selection + Event Feature Creation** model correctly recognizing 57% of students, and the **Feature Selection** model recognizing 56% of students. This was not a statistically significant difference. On average, the naïve Bayesian networks did worse than baseline while the static Bayesian networks were statistically equivalent to baseline.

### 8.3 Discussion

Overall, we see that the classification procedures generalized relatively well to the new population. The clustering approach for representing differences in problem-solving yielded similar learning gain differences as in the training corpus and appears to be a meaningful way of identifying patterns of cognitive SRL. However, the metacognitive and motivational representations of SRL did not generalize as well and did not demonstrate the same learning differences as in the primary corpus. This is likely because these classifications were identified by an even ternary split which was heavily dependent on the distribution of the original data. It suggests that identifying more meaningful ways of splitting this information may yield more generalizable results.

Of the predictive models, only the dynamic Bayesian networks reliably outperformed baseline measures, but appeared to generalize relatively well. It is common for models tested on new populations to fail to exceed predictive accuracies greater than chance so it was encouraging to see at least one model type able to do so. Furthermore, it was promising that similar patterns of model performance were seen in the testing corpus, suggesting that the conclusions about the overall approaches are likely valid. First, the dynamic models outper-

formed the naïve and static models that offered less theoretical grounding. Second, the dynamic Bayesian networks trained on the Feature Selection + Event Feature Creation tended to outperform the models without the contingency and patterned contingency features, though this difference was not significant. At the very least it suggests that the inclusion of these features does not harm the model by overfitting, which is a possible concern given how the features are created.

Overall, the performance of the models points to the need for more work in this area, with models built and trained on corpora from a variety of populations. Evidence suggests that only by including many populations can models start to overcome some of the issues commonly seen when attempting generalization (Baker et al., 2013).

## CHAPTER 9

### Conclusion

Self-regulated learning is a critical skill for success in many academic settings, including open-ended game-based learning environments. These systems require students to identify learning objectives, make complex connections and guide their own learning. Without self-regulated learning skills, students may flounder in these environments and fail to learn effectively. In order to recognize and support these skills it is necessary to develop real-time unobtrusive assessments of SRL. Trace-based stealth assessment has been highlighted as an important objective in the area of SRL but to date there have been few developments in how exactly to accomplish this task

This work presented a theory-driven machine-learning approach to real-time SRL assessment. We capitalize on the robust and informative theories of self-regulated learning that have been the subject of extensive investigation. These foundations are used to guide the selection and formation of empirical machine learning techniques. By matching the computational approaches with theoretical guidance it is hoped that the resulting models offer more meaningful, interpretable, and powerful models of self-regulated learning.

The proposed framework centers on three theoretical foundations and corresponding machine-learning approaches. First, SRL can be conceptualized as an attribute or an event. SRL events can be further described at occurrence, contingency, and patterned contingency levels. Feature selection and differential sequence mining are used to identify each of these features for predictive modeling purposes. Second, SRL is often described as having a cyclic nature comprised of three key processes: forethought, performance, and reflection. Students cycle through these processes continually during learning. These processes and the cyclic relationship are represented in the structure of the predictive Bayesian models. Finally, SRL behaviors are often described as falling into one of three categories: cognitive, metacognitive, and motivational. Students must have each of these components to be successful in learning settings. Similarly, it is important for computational models of SRL to recognize processes

related to these three components. For this reason, this framework proposes that each component should be identified and modeled to result in a robust portrait of self-regulated learning.

## 9.1 Hypotheses Revisited

This dissertation sought to evaluate the following thesis:

*The proposed theory-driven framework can guide the investigation of how machine learning and data mining techniques can be leveraged for stealth assessment of self-regulated learning.*

Here, the *usefulness* of the approach is defined by its ability of the proposed machine learning techniques to yield (1) meaningful indicators of self-regulated learning, and (2) real-time assessment models that offer significant increases in predictive accuracy. As part of this evaluation, this work explored three primary hypotheses and produced the following results:

- **Hypothesis 1:** Differential sequence mining can identify patterns of student behavior that:

(a) are indicative of self-regulated learning.

*The differential sequence mining approach yielded eight patterns of behavior that differed based on SRL level. These patterns offered indication of how highly self-regulated students used resources, drew connections and tracked knowledge. These findings point to possible means of intervention and provide a meaningful portrait of SRL behaviors in CRYSTAL ISLAND.*

(b) improve the predictive accuracy of real-time predictive models of self-regulated learning.

*Contingency and patterned contingency features were generated from the eight patterns identified by the differential sequence mining technique. These features were added to Bayesian models and compared against models with only occurrence event and personal attribute features. Results indicate that the addition of these feature sets did significantly improve predictive accuracy overall. Furthermore, there is evidence that these features do not harm predictive models' capabilities of generalizing to a new unseen population.*

- **Hypothesis 2:** Bayesian networks designed with a structure to encode the key processes of self-regulated learning will yield increased predictive power. Specifically,

(a) Static Bayesian networks that encode the key processes will achieve greater predictive accuracy than naïve Bayesian networks with no such representation.

*Static Bayesian models were designed to represent the forethought, performance, and self-reflection processes of self-regulated learning as hidden variables. Overall the static Bayesian networks offered a predictive accuracy significantly greater than the naïve Bayesian networks with no hidden variables. This indicates the benefit of encoding information about the key processes.*

(b) Dynamic Bayesian networks that encode the cyclic relationship between key processes across time will achieve greater predictive accuracy than static Bayesian networks with no such temporal representation.

*Dynamic Bayesian models were designed to extend the static Bayesian networks by representing the cyclic relationship between the three processes over time. Results indicated that this offered a*

*significant increase in predictive accuracy overall. This was especially true in the feature sets that included the features created using the differential sequence mining approach. Overall this supports the hypothesis that representing the cyclic nature of the key processes of SRL leads to better predictive assessment.*

- **Hypothesis 3:** The proposed theory-driven machine learning techniques can be used to assess the

**(a)** cognitive/behavioral

*Predictive models were trained to assess classifications of problem-solving strategy, which is a cognitive indicator of self-regulated learning skill. The best performing model correctly classified 73% of students over the entire interaction with the ability to reach this accuracy within 4 minutes of interaction. This is a significant increase in performance over the 45% baseline. This suggests that this approach is effective for recognizing this cognitive and behavioral component of SRL.*

**(b)** metacognitive

*Predictive models were trained to assess classifications of goal-setting and monitoring, which is a metacognitive indicator of self-regulated learning skill. The best performing model correctly classified 90% of students over the entire interaction with the ability to reach this accuracy within 11 minutes of interaction. This is a significant increase in performance over the 35% baseline. This suggests that this approach is effective for recognizing this metacognitive component of SRL.*

**(c)** motivational

*Predictive models were trained to assess classifications of intrinsic motivation, which is a motivational indicator of self-regulated*

*learning skill. The best performing model correctly classified 87% of students over the entire interaction with the ability to reach this accuracy within 4 minutes of interaction. This is a significant increase in performance over the 37% baseline. This suggests that this approach is effective for recognizing this motivational component of SRL. Furthermore, the approach was effective in modeling all three categories of SRL behaviors.*

aspects of self-regulated learning in real-time.

## **9.2 Summary**

This dissertation has explored the use of a novel theory-driven machine learning framework for real-time stealth assessment of self-regulated learning. Students were classified as high, medium, or low self-regulated learners based on evidence of their cognitive, metacognitive, and motivational tendencies. These classes were identified using definitions of self-regulated learning that are of particular interest in CRYSTAL ISLAND, with a focus on using a broad array of techniques. Specifically, cognitive skills of problem-solving strategy use were identified using an empirical data-mining approach. Metacognitive skills were identified through hand-annotated evidence of goal setting and monitoring. Finally, motivation was determined through a validated questionnaire.

Features for the predictive models were identified through feature selection and feature creation algorithms aimed at providing meaningful and predictive evidence of self-regulated learning. Stepwise logistic regression was used to identify the attribute and occurrence features that were most predictive of the previously defined self-regulated learning classes. Differential sequence mining was used to identify patterns of behaviors that occurred at statistically different levels based on SRL classification. These patterns provided interesting insights into how students use resources, make connections and interact with CRYSTAL ISLAND. They also provided patterns of actions that were transformed to represent contingency and patterned contingency level features.

Bayesian models were then learned and evaluated for real-time predictive performance. Models were trained with and without the features created from the differential sequence mining patterns to compare the predictive power given by these features. Three model structures were also compared to identify any benefits of encoding the cyclic process of SRL in the model. Each model was evaluated using 10-fold cross validation, a common approach for estimating how well the model will generalize. Further evaluation was conducted by evaluating the models on an entirely unseen population, a significantly more challenging problem, to determine the true generalizability of the models. The comparisons of the performance of these models was tested in order to evaluate the above hypotheses.

Results indicated that the differential sequence mining technique was useful for uncovering complex patterns of behavior that were not indicated by occurrence level events alone. For example, at the occurrence level there was no indication of any difference in use of the diagnosis worksheet. However, the differential patterns indicated that highly self-regulated students were more likely to note information immediately receiving it. Such patterns provide meaningful indicators of the ways in which self-regulated learning manifests in CRYSTAL ISLAND and can help guide future development of adaptive scaffolding. Furthermore, the features developed from these patterns yielded statistically significant gains in predictive accuracy over the models without these features when evaluated on the same population. These features showed trends of improvement in the generalization problem, though these differences were not statistically significant.

The naïve, static and dynamic models were compared to identify the benefit of encoding theoretical structure in the models. Results indicated that the more closely the structure of the model encoded the proposed theory, the better its predictive accuracy. This held true across the generalization test as well. These results highlight the important role that theory can play in guiding empirical modeling. Furthermore, they hint at how empirical modeling can be used to refine current theories.

Overall, the approach was effective for each of the three components of SRL that were explored. It is hoped that the nature of these classifications and results will allow this

framework to be adopted more broadly and in other computer-based environments. A wide variety of techniques was used to identify self-regulated behaviors of interest in CRYSTAL ISLAND. These behaviors represented very distinct areas of SRL as well. If the theory-driven framework worked well in each of these very different cases, it suggests there is a good chance it may work for other environments and other self-regulated learning constructs.

### **9.3 Limitations**

This dissertation presents a theory-driven machine learning framework for assessing self-regulated learning. The aim is to develop a framework that can be easily adapted and applied to many other aspects of self-regulated learning and different computer-based environments. However, along these lines there are several limitations. First this work only evaluates the framework in the context of CRYSTAL ISLAND, a narrative-based game environment, which is quite unique when compared to more traditional learning systems. It is unclear how well the results from this environment will generalize to other tutoring system, especially those that are more structured and problem-based.

This work also only uses data from two schools in rural North Carolina. While both schools had a reasonable level of diversity there are likely findings that are very specific to the population. Furthermore, attempts were made to identify how well the models could generalize to unseen populations by extending models trained on data from the first school to students from the second. Results indicate a reduction in performance, but still some models were able to outperform a most-frequent baseline classifier. While the data is still from a second population, it is likely that the populations are fairly similar when considering a global context. The generalizability rates to the second school are likely not reflective of how well the results would extend to urban US schools, and even less likely to extend to international students. This is a common issue in educational data mining and can only be rectified with more data and investigation in broad diverse populations

The cognitive, metacognitive, and motivational evidence of SRL used in this work were selected to be representative of common SRL processes. However, there may be some

issues of validity when using these measures as a representation of SRL. Specifically, while the behaviors indicated by the cognitive clusters appear to be related to problem solving, further investigation and validation would be necessary to concretely make such a claim. Furthermore, the annotation of metacognitive goal setting may be similarly called into question since the prompts to the students were not designed to measure metacognition. A more robust prompt designed specifically for this purpose may yield higher quality classifications.

This work makes claims that theoretical grounding improved the performance of the machine learning approaches. However, this is likely highly dependent on the theories, machine learning techniques, and interpretations that were applied. The empirical techniques and theories were matched based on the researcher's interpretation and understanding of the constructs and may consequently reflect some bias. Overall, it is hoped that the work demonstrates the feasibility of the approach; however, it will be necessary to continue exploring how additional theories and machine learning techniques can be used to assess self-regulated learning.

## 9.4 Future Directions

A key area of future work is to explore how well the framework generalizes beyond the presented findings. Generalization can be explored in a variety of directions. First, we presented three representations of SRL that are of particular interest in CRYSTAL ISLAND. However there are many other typical SRL behaviors of interest such as critical thinking and self-recording that would be interesting to explore within the framework. It is also important to see how these representations may change with other environments. It is expected that as the framework is successfully (or unsuccessfully) applied to other learning environments and representations of SRL, the more refined and robust it will become over time. This is necessary to further test the validity of the proposed framework.

Another area of extension is to continue to train the CRYSTAL ISLAND specific models on new populations of students. With data from more, and more diverse, students the models will likely improve in their ability to successfully recognize self-regulated learning early

enough into an interaction to provide feedback. Further work is necessary to ensure that the models are recognizing SRL at a predictive accuracy that is good enough to inform intervention, especially since the models would be applied to new populations. After this has been established, the next step will be to develop the adaptive scaffolding to support students who have not yet mastered some components of self-regulated learning. This support should likely be targeted to the specific level of a student's skill as well as the components (cognitive, metacognitive, or motivational) that seem to be needed the most. Current research suggests that scaffolding targeted directly at the skill level of the student is needed to prevent floundering, and to allow skilled students the opportunity to demonstrate their abilities (Alfieri et al., 2011; Easterday et al., 2011; Koedinger & Alevan, 2007). Consequently, the usefulness of the models themselves will be determined by their ability to appropriately direct scaffolding in a way that has meaningful and measurable outcomes.

Finally, another interesting area of future work is to extend the framework to explore different psychological theories of self-regulated learning. The three theories that were selected were highlighted because of their acceptance among experts as well as their clear ties to empirical machine learning approaches. However, there is no universally accepted theory, and there is still much contention about some key processes. Therefore an important area of future work is to explore many more theoretical foundations. It will be an interesting creative endeavor to see how these can be mapped to empirical machine learning approaches. It will also be interesting to see whether different theories provide different levels of improvement in predictive power. This could go a long way towards refining existing psychological theories based on empirical investigations. In this way educational data mining and machine learning communities could transition from simply being consumers of learning theories to being contributors.

## **9.5 Concluding Remarks**

This dissertation sought to address the challenge of trace-based stealth assessment of self-regulated learning. This has been highlighted as an important direction for improving educa-

tional technologies, especially those that allow a significant amount of student autonomy. This work highlighted the role that psychological theory can play in informing the selection and structure of empirical machine-learning techniques. The overarching framework that was proposed and evaluated shows promise in guiding real-time assessment of SRL both in the CRYSTAL ISLAND environment and in other more varied settings. The work presented was designed to be open and easily adapted by other researchers with different environments, different conceptualizations of SRL, and different sources of data. It is hoped that this framework will serve as inspiration to other researchers to help guide the development of new models of self-regulated learning assessment and the next generation of adaptive learning technologies.

## REFERENCES

- Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward Meta-cognitive Tutoring: A Model of Help-Seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education, 16*, 101–128.
- Aleven, V., Roll, I., McLaren, B., & Koedinger, K. (2010). Automated, Unobtrusive, Action-by-Action Assessment of Self-Regulation During Learning with an Intelligent Tutoring System. *Educational Psychologist, 45*(4), 224–233.
- Alfieri, L., Brooks, P., Aldrich, N., & Tenenbaum, H. (2011). Does Discovery-Based Instruction Enhance Learning. *Journal of Education Psychology, 103*(1), 1–18.
- Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press.
- Amershi, S., & Conati, C. (2006). Automatic Recognition of Learner Groups in Exploratory Learning Environments. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 463–472).
- Azevedo, R., Cromley, J., Winters, F., Moos, D., & Greene, J. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science, 33*, 381–412.
- Azevedo, R., Moos, D., Greene, J., Winters, F., & Cromley, J. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development, 56*(56), 45–72.
- Azevedo, R., Johnson, A., Chauncey, A., & Burkett, C. (2010). Self-Regulated Learning with MetaTutor: Advancing the Science of Learning with MetaCognitive Tools. In M. Khine & I. Saleh (Eds.), *New Science of Learning: Cognition, Computers and Collaboration in Education* (pp. 225–248). New York: Springer.
- Azevedo, R., Landis, R., Feyzi-Behnagh, R., Duffy, M., Trevors, G., Harley, J., Bouchet, F., et al. (2012). The Effectiveness of Pedagogical Agents' Prompting and Feedback in Facilitating Co-Adapted Learning with MetaTutor. *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 212–221).
- Azevedo, Roger, & Witherspoon, A. (2009). Self-regulated use of hypermedia. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 319–339). Mahwah, N.J.: Erlbaum.

- Baker, R., Corbett, A., & Aleven, V. (2008). More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems* (pp. 406–415).
- Baker, R., Corbett, A., Koedinger, K., & Wagner, A. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 383–390). ACM Press.
- Baker, R., D’Mello, S., Rodrigo, S., & Graesser, A. (2010). Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners’ Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*, 68(4), 223–241.
- Baker, R., Ocumpaugh, J., Gowda, S., & Heffernan, N. (2013). Ensuring Reliability of Educational Data Mining Detectors for Diverse Populations of Learners. *CREA: Center for Culturally Responsive Evaluation and Assessment: Inaugural Conference*.
- Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1), 3–17.
- Beal, C., Mitra, S., & Cohen, P. (2007). Modeling Learning Patterns of Students with a Tutoring System Using Hidden Markov Models. *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (pp. 238–245).
- Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010). Measuring Self-regulated Learning Skills through Social Interactions in a Teachable Agent Environment. *Research and Practice in Technology Enhanced Learning*, 5(2), 123–152.
- Biswas, G., Jeong, H., Roscoe, R., & Sulcer, B. (2009). Promoting Motivation and Self-Regulated Learning Skills through Social Interactions in Agent-Based Learning Environments. *2009 AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems*.
- Bless, H., Clore, G., Schwarz, N., Golisano, V., Rabe, C., & Wolk, M. (1996). Mood and the Use of Scripts: Does a Happy Mood Really Lead to Mindlessness? *Journal of Personality and Social Psychology*, 71, 665–679.
- Bloom, B. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 13(6), 4–16.

- Clarke, J., & Dede, C. (2009). Design for Scalability: A Case Study of the River City Curriculum. *Journal of Science Education and Technology*, 18(4), 353–365.
- Cocea, M., Hershkovitz, A., & Baker, R. (2009). The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate. *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 507–514). IOS Press.
- Conati, C. (2002). Probabilistic Assessment of User's Emotions in Educational Games. *International Journal of Applied Artificial Intelligence, Special Issue on Merging Cognition and Affect in HCI*, 16(1), 555–575.
- Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19(3), 267–303.
- Corbett, A., & Anderson, J. (1994). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.
- Cuevas, P., Lee, O., Hart, J., & Deaktor, R. (2005). Improving Science Inquiry with Elementary Students of Diverse Backgrounds. *Journal of Research in Science Teaching*, 42(3), 337–357.
- Davis, E. (2003). Prompting Middle School Science Students for Productive Reflection: Generic and Directed Prompts. *Journal of the Learning Sciences*, 12(1), 91–142.
- DiBenedetto, M., & Zimmerman, B. (2013) Construct and Predictive Validity of Microanalytic Measures of Students' Self-Regulation of Science Learning. *Learning and Individual Differences*, 26, 30-41.
- Easterday, M., Aleven, V., Scheines, R., & Carver, S. (2011). Using Tutors to Improve Educational Games. *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 63–71). Berlin Heidelberg: Springer Verlag.
- Elliot, A., & McGregor, H. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology*, 80, 501–519.
- Ellis, D., & Zimmerman, B. (2001). Enhancing self-monitoring during self-regulated learning of speech. In H. J. Hartman (Ed.), (pp. 205–228). Dordrecht, The Netherlands: Kluwer.
- Fiorella, L., & Mayer, R. (2012). Paper-based aids for learning with a computer-based game. *Journal of Educational Psychology*, 104(4), 1074–1082.

- Gee, J. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.
- Gernefski, N., & Kraati, V. (2006). Cognitive Emotion Regulation Questionnaire: Development of a Short 18-Item Version. *Personality and Individual Differences, 41*, 1045–1053.
- Gertner, A., Conati, C., & VanLehn, K. (1998). Procedural help in Andes: Generating hints using a Bayesian network student model. *Proceedings of the 15th National Conference on Artificial Intelligence*.
- Gong, Y., Beck, J., Heffernan, N., & Forbes-Summers, E. (2010). The Fine-Grained Impact of Gaming (?) on Learning. *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 194–203). Springer-Verlag.
- Greene, J., & Azevedo, R. (2010). The Measurement of Learners' Self-Regulated Cognitive and Metacognitive Processes while Using Computer-Based Learning Environments. *Educational Psychologist, 45*(4), 203–209.
- Hall, M., Frank, E., Holmes, G., Pfajhringer, B., Reutmann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1).
- Hallinen, N., Walker, E., Wylie, R., Ogan, A., & Jones, C. (2009). I was playing when I learned: a narrative game for french aspectual distinctions. *Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education* (pp. 117–120).
- Harp, S., & Mayer, R. (1998). How Seductive Details Do Their Damage: A Theory of Cognitive Interest in Science Learning. *Journal of Educational Psychology, 90*(3), 414–434.
- Herold, J., Zundel, A., & Stahovich, T. (2013). Mining Meaningful Patterns from Students' Handwritten Coursework. *Proceedings of the 6th International Conference on Educational Data Mining*.
- Ifenthaler, D. (2012). Determining the Effectiveness of Prompts for Self-Regulated Learning in Problem-Solving Scenarios. *Educational Technology & Society, 15*(1), 38–52.
- Johnson, W. (2010). Serious Use of a Serious Game for Language Learning. *International Journal of Artificial Intelligence in Education, 20*(2), 175–195.

- Kanfer, R., & Ackerman, P. (1989). Motivation and Cognitive Abilities: An Integrative/Aptitude-Treatment Interaction Approach to Skill Acquisition. *Journal of Applied Psychology, 74*, 657–690.
- Kapp, K. (2012). *The Gamification of Learning and Instruction*. San Francisco: Pfeiffer.
- Kauffman, D. (2004). Self-Regulated Learning in Web-Based Environments: Instructional Tools Designed to Facilitate Cognitive Strategy Use, Metacognitive Processing, and Motivational Beliefs. *Journal of Educational Computing Research, 30*, 139–161.
- Ketelhut, D. J. (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in “River City”, a multi-user virtual environment. *Journal of Science Education and Technology, 16*(1), 99–111.
- Kim, J., Hill, R., Durlach, P., Lane, H., Forbell, E., Core, M., Marsella, S., & et al. (2009). BiLAT: A game-based environment for practicing negotiation in a cultural context. *International Journal of Artificial Intelligence in Education, 19*(3), 289–308.
- Kinnebrew, J., & Biswas, G. (2012). Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. *Proceedings of the 5th International Conference on Educational Data Mining*.
- Kinnebrew, J., Loretz, K., & Biswas, G. (2013). A Contextualized, Differential Sequence Mining Method to Derive Students’ Learning Behavior Patterns. *Journal of Educational Data Mining, In Press*.
- Kinnebrew, J., Mack, D., & Biswas, G. (2013). Mining Temporally-Interesting Learning Behavior Patterns. *Proceedings of the 6th International Conference on Educational Data Mining*.
- Kirschner, P., Sweller, J., & Clark, R. (2006). Why Minimal Guidance during instruction does not work: An analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist, 41*, 75–86.
- Koedinger, K., & Aleven, V. (2007). Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educational Psychology Review, 19*, 239–364.
- Kostons, D., van Gog, T., & Paas, F. (2012). Training Self-Assessment and Task-Selection Skills: A Cognitive Approach to Improving Self-Regulated Learning. *Learning and Instruction, 22*, 121–132.

- Land, S. (2000). Cognitive requirements for learning with open-ended learning environments. *Educational Technology Research and Development*, 48(3), 61–78.
- Martinez-Maldonado, R., Yacef, K., & Kay, J. (2013). Data Mining in the Classroom: Discovering Groups' Strategies at a Multi-tabletop Environment. *Proceedings of the 6th International Conference on Educational Data Mining*.
- Mayer, R., & Johnson, C. (2010). Adding instructional features that promote learning in a game-like environment. *Journal of Educational Computing Research*, 42(3), 241–265.
- McAuley, E., Duncan, T., & Tammen, V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60(1), 48–58.
- McCrae, R., & Costa, P. (1993). *Personality in Adulthood: A Five-Factor Theory Perspective* (2nd ed.). New York: Guilford Press.
- McNamara, D., Jackson, G., & Graesser, A. (2009). Intelligent tutoring and games (ITaG). *Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education/Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education* (pp. 1–10).
- Meyer, D., & Turner, J. (2006). Re-conceptualizing Emotion and Motivation to Learn in Classroom Contexts. *Educational Psychology Review*, 18(4), 377–390.
- Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & L. C. Jain (Eds.), *Serious Games and Edutainment Applications* (pp. 169–195). London: Springer-Verlag.
- Paiva, A., Dias, J., Sobral, D., Aylett, R., Woods, S., Hall, L., & Zoll, C. (2005). Learning by feeling: Evoking empathy with synthetic characters. *Applied Artificial Intelligence*, 19, 235–266.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. (2002). Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist*, 37(2), 91–105. doi:10.1207/S15326985EP3702\_4
- Picard, R., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., et al. (2004). Affective Learning — A Manifesto. *BT Technology Journal*, 22(4), 253–269.

- Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review*, 16, 385–407.
- Ramirez, O. M., & Dockweiler, C. J. (1987). Mathematics Anxiety: A Systematic Review. In R. Schwarzer, H. M. Ploeg, & C. D. Spielberger (Eds.), *Advances in test anxiety research* (pp. 157–175). Hillsdale, NJ: Erlbaum Associates.
- Roll, I., Alevan, V., & Koedinger, K. (2010). The Invention Lab: Using a Hybrid of Model Tracing and Constraint-Based Modeling to Offer Intelligent Support in Inquiry Environments. *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 115–124).
- Rowe, J., Shores, L., Mott, B., & Lester, J. (2010). Individual Differences in Gameplay and Learning: A Narrative-Centered Learning Perspective. *Proceedings of the 5th International Conference on Foundations of Digital Games* (pp. 171–178). Monterey, CA: ACM.
- Rowe, J., McQuiggan, S., Robison, J., & Lester, J. (2009). Off-Task Behavior in Narrative-Centered Learning Environments. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 99–106). Brighton, UK: IOS Press.
- Rowe, J., Shores, L., Mott, B., & Lester, J. (2011). Integrating Learning, Problem Solving, and Engagement in Narrative-Centered Learning Environments. *International Journal of Artificial Intelligence in Education*, 166–177.
- Ryan, R., Connell, J., & Plant, R. (1990). Emotions in non-directed text learning. *Learning and Individual Differences*, 2, 1–17.
- Ryan, R., Koestner, R., & Deci, E. (1991). Varied forms of persistence: When free-choice behavior is not intrinsically motivated. *Motivation and Emotion*, 15, 185–205.
- Sabourin, J., Mott, B., & Lester, J. (2011). Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction* (pp. 286–295). Springer-Verlag.
- Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2011). When Off-Task in On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. *Proceedings of the 15th International Conference on Artificial Intelligence and Education* (pp. 534–536). Springer-Verlag.

- Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2013). Considering Alternate Futures to Classify Off-Task Behavior as Emotion Self-Regulation: A Supervised Learning Approach. *Journal of Educational Data Mining, Special Is.*
- Sabourin, J., Mott, B., & Lester, J. (2013). Discovering Behavior Patterns of Self-Regulated Learners in an Inquiry-Based Learning Environment. *In Proceedings of the 16th International Conference on Artificial Intelligence in Education*. Retrieved from <http://www4.ncsu.edu/~jlrobiso/papers/aied2013.pdf>
- Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2012a). Exploring Affect and Inquiry in Open-Ended Game-Based Learning Environments. *In Workshop on Emotions in Games for Learning in conjunction with the 11th International Conference on Intelligent Tutoring Systems*.
- Sabourin, J., Rowe, J. P., Mott, B. W., & Lester, J. C. (2012b). Exploring Inquiry-based Problem-Solving Strategies in Game-based Learning Environments. *Proceedings of the Eleventh International Conference on Intelligent Tutoring Systems* (pp. 470–475).
- Sabourin, J., Shores, L., Mott, B., & Lester, J. (2012). Predicting Student Self-Regulation Strategies in Game-Based Learning Environments. *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 470–475).
- Sabourin, J., & Lester, J. . (in press). Affect and Engagement in Game-Based Learning Environments. *Transactions on Affective Computing*, (Special Issue on Emotion in Games).
- Sabourin, J., Mott, B., & Lester, J. (2012). Early Prediction of Student Self-Regulation Strategies by Combining Multiple Models. *Proceedings of the Fifth International Conference on Educational Data Mining* (pp. 156–159).
- Sabourin, J., Mott, B., & Lester, J. (2013). Utilizing Dynamic Bayes Nets to Improve Early Prediction Models of Self-Regulated Learning. *Proceedings of the 21st International Conference on User Modeling, Adaptation and Personalization* (pp. 228–241).
- Sabourin, J., Shores, L. R., Mott, B., & Lester, J.. (in press). Understanding and Predicting Student Self-Regulated Learning Strategies in Game-Based Learning Environments. *International Journal of Artificial Intelligence in Education*, 23(1).
- Schraw, G. (2010). Measuring Self-Regulation in Computer-Based Learning Environments. *Educational Psychologist*, 45(4), 258–266.

- Schraw, G., Crippen, K., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*, 111–139.
- Schunk, D., & Swartz, C. (1993a). Writing strategy instruction with gifted students: Effects of goals and feedback on self-efficacy and skills. *Roeper Review, 15*(4), 225–230.
- Schunk, D., & Swartz, C. W. (1993b). Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology, 18*(3), 337–354.
- Schunk, D., & Zimmerman, B. (2003). Self-regulation and learning. In W. M. Reynolds & G. E. Miller (Eds.), (Vol. 7, pp. 59–78). New York, NY: Wiley & Sons.
- Schunk, D. & Ertmer, P. (2000). Self-Regulation and Academic Learning: Self-Efficacy Enhancing Interventions. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 631–646).
- Shaffer, D. (2006). *How computer games help children learn*. New York, NY: Palgrave Macmillan.
- Shores, L., Rowe, J., & Lester, J. (2011). Early Prediction of Cognitive Tool Use in Narrative-Centered Learning Environments. *Proceedings of the Fifteenth International Conference on Artificial Intelligence in Education*. Auckland, New Zealand.
- Shute, V. (2011). Stealth Assessment in Computer-Based Games to Support Learning. *Computer Games and Instruction* (pp. 503–523). Information Age Publishing.
- Shute, V., & Ventura, M. (2013). *Measuring and Supporting Learning in Games: Stealth Assessment*. Cambridge, MA: The MIT Press.
- Tang, T., & McCalla, G. (2004). On the Pedagogically Guided Paper Recommendation for an Evolving Web-Based Learning System. *Proceedings of the 17th Annual Florida Artificial Intelligence Research Society Conference*.
- Vanlehn, K. (2006). The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265.
- Veermans, K., van Joolingen, W., & de Jong, T. (2000). Promoting Self Directed Learning in Sumulation Based Discovery Learning Environments through Intelligent Support. *Interactive Learning Environments, 8*(3), 229–255.

- White, B., & Frederiksen, J. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. *Cognition & Instruction, 16*(1), 3–118.
- Winne, P. (2010). Improving Measurements of Self-Regulated Learning. *Educational Psychologist, 45*(4), 267–276.
- Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), (pp. 227–304). Mahwah, NJ: Erlbaum.
- Winne, P., & Perry, N. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), (pp. 532–546). San Diego, CA: Academic Press.
- Witmer, B., & Singer, M. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments, 7*(3), 225–240.
- Woolf, B. (2009). *Building Intelligent Interactive Tutors: Student-centered strategies for revolutionizing e-learning*. Burlington, MA: Morgan Kaufmann Publishers Inc.
- Woolf, B., Murray, T., Marshall, D., Dragon, T., Kohler, K., Mattingly, M., Bruno, M., et al. (2005). Critical Thinking Environments for Science Education. *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (pp. 515–522).
- Young, J. D. (1996). The Effect of Self-Regulated Learning Strategies on Performance in Learner Controlled Computer-Based Instruction. *Educational Technology Research and Development, 44*(2), 17–27.
- Zimmerman, B. (2008). Goal Setting: A Key Proactive Source of Academic Self-Regulation. In Dale H. Schunk & B. J. Zimmerman (Eds.), *Motivation and Self-Regulated Learning: Theory, Research, and Applications* (pp. 267–286). New York: Routledge.
- Zimmerman, B. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*, 3–17.
- Zimmerman, B. (2000). Attaining self-regulation: A social cognitive perspective. In and M. Z. M. Boekaert, P. R. Pintrich (Ed.), *Handbook of self-regulated learning* (pp. 13–39). San Diego, CA: Academic Press.
- Zimmerman, B., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology, 94*(4), 660–668.

## **APPENDICES**

## Appendix A – SRL Tagging Protocol

There are four labels, three of which represent behavior that may be a type of self-regulation but at different levels of sophistication. The final fourth group is a catch-all category.

**Specific Reflection**– This label is for statements in which students mention a specific goal or obstacle or mental belief. It may be positive or negative in valence but must reflect a specific consideration. This may include plans for future actions if it is phrased in a way that suggests that the student is setting goals. For example, “I guess I need to run more tests” would reference a specific goal whereas “About to run tests” is just a statement of what the student is doing. Other examples:

- I need to know more about viruses
- I can’t figure out the diagnosis worksheet
- I think it is the eggs
- I know the disease, but what is the source
- Where is Bryce?

**General Reflection** – The student is still evaluating their knowledge or progress but they do not make reference to a specific goal. They must provide additional information above and beyond the emotion self-report, therefore “confused” does not count as a general comment. Additionally referencing the goal of “solving the mystery” or “finding out why they are sick” is not a specific goal as this is the goal of the entire interaction. Students may reference feeling like they are almost finished, but without reference to what is left remaining or what they have recently accomplished that causes them to feel this way, it is still a general statement.

Other examples:

- I don’t know what to do
- I wonder what happened
- I’m almost there
- I figured it out!
- This is hard

**Non-reflective Statement** – This is a statement of what the student is doing or what is happening in the environment without referencing mental states or goals. If it is information that you could get from a log, it probably fits the non-reflective category. Examples:

- Talking to people
- Running tests
- The egg tested positive
- Looking for more items

**Unrelated** – This category is the catch-all for statements that do not fit within the others. This also includes statements that are non-word (ex. “!!!”, “arghh”), are not related to the game play (ex. “I’m hungry”, “My back hurts” ) or are repeats of the affective states without additional information. Additionally, if students comment on aspects of the game that are not related to what they are objectively doing or a cognitive state these also fall in the catch-all group. Examples:

- This game is fun
- I’m really excited
- Leave me alone, Kim!!

**When in doubt:**

If you’re debating between General and Specific Reflection, consider if you can tell exactly what part of the mystery the student is working on and if you could provide some guidance based on what they’ve said. If you know enough about their current goal or frustration then it should be labeled Specific. If you can’t really tell what aspect of the game they are at or how you could help, it’s probably a General statement.

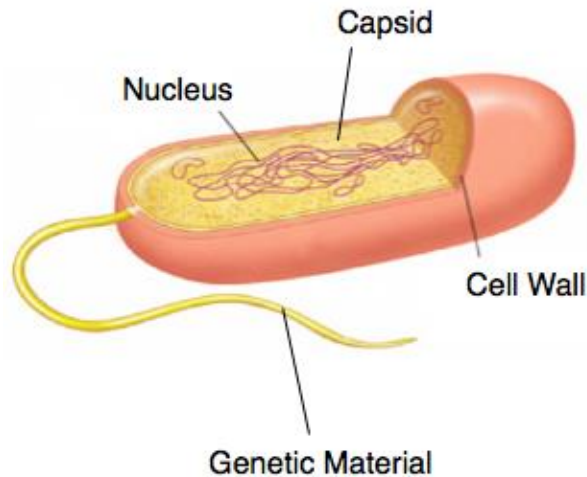
If you’re debating between Specific Reflection and Non-reflective, consider if there is evidence of the student doing some type of evaluative behavior. If the statement seems like it could be a response to the question “What are you doing right now?” then it’s probably a Non-reflective statement. If it is closer to a response to “How are you progressing in the game?” then it’s probably Specific.

**Appendix B – Content Test for CRYSTAL ISLAND Study**

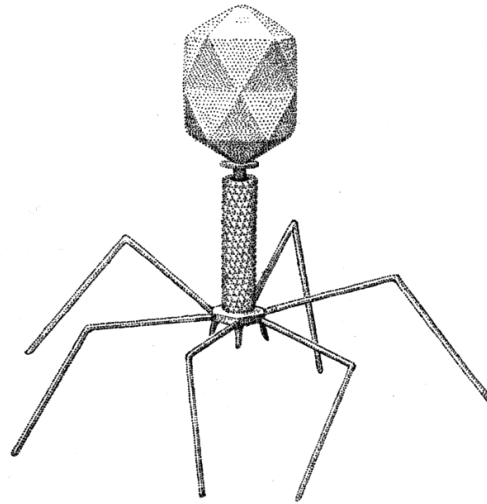
1. Which of the following statements best describes bacteria and viruses?
  - a. Bacteria and viruses are BOTH considered alive.
  - b. Bacteria are considered alive, but viruses are NOT considered alive.
  - c. Viruses are considered alive, but bacteria are NOT considered alive.
  - d. Viruses are NOT considered alive, and bacteria are NOT considered alive.
  
2. Which of the following sequences is in order from smallest size to largest size?
  - a. Bacteria, viruses, human hair
  - b. Human hair, bacteria, viruses
  - c. Viruses, bacteria, human hair
  - d. Viruses, human hair, bacteria
  
3. You place a biological agent under an electron microscope and observe that it does NOT have a nucleus. What type of agent might you be looking at?
  - a. Bacterium
  - b. Carcinogen
  - c. Virus
  - d. Either a virus or a bacterium
  
4. Your friend is feeling ill and goes to the doctor. The doctor gives your friend an anti-biotic and as a result your friend begins to feel better soon afterwards. What infectious agent likely made your friend sick?
  - a. Bacterium
  - b. Carcinogen
  - c. Virus
  - d. Either a virus or a bacterium
  
5. Your lab partners are examining a pathogen through a microscope and have observed that it is smooth and round in shape. What pathogen are your lab partners probably looking at?
  - a. Bacterium
  - b. Carcinogen
  - c. Virus
  - d. Either a virus and a bacterium

6. Which of the following statements about viruses is true?
  - a. Viruses are considered the smallest living cells.
  - b. Viruses consist of genetic material within a capsid.
  - c. Viruses reproduce through binary fission.
  - d. Virus specimens can be viewed through an optical microscope.
  
7. Which of the following diseases is caused by a bacterial infection?
  - a. Ebola Hemorrhagic Fever
  - b. Influenza
  - c. Salmonellosis
  - d. Smallpox
  
8. You are part of a team of scientists investigating an illness that may have been caused by contaminated food. Your team has asked the question, "What food source could be causing the illness?" What step should be taken next?
  - a. Formulate a hypothesis about the source of the illness.
  - b. Gather information about the spreading illness.
  - c. Perform a test to identify the food source.
  - d. Report findings about the cause of the illness.
  
9. What type of pathogen is considered the smallest living microorganism?
  - a. Bacterium
  - b. Carcinogen
  - c. Fungi
  - d. Virus
  
10. An illness has been spreading through a town, and doctors have told sick people to remain at home until they are well. The spread of the illness appears to have been stopped due to the doctors' instructions. What does this tell you must be true about the illness?
  - a. It was caused by a mutagen because it spread from one organism to another.
  - b. It was caused by a pathogen because it spread from one organism to another.
  - c. It was caused by a carcinogen because it spread from one organism to another.
  - d. It was not caused by a mutagen, carcinogen, or pathogen.
  
11. Which of the following statements about pathogens is true?
  - a. Any pathogen can be treated with antibiotics.
  - b. Pathogens are considered living microorganisms.
  - c. Pathogens are only responsible for a few hundred deaths each year.
  - d. Pathogens spread from person to person.

12. You have determined that your patients' disease has been caused by a genetic mutation. Knowing this you can determine that the disease was caused by
- A bacterium.
  - A virus.
  - A mutagen.
  - None of the above.
13. Which of the following treatments is generally considered the most effective way to reduce the likelihood of a viral infection?
- Antibiotic
  - Chemotherapy
  - Surgery
  - Vaccine
14. Your friend began to feel sick this morning and is showing the following symptoms: stomach cramps, fever, and severe diarrhea. She suspects that the source of her illness was some suspicious-looking hamburger meat she ate yesterday. Which of the following diseases is she likely suffering from?
- Anthrax
  - Botulism
  - Influenza
  - Salmonellosis
15. Bacteria come in several different shapes, but they share common structural characteristics. An illustration of one type of bacterium is shown below. Which of the illustration's labels is CORRECT?
- Cell Wall
  - Nucleus
  - Genetic Material
  - Capsid



16. Which of the following statements about viruses and bacteria is TRUE?
- a. All viruses and bacteria are harmful to humans.
  - b. All viruses and bacteria are considered pathogens.
  - c. Both viruses and bacteria are composed of small cells.
  - d. Some viruses and bacteria are not harmful to humans.
17. Which of the following characteristics do optical microscopes have in common with electron microscopes?
- a. They both can achieve magnifications of 1,000,000x.
  - b. They both can be used to view living specimens.
  - c. They both can be used to view parts of bacteria specimens.
  - d. They both use light to produce images.
18. An illustration of one type of infectious agent is shown below. Note that its shape resembles a lunar landing pod. Which type of infectious agent does the illustration most likely represent?
- a. Bacterium
  - b. Mutagen
  - c. Virus
  - d. Either a bacterium or a virus.



19. Which of the following diseases can cause black skin lesions?
- a. Anthrax
  - b. Ebola Hemorrhagic Fever
  - c. Influenza
  - d. Salmonellosis

## Appendix C – Pre-Experiment Materials for CRYSTAL ISLAND Study

### Demographic and Game-Playing Experience

1. What is your gender?
  - Male
  - Female
2. What is your age? (free response)
3. What is your race?
  - American Indian or Alaska Native
  - Asian
  - Black or African American
  - Hispanic or Latino
  - Native Hawaiian or Other Pacific Islander
  - White
  - Other (please specify)
4. How frequently do you play video games?
  - Not at all
  - Rarely
  - Occasionally
  - Frequently
  - Very frequently
5. How skilled are you when playing video games?
  - Not at all
  - Limited skills
  - Average
  - Skilled
  - Very skilled
6. How many hours per week do you typically play video games?
  - 0 - 2 hours
  - 2 - 5 hours
  - 5 - 10 hours
  - 10 - 20 hours
  - Over 20 hours

7. If you do play video games, please list up to three games that you enjoy playing.(free response)

### Achievement Goals Questionnaire

Please indicate your opinion about each of the statements below in reference to your current situation

|              |   |   |   |   |   |   |               |
|--------------|---|---|---|---|---|---|---------------|
| Almost Never |   |   |   |   |   |   | Almost Always |
| 1            | 2 | 3 | 4 | 5 | 6 | 7 |               |

It is important for me to do better than other students.

It is important for me to do well compared to others completing this exercise.

My goal in completing this exercise to get a better score than most of the other students.

I worry that I may not learn all that I possibly could while completing this exercise.

Sometimes I'm afraid that I may not understand the content of things as thoroughly as I'd like.

I am often concerned that I may not learn all that there is to learn.

I want to learn as much as possible from this exercise.

It is important for me to understand the content of this exercise as thoroughly as possible.

I desire to completely master this exercise.

I just want to avoid doing poorly while completing this exercise.

My goal during this exercise is to avoid performing poorly.

My fear of performing poorly is what motivates me.

### Big Five Inventory

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who *likes to spend time with others*? Please write a number next to each statement to indicate the extent to which **you agree or disagree with that statement**.

| 1                    | 2                    | 3                             | 4                 | 5                 |
|----------------------|----------------------|-------------------------------|-------------------|-------------------|
| Disagree<br>strongly | Disagree<br>a little | Neither agree<br>nor disagree | Agree<br>a little | Agree<br>strongly |

#### I am someone who...

1. \_\_\_\_\_ Is talkative
2. \_\_\_\_\_ Tends to find fault with others
3. \_\_\_\_\_ Does a thorough job
4. \_\_\_\_\_ Is depressed, blue
5. \_\_\_\_\_ Is original, comes up with new ideas
6. \_\_\_\_\_ Is reserved
7. \_\_\_\_\_ Is helpful and unselfish with others
8. \_\_\_\_\_ Can be somewhat careless
9. \_\_\_\_\_ Is relaxed, handles stress well.
10. \_\_\_\_\_ Is curious about many different things
11. \_\_\_\_\_ Is full of energy
12. \_\_\_\_\_ Starts quarrels with others
13. \_\_\_\_\_ Is a reliable worker
14. \_\_\_\_\_ Can be tense
15. \_\_\_\_\_ Is ingenious, a deep thinker
16. \_\_\_\_\_ Generates a lot of enthusiasm
17. \_\_\_\_\_ Has a forgiving nature
18. \_\_\_\_\_ Tends to be disorganized
19. \_\_\_\_\_ Worries a lot
20. \_\_\_\_\_ Has an active imagination
21. \_\_\_\_\_ Tends to be quiet
22. \_\_\_\_\_ Is generally trusting
23. \_\_\_\_\_ Tends to be lazy
24. \_\_\_\_\_ Is emotionally stable, not easily upset
25. \_\_\_\_\_ Is inventive
26. \_\_\_\_\_ Has an assertive personality
27. \_\_\_\_\_ Can be cold and aloof
28. \_\_\_\_\_ Perseveres until the task is finished
29. \_\_\_\_\_ Can be moody
30. \_\_\_\_\_ Values artistic, aesthetic experiences
31. \_\_\_\_\_ Is sometimes shy, inhibited

32. \_\_\_\_\_ Is considerate and kind to almost everyone
33. \_\_\_\_\_ Does things efficiently
34. \_\_\_\_\_ Remains calm in tense situations
35. \_\_\_\_\_ Prefers work that is routine
36. \_\_\_\_\_ Is outgoing, sociable
37. \_\_\_\_\_ Is sometimes rude to others
38. \_\_\_\_\_ Makes plans and follows through with them
39. \_\_\_\_\_ Gets nervous easily
40. \_\_\_\_\_ Likes to reflect, play with ideas
41. \_\_\_\_\_ Has few artistic interests
42. \_\_\_\_\_ Likes to cooperate with others
43. \_\_\_\_\_ Is easily distracted
44. \_\_\_\_\_ Is sophisticated in art, music, or literature

### Cognitive Emotion Regulation Questionnaire – Short

When faced with a difficult situation in school (e.g., poor performance on an exam)...

|              |   |           |   |              |
|--------------|---|-----------|---|--------------|
| Almost Never |   | Sometimes |   | Almost Never |
| 1            | 2 | 3         | 4 | 5            |

I feel that I am the one who is responsible for what has happened

I think that I have to accept that this has happened

I often think about what I have experienced

I think of pleasant things that have nothing to do with it

I think about how to change the situation

I think I can learn something from the situation

I think that it hasn't been too bad compared to other things

I keep thinking about how terrible what I have experienced is

I think that basically the cause must lie within myself

I think that I have to accept the situation

I am preoccupied with what I think and feel about what I have experienced

I think of something nice instead of what has happened

I think about a plan of what I can do better

I think that I can become a stronger person as a result of what has happened

I tell myself that there are worse things in life

I continually think how horrible the situation has been

## Appendix D – Post-Experiment Materials for CRYSTAL ISLAND Study

### Intrinsic Motivation Inventory

For each of the following statements, please indicate how true it is for you, using the following scale:

|                    |   |   |                  |   |   |              |
|--------------------|---|---|------------------|---|---|--------------|
| 1                  | 2 | 3 | 4                | 5 | 6 | 7            |
| not at<br>all true |   |   | somewhat<br>true |   |   | very<br>true |

#### Interest/Enjoyment

I enjoyed doing this activity very much.

This activity was fun to do.

I thought this was a boring activity.

This activity did not hold my attention at all.

I would describe this activity as very interesting.

I thought this activity was quite enjoyable.

While I was doing this activity, I was thinking about how much I enjoyed it.

#### Perceived Competence

I think I am pretty good at this activity.

I think I did pretty well at this activity, compared to other students.

After working at this activity for awhile, I felt pretty competent.

I am satisfied with my performance at this task.

I was pretty skilled at this activity.

This was an activity that I couldn't do very well.

#### Effort/Importance

I put a lot of effort into this.

I didn't try very hard to do well at this activity.

I tried very hard on this activity.

It was important to me to do well at this task.

I didn't put much energy into this.

#### Pressure/Tension

I did not feel nervous at all while doing this.

I felt very tense while doing this activity.

I was very relaxed in doing this task.

I was anxious while working on this task.

I felt pressured while doing this task.

**Value/Usefulness**

I believe this activity could be of some value to me.

I think that doing this activity is useful for learning science.

I think this is important to do because it can help me learn science.

I would be willing to do this again because it has some value to me.

I think doing this activity could help me to learn science.

I believe doing this activity could be beneficial to me.

I think this is an important activity.

### Presence Questionnaire

Characterize your experience in the environment, by marking an "X" in the appropriate box of the 7-point scale, in accordance with the question content and descriptive labels. Please consider the entire scale when making your responses, as the intermediate levels may apply. Answer the questions independently in the order that they appear. Do not skip questions or return to a previous question to change your answer.

#### With regard to the environment:

1. How much were you able to control events?

|            |       |       |          |       |       |            |
|------------|-------|-------|----------|-------|-------|------------|
| _ _ _      | _ _ _ | _ _ _ | _ _ _    | _ _ _ | _ _ _ | _ _ _      |
| NOT AT ALL |       |       | SOMEWHAT |       |       | COMPLETELY |

2. How responsive was the environment to actions that you initiated (or performed)?

|                   |       |       |                          |       |       |                          |
|-------------------|-------|-------|--------------------------|-------|-------|--------------------------|
| _ _ _             | _ _ _ | _ _ _ | _ _ _                    | _ _ _ | _ _ _ | _ _ _                    |
| NOT<br>RESPONSIVE |       |       | MODERATELY<br>RESPONSIVE |       |       | COMPLETELY<br>RESPONSIVE |

3. How natural did your interactions with the environment seem?

|                         |       |       |           |       |       |                       |
|-------------------------|-------|-------|-----------|-------|-------|-----------------------|
| _ _ _                   | _ _ _ | _ _ _ | _ _ _     | _ _ _ | _ _ _ | _ _ _                 |
| EXTREMELY<br>ARTIFICIAL |       |       | BODERLINE |       |       | COMPLETELY<br>NATURAL |

4. How much did the visual aspects of the environment involve you?

|            |       |       |          |       |       |            |
|------------|-------|-------|----------|-------|-------|------------|
| _ _ _      | _ _ _ | _ _ _ | _ _ _    | _ _ _ | _ _ _ | _ _ _      |
| NOT AT ALL |       |       | SOMEWHAT |       |       | COMPLETELY |

5. How much did the auditory aspects of the environment involve you?

|            |       |       |          |       |       |            |
|------------|-------|-------|----------|-------|-------|------------|
| _ _ _      | _ _ _ | _ _ _ | _ _ _    | _ _ _ | _ _ _ | _ _ _      |
| NOT AT ALL |       |       | SOMEWHAT |       |       | COMPLETELY |

6. How natural was the mechanism which controlled movement through the environment?

|                         |       |       |           |       |       |                       |
|-------------------------|-------|-------|-----------|-------|-------|-----------------------|
| _ _ _                   | _ _ _ | _ _ _ | _ _ _     | _ _ _ | _ _ _ | _ _ _                 |
| EXTREMELY<br>ARTIFICIAL |       |       | BODERLINE |       |       | COMPLETELY<br>NATURAL |

7. How compelling was your sense of objects moving through space?

|                          |       |       |                          |       |       |                    |
|--------------------------|-------|-------|--------------------------|-------|-------|--------------------|
| _ _ _                    | _ _ _ | _ _ _ | _ _ _                    | _ _ _ | _ _ _ | _ _ _              |
| NOT AT ALL<br>COMPELLING |       |       | MODERATELY<br>COMPELLING |       |       | VERY<br>COMPELLING |

8. How much did your experiences in the virtual environment seem consistent with your real world experiences?

|            |  |            |  |  |            |  |
|------------|--|------------|--|--|------------|--|
|            |  |            |  |  |            |  |
| NOT AT ALL |  | MODERATELY |  |  | VERY       |  |
| CONSISTENT |  | CONSISTENT |  |  | CONSISTENT |  |

9. Were you able to anticipate what would happen next in response to the actions that you performed?

|            |  |          |  |  |            |  |
|------------|--|----------|--|--|------------|--|
|            |  |          |  |  |            |  |
| NOT AT ALL |  | SOMEWHAT |  |  | COMPLETELY |  |

10. How completely were you able to actively survey or search the environment using vision?

|            |  |          |  |  |            |  |
|------------|--|----------|--|--|------------|--|
|            |  |          |  |  |            |  |
| NOT AT ALL |  | SOMEWHAT |  |  | COMPLETELY |  |

11. How well could you identify sounds?

|            |  |          |  |  |            |  |
|------------|--|----------|--|--|------------|--|
|            |  |          |  |  |            |  |
| NOT AT ALL |  | SOMEWHAT |  |  | COMPLETELY |  |

12. How well could you localize sounds?

|            |  |          |  |  |            |  |
|------------|--|----------|--|--|------------|--|
|            |  |          |  |  |            |  |
| NOT AT ALL |  | SOMEWHAT |  |  | COMPLETELY |  |

13. How well could you actively survey or search the virtual environment using touch?

|            |  |          |  |  |            |  |
|------------|--|----------|--|--|------------|--|
|            |  |          |  |  |            |  |
| NOT AT ALL |  | SOMEWHAT |  |  | COMPLETELY |  |

14. How compelling was your sense of moving around inside the virtual environment?

|            |  |            |  |  |            |  |
|------------|--|------------|--|--|------------|--|
|            |  |            |  |  |            |  |
| NOT        |  | MODERATELY |  |  | VERY       |  |
| COMPELLING |  | COMPELLING |  |  | COMPELLING |  |

15. How closely were you able to examine objects?

|            |  |                |  |  |              |  |
|------------|--|----------------|--|--|--------------|--|
|            |  |                |  |  |              |  |
| NOT AT ALL |  | PRETTY CLOSELY |  |  | VERY CLOSELY |  |

16. How well could you examine objects from multiple viewpoints?

|            |  |          |  |  |             |  |
|------------|--|----------|--|--|-------------|--|
|            |  |          |  |  |             |  |
| NOT AT ALL |  | SOMEWHAT |  |  | EXTENSIVELY |  |



25. How completely were your senses engaged in this experience?

NOT MILDLY COMPLETELY  
 ENGAGED ENGAGED ENGAGED

26. How easy was it to identify objects through physical interaction; like touching an object, walking over a surface, or bumping into a wall or object?

IMPOSSIBLE MODERATELY VERY EASY  
 DIFFICULT

27. Were there moments during the virtual environment experience when you felt completely focused on the task or environment?

NONE OCCASIONALLY FREQUENTLY

28. How easily did you adjust to the control devices used to interact with the virtual environment?

DIFFICULT MODERATE EASILY

29. Was the information provided through different senses in the virtual environment (e.g., vision, hearing, touch) consistent?

NOT MODERATELY VERY  
 CONSISTENT CONSISTENT CONSISTENT

30. To what extent did you feel completely surrounded by and enveloped by the virtual environment?

NOT AT ALL SOME EXTENT VERY MUCH

31. As you moved through the virtual environment and interacted with it, did you feel like you were inside the virtual environment, affecting or being affected by objects and events in that environment.

NOT AT ALL SOMEWHAT COMPLETELY

32. How much did your experience in the virtual environment seem like you were in a real place, able to directly sense and interact with the environment?

NOT AT ALL MODERATELY SO VERY MUCH

33. In the virtual environment, how strong was your sense of “being there”?

|        |  |  |            |  |  |        |
|--------|--|--|------------|--|--|--------|
|        |  |  |            |  |  |        |
| NOT    |  |  | MODERATELY |  |  | VERY   |
| STRONG |  |  | STRONG     |  |  | STRONG |

## Appendix E – Handout Materials for CRYSTAL ISLAND Study

### Crystal Island Storyworld

The CRYSTAL ISLAND storyworld is situated on a recently discovered volcanic island where a research station (Figure 1) has been established to study the island’s unique flora and fauna. There are eight main characters in the CRYSTAL ISLAND storyworld (see the “Crystal Island Characters” handout): Robert Campbell (bacteria specialist), Elise Johnson (lab technician), Kim Lee (camp nurse), Teresa Moore (senior scientist), Quentin Nash (cook and custodian), Alex Reid (player), Bryce Reid (lead scientist), and Ford Patterson (virus specialist). The user plays the role of Alex Reid visiting her father, Bryce Reid, who serves as the research station’s lead scientist.



**Figure 1: Crystal Island Research Station**

The research camp includes the following buildings (see the “Crystal Island Virtual Environment Map” handout): *Bryce’s Quarters*, the *Dining Hall*, the *Infirmary*, the *Laboratory*, and the *Living Quarters*. There is also a *waterfall* at one end of the camp.

The CRYSTAL ISLAND virtual environment features a science mystery in which the user plays the role of a “medical detective”. As members of the research team fall ill, it is the user’s task to discover the cause of the outbreak and its source. Solve the mystery before you, and the team, run out of time!

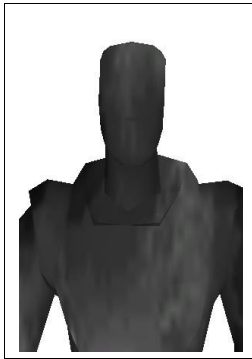
You will be given 50 minutes to interact with the CRYSTAL ISLAND virtual environment.

Feel free to use the “Interacting with the Virtual Environment” handout (describing the keyboard and mouse controls), as well as the “Crystal Island Virtual Environment Map” handout and the “Crystal Island Characters” handout, throughout your interaction.

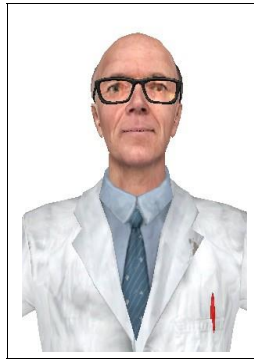
If you have any questions at this time, please ask.

## Crystal Island Characters

**Alex**  
(User)



**Bryce**  
(Lead Scientist)



**Elise**  
(Lab Technician)



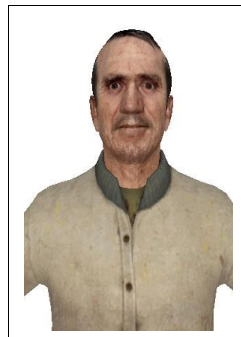
**Ford**  
(Virus Specialist)



**Kim**  
(Nurse)



**Quentin**  
(Cook/Custodian)



**Robert**  
(Bacteria Specialist)



**Teresa**  
(Senior Scientist)



## Interacting with the Virtual World

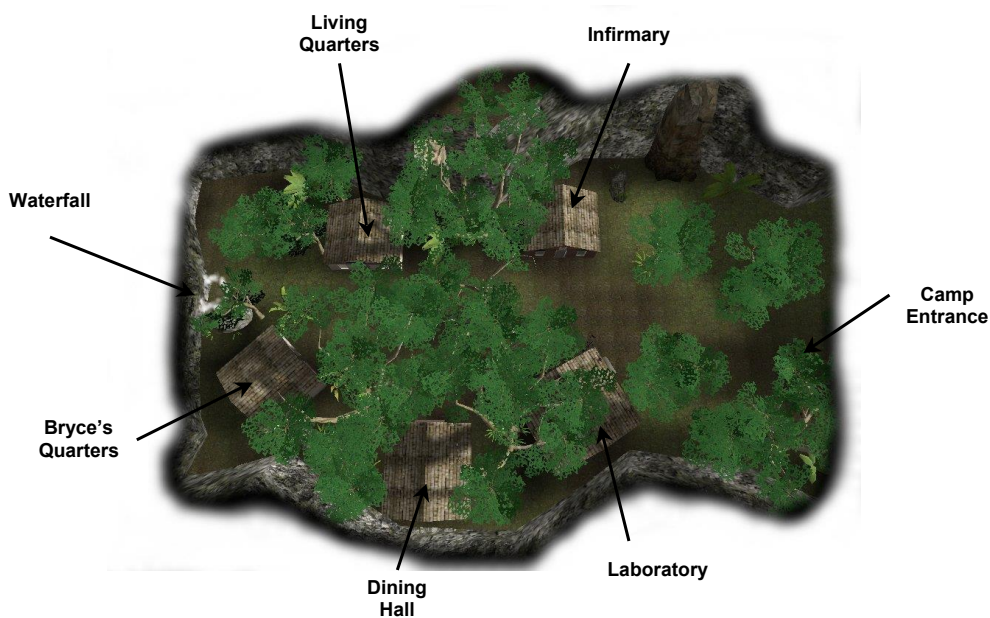
### Keyboard Controls

|                                  |                          |                                |
|----------------------------------|--------------------------|--------------------------------|
| <b>W</b> – Diagnosis worksheet   |                          |                                |
| <b>C</b> – Handheld Communicator | <b>↑</b> – Move Forward  |                                |
| <b>←</b> – Move Left (strafe)    | <b>↓</b> – Move Backward | <b>→</b> – Move Right (strafe) |

### Mouse Controls

|  |   |
|--|---|
| <b>Left Mouse Button</b> – Action Button | <b>Right Mouse Button</b> – Stow/retrieve from backpack |
| <b>Mouse Up</b> – Look Up                |   |
| <b>Mouse Left</b> – Turn Left            | <b>Mouse Right</b> – Turn Right                         |
| <b>Mouse Down</b> – Look Down            |   |

### Crystal Island Virtual Environment Map



|                          |  |                         |                                  |
|--------------------------|--|-------------------------|----------------------------------|
| <b>Waterfall:</b>        | A nice place to cool off.                  | <b>Laboratory:</b>      | Research lab facilities          |
| <b>Bryce's Quarters:</b> | Living quarters for Bryce, lead scientist. | <b>Living Quarters:</b> | Living quarters for team members |
| <b>Infirmary:</b>        | Medical facilities for the camp            | <b>Camp Entrance:</b>   | Gateway to camp                  |
| <b>Dining Hall:</b>      | Camp's dining facilities                   |                         |                                  |