

ABSTRACT

CARPENTER, DAN. Natural Language Analysis of Student Reflection in Game-Based Learning. (Under the direction of Dr. James Lester).

Reflection is a critical skill for 21st century learners. As learning continues to become increasingly autonomous with the rise of digital learning platforms, such as game-based learning environments, students must effectively regulate their own learning processes to achieve the desired learning outcomes. This self-regulated approach to learning relies on students' ability to engage in meaningful reflection, where they look back on their past learning and usage of learning strategies to evaluate the progress made toward their goals and to inform adaptations for future learning. However, many students do not have strong reflection skills, and these skills are often not explicitly taught in the classroom. Thus, there is a need for educational systems that can help students acquire reflection skills. Integrating reflection into a game-based learning environment presents an opportunity to collect data that captures evidence of students' reflection behaviors during complex learning scenarios that require effective self-regulation to achieve success. Then, leveraging recent advances in machine learning and natural language processing, we can construct automated systems that analyze this data in real time to drive adaptive scaffolds for reflection during game-based learning.

This dissertation introduces a natural language processing-based framework for automatically analyzing students' reflections during game-based learning in terms of how effectively they reflect and what they choose to reflect on. The framework leverages recent advances in natural language processing, namely Transformer-based language models, to investigate the reflection behaviors of 153 middle school students as they interacted with CRYSTAL ISLAND, a game-based learning environment for middle school microbiology. First, we explore relationships between students' reflection behaviors and their learning outcomes to empirically evaluate the theoretical benefits of reflection on learning. To this end, we operationalize a measure of reflection depth during science problem solving, extract common topics of students' reflections, align students' reflections with scientific inquiry processes to support a broader exploration of the relationship between students' reflection behaviors and their science problem-solving strategies, and investigate the extent to which students' reflection behaviors change over time as they interact with CRYSTAL ISLAND. Next, we perform contextual analyses that elucidate the relationship between students' learning activities in CRYSTAL ISLAND

and their reflection behaviors. Most of the existing work done to analyze students' reflection behaviors deals with reflections written in a post-learning setting, so improving our understanding of the ways in which students' reflection behaviors are related to their learning activities addresses a critical gap in the current literature. Finally, we leverage Transformer-based language models to train robust machine learning models for automatically assessing the depth of students' reflections. Given the small size of our reflection dataset and the high cost of collecting additional reflection data, we explore the use of large language models to synthesize new reflections and find that models trained on synthetic reflections with human-assigned labels significantly improve predictive performance compared to models that only used genuine student reflections.

This dissertation presents insights into students' reflection behaviors during game-based learning and the automated reflection analysis framework provides the foundation for AI-enabled adaptive support for reflection during game-based learning.

© Copyright 2023 by Dan Carpenter

All Rights Reserved

Natural Language Analysis of Student Reflection in Game-Based Learning

by
Dan Carpenter

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
2023

APPROVED BY:

Dr. James Lester II
Committee Chair

Dr. Jonathan Rowe

Dr. Collin Lynch

Dr. Shiyang Jiang

BIOGRAPHY

Dan Carpenter was born in Glens Falls, New York on July 22, 1996. He completed his undergraduate studies at Siena College in Loudonville, New York, earning degrees in computer science and mathematics in the spring of 2018. Later that year, in the fall of 2018, Dan fled the cold northeast weather to head to Raleigh, North Carolina in pursuit of a PhD in computer science at North Carolina State University. He was always a fan of learning and building things, so research in educational technology was a natural fit for him. Over the course of his PhD studies, he enjoyed designing and developing digital game-based learning environments and interacting with students in authentic classroom settings as they tested the technology and provided feedback with unmitigated honesty.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. James Lester, for all of his support and guidance during my PhD. He has been an excellent example for how to conduct research in a richly interdisciplinary field, which requires a huge breadth of knowledge and an ability to collaborate with individuals from so many diverse backgrounds. I have learned a lot from watching him successfully juggle such a wide range of projects. I would also like to thank the other members of my committee – Dr. Shiyan Jiang, Dr. Collin Lynch, and Dr. Jon Rowe – who helped shape the direction of this dissertation work. In particular, Jon has been involved with my research from the very beginning and has provided invaluable mentoring and guidance during my PhD.

I was lucky to work alongside a great group of fellow PhD students as I completed my degree in the IntelliMedia Group – Halim Acosta, Andrew Emerson, Fahmid Fahid, Alex Goslen, Anisha Gupta, Nathan Henderson, Jay Pande, and Kyungjin Park were wonderful colleagues who provided consistent support and fostered a fun and collaborative research environment. I also had the benefit of working with a first-rate group of researchers during my PhD – Yeojin Kim, Seung Lee, Wookhee Min, Brad Mott, Andy Smith, Randy Spain, and Jessica Vandenberg were stellar examples for how to conduct solid and impactful research. IntelliMedia also has an exceptional team of staff members who I had the pleasure of working with – Courtney Barron, Kara Cassell, Kirby Culbertson, Vikram Kumaran, Barry Liu, Justin Philips, Sarah Reaves, Rob Taylor, and Sandy Taylor generously provided support for a lot of my work. Finally, this work greatly benefited from many years of collaboration with an exemplary team of educational psychology researchers at the University of Central Florida – Roger Azevedo, Elizabeth Cloude, Daryn Dever, Michelle Taub, and Megan Wiedbusch were always there to support my journey into the world of self-regulated learning and metacognition.

TABLE OF CONTENTS

List of Tables	vii
List of Figures.....	ix
Chapter 1 Introduction.....	1
1.1 Thesis Statement and Hypotheses.....	4
1.2 Contributions	5
1.3 Organization.....	6
Chapter 2 Background and Related Work.....	7
2.1 Self-Regulated Learning.....	7
2.2 Learning Analytics for Self-Regulated Learning.....	9
2.3 Natural Language Processing	11
2.3.1 Natural Language Processing in Education.....	12
2.3.2 Automated Reflection Analysis.....	15
Chapter 3 Reflection-Enhanced Game-Based Learning.....	18
Chapter 4 CRYSTAL ISLAND: REFLECT Dataset.....	21
4.1 Participants	24
4.2 Study Procedure.....	24
Chapter 5 Exploring Students’ Reflection Behaviors during Game-Based Learning	25
5.1 Analyzing the Depth of Students’ Reflections.....	25
5.1.1 Examining the Predictive Relationship Between Gameplay Behaviors and Reflection Depth.....	28
5.2 Analyzing the Content of Students’ Reflections using Automated Reflection Topic Modeling.....	31
5.2.1 Method	32
5.2.2 Results	32

5.3 Identifying Links Between Reflections and In-Game Sources of Information.....	37
5.3.1 Method	37
5.3.2 Results	40
5.4 Examining Relationships Between Reflection Content and Reflection Depth.....	43
5.4.1 Method	43
5.4.2 Results	44
5.5 Investigating Relationships Between Reflection Behaviors and Learning Outcomes.....	46
5.5.1 Method	46
5.5.2 Results	48
5.6 Examining Changes in Students’ Reflection Behaviors Over Time	52
5.6.1 Method	52
5.6.2 Results	53
Chapter 6 Analyzing Students’ Reflections in the Context of their Gameplay	
Interactions.....	56
6.1 Sequential Pattern Mining to Identify Associations Between Gameplay Patterns and Reflection Content.....	57
6.1.1 Method	57
6.1.2 Results	58
6.2 Predicting Reflection Content Based on Gameplay Interactions	62
6.2.1 Method	62
6.2.2 Results	64
6.3 Automatically Linking Reflection Content to Sources of In-Game Information.....	66
6.3.1 Method	67
6.3.2 Results	68
Chapter 7 Automated Assessment of Reflection Depth using Data Augmentation and Transformer-Based Language Models	71

7.1 Comparison of Natural Language Representations for Predicting Reflection Depth	72
7.1.1 Method	72
7.1.2 Results	75
7.2 Predicting Components of Reflection Depth	78
7.2.1 Method	79
7.2.2 Results	81
7.3 Improving Automated Reflection Depth Assessment Models using Data Augmentation Techniques	83
7.3.1 Baseline Data Augmentation Techniques	84
7.3.2 Reflection Synthesis using Large Language Models.....	85
7.3.3 Exploring Synthetic Reflections	88
7.3.4 Predictive Modeling with Augmented Reflection Data	89
7.3.5 Improving Predictive Models with Manually Labeled Synthetic Reflections.....	92
Chapter 8 Conclusion	96
8.1 Hypotheses Revisited	96
8.2 Summary	99
8.3 Future Work	100
References	103

LIST OF TABLES

Table 4.1 Triggers used to prompt reflection.....	23
Table 5.1 Reflection depth rubric.....	27
Table 5.2 Hierarchical regression results for predicting average reflection depth. All regression coefficients are from the final step in the analyses.	29
Table 5.3 Descriptions of the reflection topic clusters identified by BERTopic.....	35
Table 5.4 Rubric used to label pairs of reflections and text representations of gameplay interactions. Underlined text indicates specific parts of the text that are considered to be related.	39
Table 5.5 Average reflection depth scores across phases of scientific inquiry.	44
Table 5.6 Average reflection depth scores across different in-game sources of information that were referenced in the reflection.	45
Table 5.7 Hierarchical regression for predicting post test score and whether the student solved the mystery. All regression coefficients are from the final step in the analyses.....	50
Table 5.8 Model performance for predicting students' learning outcomes using different sets of features derived from gameplay interactions and reflections.....	51
Table 5.9 Growth model parameter estimates with standard error in parentheses and fit indices. * $p < .05$	53
Table 6.1 Differentially frequent patterns – Experimental Design vs Other	60
Table 6.2 Differentially frequent patterns - Evidence Evaluation vs Other	60
Table 6.3 Differentially frequent patterns – Evidence Evaluation vs Experimental Design	61
Table 6.4 Differentially frequent patterns - Poster vs Other.....	61
Table 6.5 Results for predicting the type of gameplay interaction that a student would reflect on based on their recent gameplay interactions.	64
Table 6.6 Results for predicting the scientific inquiry phase of a student's next reflection based on their recent gameplay interactions.	65

Table 7.1 Accuracy scores for reflection depth assessment models.	75
Table 7.2 F1 scores for reflection depth assessment models.	76
Table 7.3 Quadratic weighted kappa scores for reflection depth assessment models.....	78
Table 7.4 Rubric for components of reflection depth.....	80
Table 7.5 Accuracy for reflection depth component detection models.	81
Table 7.6 F1 score for reflection depth component detection models.....	82
Table 7.7 Depth assessment results for FLAN-T5 and BERT-family models.	89

LIST OF FIGURES

Figure 3.1 Overview of a reflection-enhanced game-based learning environment.	18
Figure 3.2 Overview of the proposed automated reflection analysis framework for a game-based learning environment.	19
Figure 4.1 CRYSTAL ISLAND game-based learning environment.	21
Figure 4.2 Embedded reflection prompt.	22
Figure 5.1 Visualization of the reflection topic clusters identified by BERTopic. Smaller greyed out data points represent reflections that were considered outliers.	34
Figure 5.2 Distribution of types of gameplay interactions that students reflected on based on splits between high versus low normalized learning gain (NLG) and solving versus not solving the mystery.	42
Figure 5.3 Students' reflection depth over the course of their interaction with CRYSTAL ISLAND. Darker lines represent more common trajectories, and the red line represents the model implied trajectory over all students.	54
Figure 6.1 Prompt used to instruct FLAN-T5 on how to link students' reflections with relevant gameplay interactions.	67
Figure 6.2 Accuracy and F1 scores for automatically linking reflections and gameplay interactions using FLAN-T5.	69
Figure 7.1 Background information about CRYSTAL ISLAND that was included in the reflection synthesis prompt.	85
Figure 7.2 Prompt instructions for synthesizing new reflections based on the reflection depth rubric, examples of previous students' reflections, and a sample gameplay interaction sequence.	87
Figure 7.3 QWK for baseline data augmentation techniques across different amounts of synthetic data.	90
Figure 7.4 QWK for FLAN-T5 models trained with reflections generated via back translation and synthetic reflections created by GPT-3.5.	92

Figure 7.5 QWK for FLAN-T5 models trained with reflections generated via back translation, synthetic reflections created by GPT-3.5, and manually relabeled synthetic reflections.....95

CHAPTER 1

INTRODUCTION

Self-regulated learning (SRL) is a critical competency in 21st century education. Due to the increase in autonomy enabled by digital learning platforms (Carter et al., 2020), learners must be able to break down an open-ended learning scenario to identify problems that need to be addressed, set learning goals and construct effective plans for achieving them, enact learning strategies and continuously monitor progress toward those goals, and reflect on performance to drive adaptation for future learning. Unfortunately, these skills are complex and many students do not successfully employ them in their own learning (Greene & Azevedo, 2010). Thus, there is a need to develop systems and tools to support students in the development and deployment of self-regulated learning skills. By leveraging recent advances in natural language processing to automatically analyze students' self-regulated learning behaviors while interacting with a digital learning environment, there is potential to create adaptive systems that can provide scalable real-time support for the development of self-regulated learning skills.

A first step toward building systems to support self-regulated learning is to develop an understanding of how students utilize these skills while learning and to identify opportunities for filling in gaps in their usage of these skills. Think-aloud data has frequently been used to observe SRL, since it allows students to articulate their internal thoughts (Engelmann & Bannert, 2021; Greene & Azevedo, 2009; Sonnenberg & Bannert, 2016). Additionally, a substantial body of work has analyzed sequences of micro-level activities extracted from trace data collected in digital learning environments (e.g., clicking to interact with a tool or a resource) to obtain evidence of SRL skill usage (Fincham et al., 2019; Lust et al., 2013; Siadaty et al., 2016; Winne et al., 2019). Based on these micro-level sequences, learning analytics techniques like process mining can be used to identify SRL processes (Saint et al., 2020) and to relate those SRL processes to students' learning tactics and learning outcomes (Fan et al., 2021; Raković et al., 2022; Srivastava et al., 2022).

Beyond deriving data-driven insights to advance our understanding of SRL, learning analytics techniques have also been used to directly present students with information about their SRL processes. These learning analytics-based interventions often include a

recommendation of changes that the student should make as well as instructions for making those changes (Winne, 2017). The most common types of interventions include visual dashboards (Law et al., 2016; Molenaar et al., 2020) and personalized feedback (Jensen et al., 2021; Lim, Gentili, et al., 2021; Timmers et al., 2015; D. Wang & Han, 2021). By directly providing learning analytics-based information to students, these studies have begun to move toward adaptive support for SRL skills.

The work laid out in this dissertation specifically focuses on providing adaptive support for the SRL skill of reflection, which is both a backward and forward-looking process in which learners evaluate their knowledge and learning processes to inform adaptations to goals and strategies moving forward (Winne & Hadwin, 2008). While prior research has shown that reflection can lead to more effective and more efficient learning (Chen et al., 2016; Hsiao et al., 2017), it has also been observed that many students lack strong reflection skills (Cavilla, 2017; van Velzen, 2016). Game-based learning presents a unique opportunity to train SRL skills like reflection during learning, since it is possible to create complex problem-solving scenarios that are challenging to complete without the successful application of SRL skills. For example, authentic game-based learning scenarios with ill-defined problems and goals force students to explore on their own, determine what problems they need to solve, and identify goals and subgoals that they must accomplish in order to solve those problems. In such a scenario, periodic reflection is critical for success because learners must evaluate the progress they are making toward their goals, determine if their learning strategies are working, and decide whether to continue working toward the same goals or redirect toward new goals. By prompting learners to externalize their reflection processes during game-based learning, the learning environment can potentially analyze their reflections and provide adaptive support to each learner to help develop their reflection skills. This could in turn lead students to adopt more effective learning processes, thereby helping a larger number of students realize the full benefits of game-based learning.

Students' reflections are typically captured by asking the student to respond to an open-ended prompt (Ullmann, 2015). Reflections have been collected at specific points during a learning experience as a reflection on progress as well as after the learning experience has been completed as an overarching reflection on the learning process as a whole (Akbari, 2007). Then, leveraging recent advances in natural language processing, researchers have trained models to

automatically analyze and assess different aspects of students' reflections. These models often assess components of reflection breadth, such as description, analysis, feeling, perspective, motive, or goal (Jung & Wise, 2020; Kovanović et al., 2018). They also assess reflection depth, such as "not reflective", "somewhat reflective", or "critically reflective" (Barthakur et al., 2022; Ullmann, 2015). Automated approaches to these analyses include dictionary-based and rule-based approaches, where an expert manually identifies keywords or defines linguistic rules that are hypothesized to be indicative of some aspect of reflection (Barthakur et al., 2022; Gibson et al., 2017; Ullmann, 2015). More recently, researchers have investigated machine learning approaches using deep language model-based representations of students' reflections to automatically evaluate reflection depth (Magooda et al., 2022; Nehyba & Štefánik, 2022). Considering how Transformer-based language models, such as the BERT (Devlin et al., 2018) and GPT (Brown et al., 2020) families of language models, have dominated the field of natural language processing (NLP) in recent years, there is significant promise in this direction.

However, there exists a gap in the body of automated reflection analysis research in that it has mostly focused on post-secondary students who produce extended segments of reflective writing, such as essays, as a post-learning activity (Jung & Wise, 2020; Kovanović et al., 2018; Ullmann, 2019). In contrast, reflections from middle school students collected during game-based learning are much shorter and messier (e.g., grammatically incorrect, incorrect use of domain terminology, and frequent misspellings), thus presenting potential challenges for applying modern NLP techniques to this task. Since higher-order thinking skills such as reflection are increasingly becoming important even from a young age (Molenaar et al., 2019; Spires et al., 2011), there is a need for work that focuses on understanding and scaffolding younger students' reflective processes. Additionally, this data allows us to analyze students' reflections in the context of rich data capturing their learning processes during science problem solving, which presents an opportunity to better understand the relationship between students' learning activities and how they engage in mid-process reflection.

This dissertation presents research that addresses two main goals: (1) developing a deeper understanding of middle school students' reflection processes during game-based learning and (2) constructing a framework for automatically analyzing students' reflections to enable adaptive support for reflection during game-based learning. To achieve these goals, we have analyzed data collected from middle school students as they interacted with a reflection-enabled version

of CRYSTAL ISLAND, a game-based learning environment for microbiology. To address the first goal, the dissertation presents a learning analytics-based investigation into students' reflection processes during game-based learning. We have analyzed students' reflections to understand how reflection depth (i.e., ranging from non-reflective to highly reflective) relates to the central learning outcomes of CRYSTAL ISLAND. We have also analyzed the content of students' reflections relative to their recent gameplay interactions to identify any contextual factors that may impact students' reflection processes during game-based learning. To address the second goal, we introduce an NLP and machine learning-based framework for automatically analyzing students' written reflections. Predictive models leveraging recent Transformer-based NLP techniques were trained to automatically assess reflection depth and evaluate the topics that students addressed in their reflections given a specific gameplay context, thus enabling the creation of highly specific adaptive supports for reflection. Results from this work reveal significant predictive relationships between the depth and content of students' reflections and their learning outcomes, indicate that machine learning models can be trained to predict the content of students' reflections based on their gameplay behaviors, and demonstrate that data augmentation techniques can be used to create robust Transformer-based models that can be used to automatically assess the depth of students' reflections.

1.1 Thesis Statement and Hypotheses

This dissertation investigates the following thesis statement:

A natural language processing framework leveraging the Transformer architecture can be used to automatically analyze the depth and content of students' written reflections during game-based learning to provide insights into students' reflection behaviors and enable adaptive intervention.

To investigate this thesis statement, we evaluate the following hypotheses:

H.1. Reflection depth assessment models that incorporate Transformer-based natural language representations will outperform models that do not. This will be the case for predicting a single measure of reflection depth as well as for identifying individual components that make up a deep reflection.

- H.1.1. The use of pre-trained Transformer-based embeddings to represent students' reflections will improve performance compared to binary unigrams, tf-idf, GloVe, and ELMo.
- H.1.2. Fine-tuning Transformer-based embeddings on text extracted from CRYSTAL ISLAND will improve predictive performance compared to off-the-shelf pre-trained embeddings.
- H.2. Data augmentation and reflection synthesis techniques will improve performance of reflection depth assessment models compared to models trained solely on genuine student reflections.
 - H.2.1. The use of transformation-based data augmentation techniques (i.e., mask-filling using BERT, noise injection, and back-translation) to provide more data for model training will improve predictive performance.
 - H.2.2. The use of synthetic reflections generated by large language models based on sample gameplay interaction sequences and desired reflection characteristics will significantly improve predictive performance over transformation-based augmentation techniques.
- H.3. Zero-shot Transformer-based approaches for automatically identifying which gameplay interactions students referred to in their reflections will achieve equivalent performance compared to supervised Transformer-based approaches.
- H.4. Predictions of students' learning outcomes based on features that capture the content of students' reflections will outperform predictive models based on gameplay trace data or non-content-based reflection features.

1.2 Contributions

The work described in this dissertation makes the following contributions:

- C.1. Insights into students' reflection behaviors during game-based learning and the relationship between their reflection behaviors and learning outcomes (Carpenter et al., 2021).
- C.2. A novel learning analytics-driven approach for automatically deriving interpretable insights based on the content of students' reflections during game-based learning.

C.3. A novel approach for automatically analyzing students' reflections in the context of their problem-solving actions during game-based learning.

C.4. A novel approach for automatically assessing several dimensions of the depth of students' reflections during game-based learning through the use of Transformer-based language modeling and text data augmentation techniques.

1.3 Organization

The remainder of this dissertation is organized as follows. Chapter 2 discusses background on self-regulated learning, learning analytics, and natural language processing. Chapter 3 presents an overview of the proposed automated reflection support framework, which leverages machine learning models of reflection depth and reflection content to analyze students' reflections. Chapter 4 describes the CRYSTAL ISLAND game-based learning environment, which is the testbed for studying student reflection. Chapter 5 presents a set of exploratory analyses that examine the depth and content of students' reflections and investigate the relationship between students' reflection behaviors and their learning outcomes. Chapter 6 presents work that analyzes students' reflections in the context of their recent gameplay interactions to construct predictive models that can inform adaptive interventions related to the types of reflections that students engage in at certain times during learning. Chapter 7 introduces a set of techniques for automatically assessing the depth of students' reflections during game-based learning using data augmentation techniques and Transformer-based language models. Finally, Chapter 8 summarizes the contributions of the dissertation work and presents opportunities for future research.

CHAPTER 2

BACKGROUND AND RELATED WORK

The research presented in this dissertation brings together work in the fields of self-regulated learning, learning analytics, and natural language processing. These fields have become increasingly intertwined over the years, indicated by the emergence of learning analytics designed to support SRL (Winne, 2017), a substantial body of work applying NLP techniques to educational tasks (Condor et al., 2021; Laban et al., 2022; Park et al., 2022), and even applications of NLP for evaluating students' reflections (Gibson et al., 2017; Jung & Wise, 2020; Magooda et al., 2022; Ullmann, 2015). Through the integration of these three research areas, the proposed work seeks to (1) derive insights about students' reflection processes during game-based learning to better understand how students engage in reflection and (2) build a machine learning framework for automatically analyzing the depth and content of students' reflections during game-based learning. This section describes background and related work in these three areas. Chapter 2.1 discusses work on self-regulated learning with a focus on reflection. Chapter 2.2 addresses work related to learning analytics, especially applications of learning analytics for understanding and supporting self-regulated learning. Finally, Chapter 2.3 discusses recent advances in the field of natural language processing, including applications of natural language processing techniques to education in general and assessment of students' reflections in particular.

2.1 Self-Regulated Learning

Self-regulated learning is a popular framework for understanding the cognitive, metacognitive, motivational, and affective aspects of learning (Panadero, 2017). SRL skills addressing each of these areas, such as the usage of learning strategies (cognitive), evaluation of one's own knowledge (metacognitive), development of goals that are relevant to one's own values and interests (motivational), and management of confusion and frustration during learning (affective), are important for students to learn and solve problems on their own. These skills are especially important when dealing with complex topics in STEM areas (Azevedo & Gašević, 2019). Moreover, with the growing adoption of advanced learning technologies, such as intelligent tutoring systems, hypermedia, and game-based learning environments, SRL skills are

essential for students to get the most out of their learning (Azevedo & Gašević, 2019). Unfortunately, many students, from middle schoolers to undergraduates, lack strong SRL skills (Azevedo & Gašević, 2019; Bernacki et al., 2021).

There are several competing models of self-regulated learning that have sought to establish theories about the internal processes that govern students' learning and validate them with empirical evidence (Panadero, 2017). These include Zimmerman's (2000) Cyclical Phases model, wherein students engage in forethought, performance, and self-reflection; Winne and Hadwin's (2008) model, with a strong focus on metacognitive monitoring and the goal-driven nature of SRL; and Pintrich's (2000) model, which highlights the role of motivation and goal orientation in SRL. While these models have some differences, these and nearly all other SRL models conceptualize SRL as being cyclical and featuring different phases (Panadero, 2017). A survey of several models identified three general phases that are common to most SRL models: preparatory, performance, and appraisal (Panadero, 2017; Puustinen & Pulkkinen, 2001). In the preparatory phase, students seek to understand the problem or task at hand and engage in goal setting and planning (Winne & Hadwin, 2008). In the performance phase, students enact learning strategies to complete the task and monitor their progress to ensure that they are moving toward the completion of their goals (Panadero, 2017). In the appraisal phase, students engage in self-reflection to evaluate what they have achieved so far, which informs adaptation for future performance (Winne & Hadwin, 2008; Zimmerman, 2000).

In this work, we adopt Winne and Hadwin's (2008) model as the theoretical framework for studying SRL. Winne and Hadwin's model describes four phases of SRL: problem understanding and task definition, goal setting and plan generation, enactment of learning tactics and strategies, and reflection and adaptation. Notably, the Winne and Hadwin model presents the phases of SRL as being only partially ordered and recursive, which is different from Pintrich's and Zimmerman's models. This can be explained by the observation that all of the different phases and processes involved in SRL are very closely related, so a student may move fluidly between phases without completing a full cycle.

This work focuses on deriving insights that can enable an AI-augmented system to provide feedback on the reflection and adaptation phase of SRL. In this work, we consider reflection to be defined as deliberate contemplation, and in order to reflect effectively, a learner must be

conscious of their own reflective thinking through introspection (Tarricone, 2011). The process of both reflective thinking and introspection leads to the development of self-knowledge (e.g., self-efficacy), which fosters the development of metacognition (e.g., monitoring one's understanding during learning activities) (Flavell, 1979; Tarricone, 2011). Reflection lays the foundation for developing higher-order thinking skills such as metacognitive monitoring and problem solving (Luo & Baaki, 2019; Patel et al., 2019).

2.2 Learning Analytics for Self-Regulated Learning

In recent years, data analytics techniques have fostered the emergence of the field of learning analytics. The Society for Learning Analytics Research defines learning analytics as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs” (Society for Learning Analytics Research, 2022). Being so broadly defined, the field of learning analytics can be conceived as covering nearly all data-driven analyses of educational data, and the lines between it and the neighboring fields of educational data mining and artificial intelligence in education have recently been dissolving, both in terms of the fields' goals and analytical techniques (Mavrikis et al., 2021; Romero & Ventura, 2020). Learning analytics deals with diverse forms of data, including interaction data, assessment responses, and forum posts, which are collected from many different sources, such as traditional classroom settings, intelligent tutoring systems, and MOOCs (Romero & Ventura, 2020). A wide range of techniques have been used to analyze and interpret this data, including data visualization, predictive modeling, clustering, knowledge tracing, social network analysis, process mining, and text mining (Romero & Ventura, 2020). These techniques have been used to analyze educational data for the purpose of revealing information about learning phenomena (Jivet et al., 2021; Vrzakova et al., 2020), as well as to derive data-driven insights for students and teachers, often with the goal of promoting self-regulation through the provision of feedback (Jensen et al., 2021; D. Wang & Han, 2021).

In recent years, there has been an increasing amount of work specifically devoted to the application of learning analytics techniques to students' self-regulated learning processes (Viberg et al., 2020; Winne, 2017). Systems that employ learning analytics for SRL typically include two elements: (1) an analysis of student trace data that captures learning or studying

activities, and (2) a recommendation to the learner or other stakeholder (e.g., a teacher) of changes to be made and instructions for implementing those changes (Winne, 2017). For example, prior work has investigated the use of dashboards for helping students set and monitor appropriate learning goals (Law et al., 2016; Molenaar et al., 2020). These visualizations were demonstrated to promote awareness of students' goals and helped them monitor the progress they were making toward achieving those goals, thus leading to better use of SRL strategies and improved learning outcomes. Other work has used students' logs from interactions with digital learning environments to better understand the metacognitive processes that are involved in learning (Fan et al., 2021; Raković et al., 2022). Fan et al. (2021) detected micro-level SRL processes from students' actions during online learning activities and then clustered sequences of actions to identify distinct learning tactics. This provides a way to understand learning tactics in terms of SRL processes, which allows for learning tactics to be compared across different learning environments. Raković et al. (2022) looked at metacognitive processes in the context of multi-source writing tasks and found that they were predictive of scores assigned to students' writing, indicating an empirical link between students' metacognitive processes and their writing products. Other recent work has used analyses like these to deliver feedback directly to students, such as a study by Lim et al. (2021) that explored the impact of personalized feedback generated by a learning analytics system on undergraduate students' studying and time management strategies. Although quantitative analyses revealed that the feedback did not guide students toward more effective SRL strategies, qualitative analyses showed that students perceived the feedback to be impactful on their strategy usage. These studies demonstrate ways that learning analytics techniques have been applied to students' interactions with a digital learning environment to unveil SRL processes, which can be helpful for better understanding how these processes unfold and for providing students with feedback that can help them improve their use of SRL strategies.

However, while there has been a substantial amount of recent work applying learning analytics to SRL, a review on learning analytics for SRL in online learning environments found that many studies have focused on the preparatory and performance phases of SRL, with relatively little work focusing on reflection (Viberg et al., 2020). Moreover, most of the empirical research on learning analytics for SRL that was examined in the survey was conducted in higher education settings, thereby ignoring younger students for whom the development of effective

SRL skills is critically important, especially with an ever-growing self-directed online learning component across all levels of education (Carter et al., 2020). Also, research that has looked at students' reflective processes has often focused on post-learning reflection rather than course-correcting reflection that occurs during a learning activity. Thus, there is a need for additional research on the reflective processes of K-12 students as well as methods for leveraging learning analytics techniques to understand and support these processes during an activity such as game-based learning. This research presented in this dissertation addresses these needs.

2.3 Natural Language Processing

In recent years, the field of natural language processing (NLP) has made significant advances. In particular, the use of deep neural networks to model natural language has pushed the field forward considerably. Driven by the availability of large language datasets (Lecun et al., 2015) and increases in computational power enabled by modern graphics processing units (Coates et al., 2013), it has become possible to train deep neural networks with upwards of a billion parameters (Goodfellow et al., 2016), which achieve high performance on many NLP tasks. Newer and newer models leveraging deep learning have continued to achieve improved performance on various tasks such as information retrieval, information extraction, text classification, text summarization, text generation, question answering, and language translation (Otter et al., 2021). One of the central features of modern NLP is the word embedding language representation, which is a vectorized representation of natural language that captures both syntactic and semantic features of language, including the relationships between words (Mikolov et al., 2013). Word embeddings are often created by extracting the internal state of a neural language model, and they have become the standard form of input for NLP systems (Otter et al., 2021).

Many recent advances in the field of NLP can be attributed to the Transformer neural architecture (Vaswani et al., 2017). While previous state of the art models relied on recurrent neural network architectures, such as the long short-term memory network (LSTM) (Hochreiter & Schmidhuber, 1997), Transformers instead make use of an encoder-decoder architecture in which a self-attention mechanism is a major component (Vaswani et al., 2017). In an encoder-decoder network, variable-length inputs (e.g., text) are transformed into a fixed-length vector by a neural network module, the encoder. Then, another neural network module, the decoder, takes

this condensed vector representation and reconstructs some output, such as the input text translated into another language. In previous encoder-decoder architectures that relied on recurrent neural networks, a common problem was that the network had to encode an entire text sequence as a fixed-length vector without considering whether any parts of the inputs were more important than others (Otter et al., 2021). With the introduction of self-attention, the words in a piece of text are weighted according to an estimate of their relative importance. When modeling a piece of text, a self-attention vector estimates the pairwise relationship between every two words in the sequence, implicitly capturing a measure of the connections between words (Qiu et al., 2020). As a result of this self-attention mechanism, Transformer-based models create representations of natural language that are sensitive to the context in which the words appear, as opposed to static word embeddings like word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2015).

An important factor that has helped enable the recent advances in NLP is the widespread open access to models that are pre-trained on massive text corpora, such as the BERT family of language models (Devlin et al., 2018). Now, even those without access to the sometimes billion-word corpora and the massive amounts of compute power and time that are required to train these models can use pre-trained models that have been made available via open-source communities like Hugging Face (Wolf et al., 2020). Researchers have been able to take the language representations learned by these large, general-purpose language models and transfer them to domain-specific tasks with impressive results (Chronopoulou et al., 2019; Gururangan et al., 2020). Often a fine-tuning step is used to further improve performance, where the large language model is adapted to a new domain through additional training on a small corpus of representative text (Dodge et al., 2020; Min et al., 2021). The ability to transfer the general knowledge of language captured by these large models without the need to always train new models from scratch has helped facilitate significant progress on downstream NLP tasks in many fields, including education (Cochran et al., 2022; W. Liu et al., 2022).

2.3.1 Natural Language Processing in Education

NLP techniques have been widely applied to educational tasks, including automated essay and short answer grading (Mayfield & Black, 2020; Z. Wang et al., 2018), the creation of intelligent pedagogical agents (Grivokostopoulou et al., 2018), automated question generation (Laban et al., 2022; Stasaski et al., 2021), and management of group conversation during collaborative learning

(Carpenter, Emerson, et al., 2020; Nikiforos et al., 2020; Park et al., 2022). Three common steps are often seen when applying NLP to these various educational tasks: (1) collecting and manually coding text data according to some rubric designed to score or classify some learning phenomenon; (2) performing text preprocessing and extracting various natural language features that may be predictive of the learning phenomenon; and (3) training and evaluating supervised machine learning models from a range of classical and deep learning-based algorithms to predict the learning phenomenon, using the previously identified features as inputs to the models.

Before machine learning models can be applied to these tasks, text data that has been collected from students' interactions with a learning environment must be coded. This data coding step is notably absent in much of the foundational research on NLP, since there are benchmark datasets available for many different tasks, such as SNLI for natural language inference (Bowman et al., 2015), SQuAD for question answering (Rajpurkar et al., 2016), and GLUE for several natural language understanding tasks (A. Wang et al., 2018). There do exist some publicly available education datasets that feature text data, such as the ASAP dataset for automated essay scoring (Shermis, 2014), Duolingo's spaced repetition dataset for second language acquisition (Settles & Meeder, 2016), and the TalkMoves dataset for teacher and student discursive moves (Suresh et al., 2022), but they are far less common. Moreover, it is often the case that the learning phenomena and the learning contexts that are being investigated are very specific, so new data must be collected and coded to answer the research questions being explored. For example, students' chat interactions with a pedagogical agent may be very different from their chat interactions with other students during collaborative learning, so separate datasets are likely needed to explore students' chat behaviors in these different settings. As a result of this specificity, and because it can be expensive and time consuming to collect student data, much of the work applying NLP to education utilizes very small amounts of data. For example, an educational dataset may contain data collected from hundreds of students with short pieces of text from each student (Gao et al., 2022; Jia et al., 2022), which is very small compared to datasets used in foundational NLP work which sometimes contain millions of documents. Once data is collected, the text is typically coded based on a rubric that assigns a score or qualitative label to each unit of text, which is often an individual sentence or a short text submission (Litman et al., 2021; Sha et al., 2021). For example, a rubric may be used to assign scores from 1 to 5 for different components of students' essays, such as grammatical correctness

(Crossley et al., 2019), successfully supporting an argument with relevant evidence (Rahimi et al., 2014), or having good organization and a valid flow of ideas (Boulanger & Kumar, 2020). As an example of qualitative coding, a rubric may categorize students' communications during collaborative learning based on the perceived intent of each communication, such as facilitating teamwork, regulating emotions, sharing information, or helping to establish a shared understanding of the content being learned (Park et al., 2022; Pugh et al., 2021).

Given a labeled dataset, text is optionally preprocessed and passed through a pipeline to extract natural language features. Preprocessing may include procedures such as lemmatization, removal of punctuation and stop words, or attempts to correct misspellings (Keerthi Kumar & Harish, 2018). While some of these preprocessing steps may not be necessary when using features extracted from deep learning-based language models (Qiao et al., 2019), these models may be sensitive to misspellings (Moradi & Samwald, 2021; Sun et al., 2020), so attempts at misspelling correction may still be beneficial (Hu et al., 2020). Regardless, the text is then sent through a pipeline that may extract a wide range of different language features. These could include features such as the number of words in the piece of text, the number of words per sentence, the usage of pronouns, part-of-speech tags for each word, or the most common n-grams used (Hossen et al., 2022; Kastrati et al., 2021). Additional hand-crafted features may also be used to capture information that is thought to be related to students' emotional, cognitive, or social states (Pennebaker et al., 2014) or information that is specific to the particular educational application or domain being investigated, such as the length difference between a students' answer to a question and a known correct answer (Y. Zhang et al., 2016) or the extraction of keywords associated with a specific domain (Hasanah et al., 2019). Finally, many modern NLP applications use pre-trained word embeddings to leverage general knowledge from massive text corpora for their specific educational applications. Common word embedding models used include word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2015), fastText (Bojanowski et al., 2017), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2018), which introduced the Transformer architecture and has been widely used in recent years.

The set of natural language features that is extracted from the text is then used to train and evaluate supervised machine learning models. Prior work has seen the extensive use of classical machine learning algorithms in educational applications, including logistic regression, naïve Bayes, random forests, and support vector machines (Ahmed et al., 2021; Hastings et al.,

2018; L. Zhang et al., 2022). More recently, neural networks have been used to predict learning phenomena based on natural language features, with a particular focus on recurrent neural networks and Transformers because of their ability to capture the sequential nature of language (Riordan et al., 2020; Xue et al., 2020; H. Zhang & Litman, 2021). For example, LSTM-based deep learning models were used by Park et al. (2022) to predict the presence of disruptive talk among students in a collaborative game-based learning environment. Many of these computational techniques have been used to analyze text data related to students' SRL processes, often with a focus on reflection.

2.3.2 Automated Reflection Analysis

There is a growing body of work that specifically applies NLP techniques to the analysis of students' reflections, and work in this area also follows the three-step process outlined above. After collecting students' reflections, most prior work has involved the creation of rubrics that distinguish between varying degrees of reflection depth and different characteristics of reflection breadth (Ullmann, 2015). The natural language features used to represent students' reflections have varied substantially, ranging from rhetorical moves that may be indicative of reflection (Gibson et al., 2017) to various vectorized representations of reflection text (Kovanović et al., 2018; Ullmann, 2015). Finally, several different machine learning models have been evaluated, including naïve Bayes, random forests, and XLM-RoBERTa (Nehyba & Štefánek, 2022). Some recent studies on the automated analysis of students' reflections are examined in detail below.

Liu et al. (2019) investigated several different machine learning models (random forest, SVM, naïve Bayes, and a rule-based model called PART) for classifying pharmacy Masters students' reflections as either reflective or non-reflective across seven different stages of reflection: attending to feelings; relating new knowledge to previous knowledge; integrating prior knowledge with new knowledge; addressing feelings or attitudes; self-assessing beliefs, approaches and assumptions; internalizing the knowledge or experience; and relating new knowledge to personal experiences. Features generated by the Linguistic Inquiry and Word Count software (Pennebaker et al., 2014) and features based on Gibson et al.'s (2017) rule-based parser for reflective writing were used to train all models. The random forest model outperformed all others in terms of F1 score, and qualitative analysis suggested that non-reflective students tended to use general language rather than including specific details.

Nehyba and Štefánik (2022) investigated the performance of several classical machine learning models (i.e., SVM, logistic regression, naïve Bayes, and random forest) and a deep learning model based on XLM-RoBERTa, a cross-lingual Transformer-based language model, for classifying sentences from reflective journal entries into one of nine categories related to elements of reflection. These categories were reflection, description of an experience, feelings, personal belief, awareness of difficulties, perspective, lessons learned, future intentions, and other. The classical machine learning models were trained on a unigram bag-of-words language representation and the deep learning model used SentencePiece subwords (Kudo & Richardson, 2018) as input. Results showed that the Transformer-based model outperformed all of the classical machine learning models across all reflection categories. This demonstrates the promise of using modern Transformer-based language models to analyze students' reflections.

Magooda et al. (2022) built and deployed a system to automatically assess undergraduate students' reflections and provide real-time formative feedback. BERT-based neural networks (DistilBERT, RoBERTa, and DistilRoBERTa) were used to extract textual features from students' written reflections, which were then used as input to SVM classifiers to predict reflection quality on a scale from 1 to 4. Experiments used a set of student reflections collected from four different undergraduate courses and models were evaluated using leave-one-out cross-validation, thereby demonstrating how the models might perform on reflections collected from a new course. The model using DistilBERT features was found to perform the best in terms of quadratic weighted kappa, and it also took the least amount of time to embed the text, which makes it an appealing choice for real-time deployments. This work again highlights opportunities for using Transformers to derive features for analyzing students' reflections and begins to explore the practical implications of using these techniques to deliver real-time feedback to learners.

Altogether, while some work has been done to apply NLP techniques to the task of automated reflection assessment, only very recent work has begun to leverage the power of large pre-trained language models for this task (Magooda et al., 2022; Nehyba & Štefánik, 2022). This recent work demonstrates the potential of Transformer-based language models for automatically analyzing reflections. However, reflections have most often been collected from students in higher education, and often in the form of reflective essays (Gibson et al., 2017; Jung & Wise, 2020; Kovanović et al., 2018; Nehyba & Štefánik, 2022; Ullmann, 2015). There has been limited work investigating the reflections of K-12 students (Magooda et al., 2022), which may present

challenges for deep learning-based assessment models due to grammatical errors and misspellings (Riordan et al., 2019). Moreover, these reflections are often collected as a post-learning exercise as opposed to mid-learning, and they typically do not have access to rich trace data that captures the learning activities that students were reflecting on. Thus, there is a need to investigate robust techniques such as Transformer-based models for automatically analyzing K-12 students' written reflections that occur in the midst of their learning activities, which this dissertation addresses.

CHAPTER 3

REFLECTION-ENHANCED GAME-BASED LEARNING

This chapter presents a high-level description of a game-based learning environment that supports students' reflection skills as an auxiliary goal in addition to content knowledge acquisition and problem-solving goals. Figure 3.1 illustrates how an automated reflection analysis module can be integrated into a game-based learning environment to provide adaptive scaffolding and feedback for reflection.

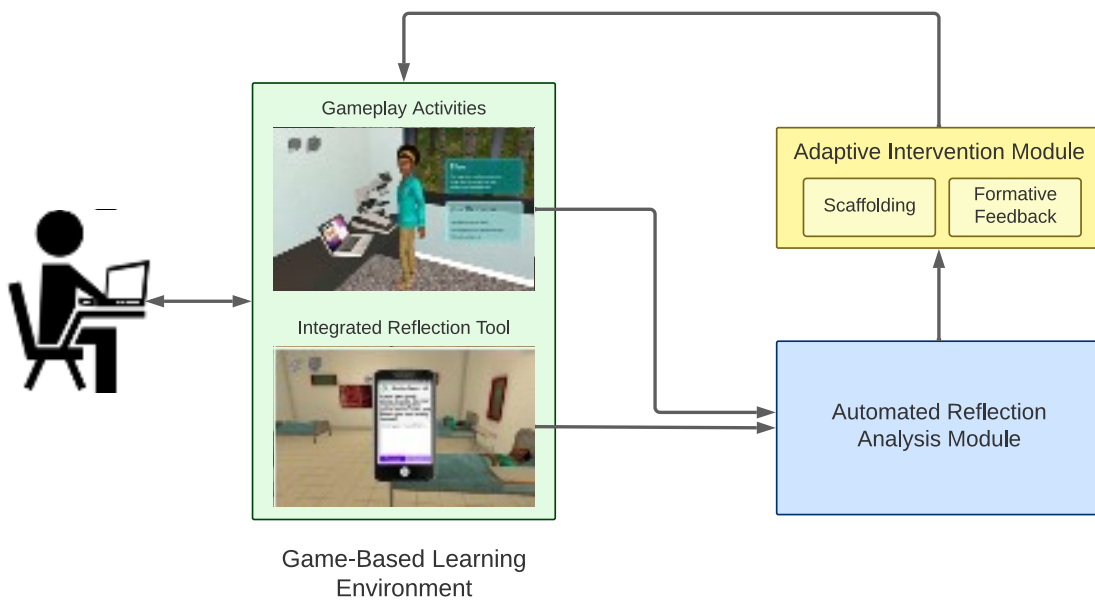


Figure 3.1 Overview of a reflection-enhanced game-based learning environment.

As students interact with the game-based learning environment, their gameplay interactions are logged by the system. This captures everything that a student does in the game, such as speaking with in-game characters, viewing educational content, or moving between locations in the virtual environment. Additionally, at several points during their gameplay, students are prompted to interact with an integrated reflection tool. This tool asks students to write a reflection that captures what they have done so far in the game as well as what they are going to do in the future. All of the data that is collected from students' gameplay interactions and in-game reflections is passed to the automated reflection analysis module, where it is used

to induce models that assess the depth and evaluate the topic of each reflection. Based on the outputs of these models, the system can deliver adaptive interventions aimed at improving students' reflection skills. For example, students can be presented with the depth score for their reflection and given suggestions on how to reflect more deeply. Alternatively, the learning environment can activate a scaffolding tool to provide support for the process of determining what information a student should include in their reflection.

The core of this system is the automated reflection analysis module (Figure 3.2), which encapsulates a machine learning framework that uses the data captured by the game-based learning environment to produce predictions that inform adaptive interventions aimed at supporting reflection.

When a student submits a reflection using the integrated reflection tool, the reflection text and the sequence of gameplay interactions logged by the system since the previous reflection are sent to the automated reflection analysis module. The gameplay sequence is then transformed from a sequence of trace data logs into natural language that can be understood by the NLP components used later in the module. This is done by taking trace data logs that capture interactions with educational resources and replacing them with relevant text (e.g., replace an interaction with an in-game book with the text of the book) or by using templates to transform gameplay interactions into natural language (e.g., moving to a different location in the game: “I

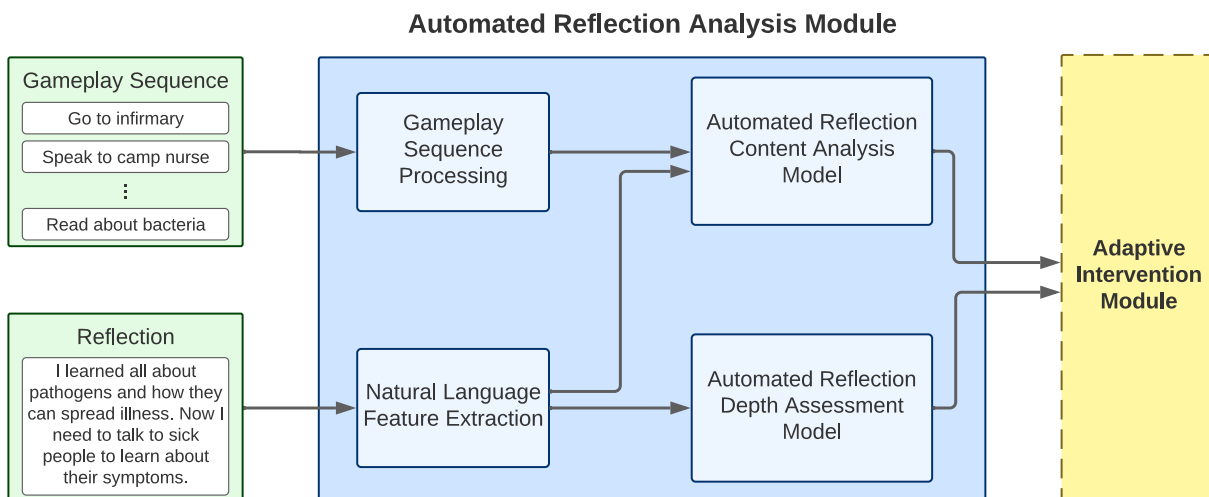


Figure 3.2 Overview of the proposed automated reflection analysis framework for a game-based learning environment.

moved to *Location*”). These text representations of students’ gameplay sequences are then converted into word embeddings using pre-trained language models. Simultaneously, the student’s reflection is passed through a pipeline that carries out preprocessing steps on the raw text and extracts linguistic features. These features may include simple rule-based features that are specific to the analysis of reflection as well as Transformer-based word embeddings. Extracted reflection features are then passed to machine learning models that assess the depth and extract information about the content of the reflection relative to the current gameplay context. The automated reflection depth assessment models take only the reflection features as input, but the automated reflection content analysis models also take in the processed gameplay sequence. Finally, the results of reflection depth assessment and reflection content analysis can be used to drive adaptive interventions to support reflection. In this dissertation, we evaluate the automated reflection analysis framework that enables this adaptive intervention, but the design, implementation, and evaluation of an adaptive intervention is beyond the scope of the proposed research.

CHAPTER 4

CRYSTAL ISLAND: REFLECT DATASET

During learning with CRYSTAL ISLAND, students are asked to investigate the mysterious outbreak of a disease on a remote island. To solve the mystery, students must identify the source of the disease (one of several food items that sick individuals have recently eaten), determine the pathogen that is spreading (either a pathogenic virus or bacteria), and recommend a prevention or treatment for the disease (either vaccination or bed rest). The specific scenario is randomly generated when the game is started, selecting the infected food item and either a virus or bacteria as the pathogen. To gather the information needed to solve the mystery, students interact with non-player characters (e.g., Elise the scientist; Figure 4.1), read books and research articles covering microbiology topics, view science posters and diagrams, and test objects for pathogens using a virtual scanner. Non-player characters (NPCs) include sick patients who describe their symptoms and recent activities prior to falling ill, bacteria and virus experts, a camp cook who provides information related to what residents have been eating lately, and a camp nurse who



Figure 4.1 CRYSTAL ISLAND game-based learning environment.



Figure 4.2 Embedded reflection prompt.

provides guidance and support to the student during their investigation. Findings related to the mystery are compiled by students in a virtual diagnosis worksheet, which is submitted to the camp nurse as a final diagnosis.

To collect evidence of students' reflection behaviors while they interact with CRYSTAL ISLAND, the game-based learning environment was augmented with embedded reflection prompts. At critical points during their investigation, students are prompted to reflect on their knowledge and problem-solving actions (Table 4.1). The prompts were designed according to Winne and Hadwin's (2008) model of SRL, so students are asked to contemplate the important information that they previously learned during their investigation as well as to set goals for solving the outbreak scenario and improving their microbiology content knowledge. Specifically, the prompts ask students, "Please describe the most important things that you've learned so far, and what is your plan moving forward?"

These prompts are triggered by a set of production rules associated with important actions taken by the student as they interact with CRYSTAL ISLAND (Table 4.1). The triggers were designed

Table 4.1 Triggers used to prompt reflection.

Trigger	Prompt
Briefed by the camp nurse	Agent, it looks like you've spoken with the camp nurse. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?
Viewed 6 microbiology texts	Agent, it looks like you've found several materials that might be useful. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?
Testing contaminated object	Agent, it looks like you found an object that tested positive for pathogenic contaminants. Before you continue, we'd like a report on your progress. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?
After submitting diagnosis worksheet with the wrong solution	Agent, it looks like you're making progress on diagnosing the illness, but you're not quite there yet. In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?
End of game (solved mystery or not)	<ul style="list-style-type: none"> • Please explain how you approached solving the mystery. • If you were asked to solve a similar problem in the future, what would you do the same and/or differently?

to align with milestones in the game's narrative (e.g., speaking with the camp nurse), students' acquisition of domain knowledge (e.g., viewing 6 microbiology texts that covered information specific to the illness plaguing the camp), and problem-solving actions (e.g., obtaining a positive test result). Based on which action triggered the reflection prompt, students received a message telling them why this would be a good time to reflect, followed by the prompt. The reflection prompts were queued immediately after each trigger, but the delivery of the prompts was scheduled to minimize disruption to students' gameplay experiences. Specifically, students were only prompted to reflect when they left a building in the game environment. Because reading materials, science diagrams, testing equipment, and virtual characters are all located in virtual buildings in CRYSTAL ISLAND, prompting for reflection only when students were outside allowed us to avoid disrupting students when they were directly in the midst of major problem-solving activities. Additionally, successive reflection prompts occurred at least 15 minutes apart since, with the triggers being event-based, it is possible that a student may activate several reflection

triggers in rapid succession. However, we wanted students to be able to maintain focus while gathering information and working on their investigation, and students need to make additional progress on their investigation before they have anything new to reflect on.

4.1 Participants

This work uses data from a classroom study conducted in 2018 and 2019. A total of 153 middle school students participated in the study, but only 118 students reported demographic information. Of these students, 51% identified as female and ranged from 13-14 years of age ($M=13.6$, $SD=0.51$), with 43 students identifying as Caucasian/White, 32 as African American, 21 as Hispanic or Latino, and 3 as Asian. None of the students reported that they had previously interacted with CRYSTAL ISLAND. Of the 153 students who participated in this study, only 105 had complete pre and post-test data. For the analyses conducted as part of the preliminary work presented in this proposal, different subsets of this dataset were used based on what data was required to answer a particular research question.

4.2 Study Procedure

Students interacted with CRYSTAL ISLAND over the course of two or three class periods, on average spending 81.4 minutes in the learning environment. In the week preceding the classroom study, we administered pre-survey instruments and a 17-item multiple-choice microbiology pre-test that covered both factual (e.g., “What is the smallest type of living organism?”) and procedural items (e.g., “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?”). On the first day of the classroom study, students were introduced to the game by a researcher and then shown a brief video detailing the problem scenario they would face in CRYSTAL ISLAND. Afterward, students interacted with CRYSTAL ISLAND until they completed the mystery or ran out of class time, which roughly allowed for 100 minutes of gameplay over two or three days. On the final day of the classroom study, students took a post-study survey including a 17-item microbiology post-test that was similar but different from the pre-test. Both tests assessed the same knowledge, but questions were presented differently. For example, the pre-test asked, “How do vaccines protect you?”, while the post-test asked, “What role do vaccines play in your immune system?” The average pre-test score was 6.90 ($SD=2.70$) and the average post-test score was 7.30 ($SD=3.35$).

CHAPTER 5

EXPLORING STUDENTS' REFLECTION BEHAVIORS DURING GAME-BASED LEARNING

This chapter describes the work done to analyze students' reflection behaviors during game-based learning with CRYSTAL ISLAND. First, we discuss the development of the reflection depth rubric that is used to evaluate the quality of students' reflections (Carpenter et al., 2021; Carpenter et al., 2020). Since reflection is highly context-dependent, it is necessary to operationalize a definition of reflection that is specific to the learning scenario being explored. As such, we developed a rubric that assesses reflection depth as it relates to science problem solving. Next, we use automated topic modeling techniques based on Transformer language models to identify common topics that students reflected on as they interacted with CRYSTAL ISLAND. Furthermore, we align the topic of students' reflections with scientific inquiry processes to enable a more general interpretation of the topic of students' reflections. This provides a way to explore not only *how well* students reflected (i.e., depth), but also what they thought was important to include in their reflections. Through a joint lens of reflection depth and reflection content, we examine relationships to students' learning outcomes and relationships between reflection depth and content. Finally, we examine changes in students' reflections over the course of their interactions with CRYSTAL ISLAND to investigate how students demonstrate different reflection behaviors over time.

Findings from these analyses present empirical evidence that aligns with existing theories linking effective reflection to positive learning outcomes. Findings also help establish the value of engaging students in reflection during science problem solving and highlight the need to help train students on how to reflect.

5.1 Analyzing the Depth of Students' Reflections

To enable adaptive systems to provide support for reflection during game-based learning, a framework for assessing reflection is necessary. In previous work, written reflections have often been assessed along two dimensions: reflection depth, which captures the extent to which the writing is reflective; and reflection breadth, which addresses the range of different topics related

to reflection (Jung & Wise, 2020; Ullmann, 2015). Depth is often evaluated on an ordinal scale, such as from non-reflective to slightly reflective to highly reflective (Kovanović et al., 2018; Van Manen, 1977), while breadth may consider aspects such as ‘attending to feelings’, ‘validation’ (Wong et al., 1995), ‘justification’ (Poldner et al., 2014), ‘analysis’, and ‘perspective’ (Jung & Wise, 2020). In this work, we evaluate reflections in terms of their depth because the reflections we collected from middle school students tend to be short ($M=18.6$ words, $SD=14.0$) and limited in reflective breadth.

To develop a rubric for assessing the depth of students’ reflections on a scale from 1 (not reflective) to 5 (highly reflective), we used a grounded theory approach (Saldaña, 2021). Looking at reflections that were collected as students played CRYSTAL ISLAND, two researchers first worked to identify reflections that were particularly shallow and discussed what stood out about them. They found that these reflections lacked any commentary on knowledge or a plan of action, were too abstract to be meaningful, or were largely unactionable. These insights formed the basis for a reflection depth score of 1 (see Table 5.1 for examples). Next, to inform a reflection depth score of 5, the researchers identified some reflections that were particularly deep and discussed their strengths. These reflections presented both a clear hypothesis regarding the current problem and either outlined a concrete plan that was supported by reasoning or laid out a high-quality sequence of abstract activities. Qualities for the other reflection depth scores (i.e., 2, 3, and 4) were similarly determined.

By assessing the extent to which students evaluated their own knowledge and articulated plans exemplifying high-quality reasoning, hypothesis formation, and metacognition, the rubric aligns with the COPES model of self-regulated learning (Winne & Hadwin, 2008). That is, students are expected to use reflection in a backward-looking manner to evaluate their progress toward solving the mystery of the spreading illness, which includes generating and evaluating hypotheses (i.e., Evaluate the Products of their learning with respect to the Standard for success), as well as in a forward-looking manner to consider how they should proceed based on that evaluation, which includes constructing plans for the future (i.e., informing changes to their Operations relative to the Conditions of the learning environment).

Table 5.1 Reflection depth rubric.

Rating	Characteristics	Examples
1	Lacks both a hypothesis and plan. The student does not demonstrate awareness of their own knowledge or goals.	“Each clue will help with solving the problem”; “Yeah cool game I learned science”
2	Presents a vague hypothesis or plan, often directly restating information that was presented in the game. The student demonstrates awareness of their own knowledge and goals but does not show that they are evaluating their knowledge to inform their future actions.	“That the illness causing the people being sick might be pathogen”; “I found out that the egg has bacteria”; “I think I am going to talk to other people”
3	Presents a clear hypothesis or plan <i>without</i> any reasoning. This demonstrates that the student has evaluated their knowledge and made connections to their goals. However, they have not articulated the reasoning behind the importance of this knowledge or its benefit toward achieving their goals.	“Getting more information off the food I think it has something to do with the food”; “The most important thing is how the illness is spreading”
4	Presents a clear hypothesis <i>or</i> plan with reasoning. This demonstrates that the student has evaluated their knowledge and made connections to their goals. However, they have only provided reasoning for the importance of this knowledge <i>or</i> its benefit toward achieving their goals, not both.	“I plan on questioning the cook as they know more about the food and how it could be contaminated with viruses or bacteria”; “I need to learn more about what the sick people do on a day to day schedule”
5	Presents both a clear hypothesis <i>and</i> plan with reasoning. This demonstrates that the student has evaluated what they have learned and made connections to their goals. Furthermore, they have provided reasoning for the importance of this knowledge, <i>and</i> they have indicated how it will help them achieve their goals.	“I think that it might have to do with salmonella because when I tested the milk it was positive with pathogenic bacteria. I think that I will test things that can be contaminated”

Once the rubric was established, each reflection was annotated by both researchers and an intraclass correlation of 0.669 was achieved, indicating moderate inter-rater reliability. Across all reflections, the average depth score was 2.41 (SD=0.86). Final reflection depth scores were

obtained by averaging the scores from the two raters and then taking the floor of that rating to align the final score with one of the scores from the reflection depth rubric. Averaging ratings is a standard approach for reconciling differences between coders' assigned ratings, although it does have limitations. For example, reflections that received the same rating from both coders (e.g., 3 and 3) and reflections that received different ratings (e.g., 2 and 4) would be rated the same even though there is disagreement in the latter case. The floor of the average was taken to err on the side of providing too much feedback, since the low depth scores indicate a general need to help students reflect more deeply.

5.1.1 Examining the Predictive Relationship Between Gameplay Behaviors and Reflection Depth

Next, we investigated whether reflection depth could be predicted by features that captured students' problem-solving actions or simple features related to their reflection behaviors. Problem-solving actions capture students' progress through the game's narrative, their acquisition of science content knowledge, and their problem-solving processes. A feature related to narrative progress was the number of plot points completed by the student. Plot points included speaking with a character for the first time, learning about viruses and bacteria, finding out what symptoms patients had and what foods they had recently eaten, and testing objects for pathogens. For science content knowledge features, we looked at the number of microbiology books that the student read, since books are a primary source of microbiology information in the game. Conversations with NPCs were also considered a major source of science content knowledge since students can speak with a virus expert, a bacteria expert, and an expert on how diseases spread between people. Finally, features related to science problem-solving processes included the number of times items were tested for pathogens, the number of times the diagnosis worksheet was submitted with a final diagnosis, and whether a positive test result was obtained by the end of the learning experience. The number of tests and diagnosis worksheet submissions have been linked to problem-solving efficiency (Taub et al., 2018), and the positive test result is directly related to problem-solving performance.

Features based on students' reflections were derived to capture information about students' interactions with the reflection tool and the content of the reflections that were submitted. First, the number of words in each reflection was calculated and the amount of time between when a student was prompted to reflect and when they submitted the reflection was

extracted from CRYSTAL ISLAND’s trace logs. Next, the number of domain-specific words used in each reflection was calculated. This feature was derived using a dictionary of keywords that was extracted from all of the text content students can interact with in CRYSTAL ISLAND (i.e., books, articles, and conversations with non-player characters). Stop words were filtered out using NLTK (Bird et al., 2009), words were stemmed to reduce them to their root form, and the 100 most common terms were selected. Candidate keywords were then manually curated and finally cross-referenced with an educational standards document that was used to develop CRYSTAL

Table 5.2 Hierarchical regression results for predicting average reflection depth. All regression coefficients are from the final step in the analyses.

Predictor	Reflection Depth		
	β	ΔR^2	F
Step 1: Prior Content Knowledge		0.0453	13.3242***
Pre Test Score	0.001	0.0453***	
Step 2: Problem-Solving Actions		0.1690	9.9371***
Number of Plot Points Completed	0.001	0.0044	
Number of Books Read	-0.005	0.0000	
Number of Worksheet Submissions	-0.029	0.0951***	
Number of Lab Tests	0.000	0.0050	
Obtained a Positive Test Result	0.204	0.0644***	
Step 3: Reflection Features		0.4796	17.6252***
Number of Words	0.036	0.4331***	
Number of Keywords	0.074	0.0111†	
Time Spent Reflecting	-0.003	0.0069	
Focuspast	0.036	0.0051	
Focusfuture	0.027	0.0001	
Authentic	0.000	0.0001	
I	-0.008	0.0069	
Analytic	0.005	0.0163*	

*** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$

ISLAND's educational content, resulting in a final set of 36 keywords. The top 5 most common keywords were *bacteria*, *disease*, *virus*, *cell*, and *infection*. In addition to these reflection features, a subset of the features included in the LIWC text analysis tool (Pennebaker et al., 2014) was used to capture information related to the content of each reflection. Rather than using all the available LIWC features, we used the five most predictive features according to a recent study on automated reflection depth assessment (Jung & Wise, 2020). The features used were *Focuspast* and *Focusfuture* (the extent to which the text focuses on the past or the future), *Authentic* (how honest, humble, and vulnerable the text is), *I* (the extent to which the text includes first person pronouns), and *Analytic* (how much the text focuses on formal, logical, and hierarchical thinking patterns).

To examine the predictive relationship between these features and reflection depth, we utilized hierarchical regression analysis. Hierarchical regression analysis provides a framework for determining which variables explain statistically significant variance in a dependent variable after accounting for all predictors. We identified several problem-solving actions and reflection features that were predictive of average reflection depth (Table 5.2). For problem-solving actions, the number of times that students submitted the diagnosis worksheet and whether they received a positive test result were predictive of reflection depth. These actions, which can be viewed as measures of science problem solving efficiency and science problem solving ability, respectively, align with previous work suggesting that problem-solving processes are related to higher-order thinking skills like goal setting and adaptation, which are related to reflection (Taub et al., 2018). However, since these actions often occur at the end of students' interactions with CRYSTAL ISLAND, they do not hold much promise for adaptively identifying students who are not engaging in deep reflection.

Some features extracted from students' reflections were found to be predictive of reflection depth, as was the case in Jung & Wise (2020). First, the number of words per reflection was predictive of reflection depth, with longer reflections tending to have higher scores. This is not very surprising since it seems likely that a student's reflection would need to exceed some minimum length for them to adequately evaluate their knowledge and consider adaptations for the future. However, it may be the case that the "correct length" for a reflection could be a range rather than a lower bound. That is, if a student's reflection covers too much information, they may fail to derive actionable insights regarding the changes they should make to their learning

processes. It will be important to identify guidelines for determining when a reflection is too short or too long to support meaningful adaptation during game-based learning. Nevertheless, the number of words in a reflection may be useful as a baseline indicator to encourage students to engage in deeper reflection. For example, when students try to submit a reflection that is extremely short, the system can simply ask them to provide some more information. Specifically looking at the features borrowed from Jung & Wise (2020), the LIWC Analytic feature was predictive of reflection depth. This feature presents an opportunity to ask students to focus more on logic and reasoning when reflecting. Leveraging these insights, we can begin to enable future versions of CRYSTAL ISLAND that can adaptively support reflection by providing feedback based on how students respond to the embedded reflection prompts.

5.2 Analyzing the Content of Students' Reflections using Automated Reflection Topic Modeling

Beyond reflecting deeply, an important component of effective reflection is the actual content of the reflection, since a deep reflection on a topic that is not relevant to the student's goals would likely not be very beneficial for helping the student achieve those goals. To explore the content of students' reflections, we extracted topics that students commonly reflected on while interacting with CRYSTAL ISLAND. While such a content analysis is possible to perform manually, it can be time consuming and challenging to extract meaningful insights from a wide range of diverse reflections. A useful property of modern embedding-based representations of natural language is that similar words will have similar embeddings, so they can be used to identify clusters of text with similar characteristics. To extract common reflection topics automatically, we used SBERT embeddings to represent all reflections in the same high-dimensional vector space and applied a clustering method to identify reflections with similar semantic content. We qualitatively analyzed the different topics discovered by the clustering algorithm to see what students reflected on. Then, to support interpretation of the content of students' reflections in a broader context, we mapped each topic to scientific inquiry processes. This allowed us to investigate how students' reflections indicated the general strategies they used to solve the mystery.

5.2.1 Method

We utilized the BERTopic pipeline (Grootendorst, 2020) to automatically extract topics from students' reflections. This approach utilizes SBERT (Reimers & Gurevych, 2019), which is a variation of the BERT language model that was trained to create representations of sentences rather than individual words. SBERT models were trained on very large natural language datasets and the learned representations can be transferred to new domains, thereby leveraging the language comprehension capabilities in settings without large amount of data. First, we embed each reflection in a high-dimensional vector space using SBERT. Next, these high-dimensional embeddings are projected down into five dimensions using UMAP (Uniform Manifold Approximation and Projection; McInnes et al., 2020), since clustering algorithms struggle with very high-dimensional data and UMAP has been suggested as an effective approach for addressing this challenge (Allaoui et al., 2020). Then, the HDBSCAN algorithm (McInnes et al., 2017) is used to cluster the lower-dimension representation of students' reflections. HDBSCAN is preferred over popular clustering methods such as k-means because it is a density-based method that will not force data points into clusters if they may actually be outliers, and it does not require the number of clusters to be specified as a parameter. Next, BERTopic also identifies the most common words from each extracted topic as well as a few reflections from each cluster that are representative of the topic's semantic content. Finally, GPT-3.5 is used to generate a short summary of the reflections in each topic to allow them to be easily interpreted.

5.2.2 Results

A visualization of the nineteen identified topics can be seen in Figure 5.1. This visualization uses a two-dimensional UMAP projection while clustering was performed on five-dimensional data, so the clusters may not appear to be as clearly defined in two dimensions. The smaller, greyed out data points represent reflections that did not fit into any topic clusters.

The following eighteen topics were identified from students' reflections: (1) understanding microbiology and disease transmission, (2) understanding disease spreading and pathogens, (3) diagnosis process and plans, (4) understanding of mutagens and carcinogens, (5) learning through diligence and interactions with characters and clues, (6) pursuit of clarity, (7) identifying bacterial contamination in food, (8) identifying pathogenic virus in egg, (9) investigating illness in camp through communication with characters, (10) emphasis on food testing for disease

identification, (11) investigation of foodborne illness on the island, (12) identifying bread as a source of pathogenic virus, (13) difficulty understanding disease origin, (14) testing consumed food for bacteria, (15) unclear writing, (16) identifying disease symptoms and food transmission, (17) identifying pathogenic bacteria in milk, and (18) understanding disease transmission and prevention. Table 5.3 shows some basic information for each topic cluster, including the number of reflections in the cluster and a sample reflection that was identified as being representative of the cluster overall.

Looking at the reflection topics that were extracted by BERTopic, we can quickly gather insights into what students are focusing on when they are engaged with CRYSTAL ISLAND. Compared to manually analyzing hundreds of reflections, it is much more tractable to interpret ten sets of keywords and representative reflections to help identify general trends in students' reflections. Presented in a dashboard, this information could allow teachers to easily understand what students have been learning about and how they have been approaching the problem-solving scenario. For example, observing that many reflections mentioned that the student had learned about pathogens and how diseases spread (cluster 2) demonstrates that students were successfully engaging with the game's educational content.

Then, to interpret the content of students' reflections more generally, we mapped the identified reflection topics onto three phases of scientific inquiry, as defined in Klahr and Dunbar's theory of scientific discovery as dual search (Klahr & Dunbar, 1988). These phases are Hypothesis Generation (i.e., working with some knowledge about a domain to construct a hypothesis that fits with that knowledge), Experimental Design (i.e., constructing a scientifically valid experiment that will provide new information about the hypothesis), and Evidence Evaluation (i.e., comparing the predictions associated with the hypothesis with the results of the experiment). This allows us to interpret students' reflection behaviors from a perspective that is applicable beyond CRYSTAL ISLAND, which is useful for creating a reflection analysis framework that can potentially be applied to reflections in other learning environments. We interpreted reflections belonging to clusters 1, 2, 3, 4, 5, 6, 9, 13, 16, or 18, where students generally described their plans for learning more about the mystery, talked about theoretical knowledge they had gained, or reported on information they had learned about the specific situation on the island, as evidence that the student was in the Hypothesis Generation phase. A reflection from clusters 10, 11, or 14, where students discussed their focus on collecting foods and running tests to

evaluate whether the foods were contaminated, showed that the student was in the Experimental Design phase. Finally, a reflection belonging to clusters 7, 8, 12, or 17, where students reported on results that they had received from using the scanner, indicated that the student was in the Evidence Evaluation phase, since they were drawing conclusions from what they had learned in the game.

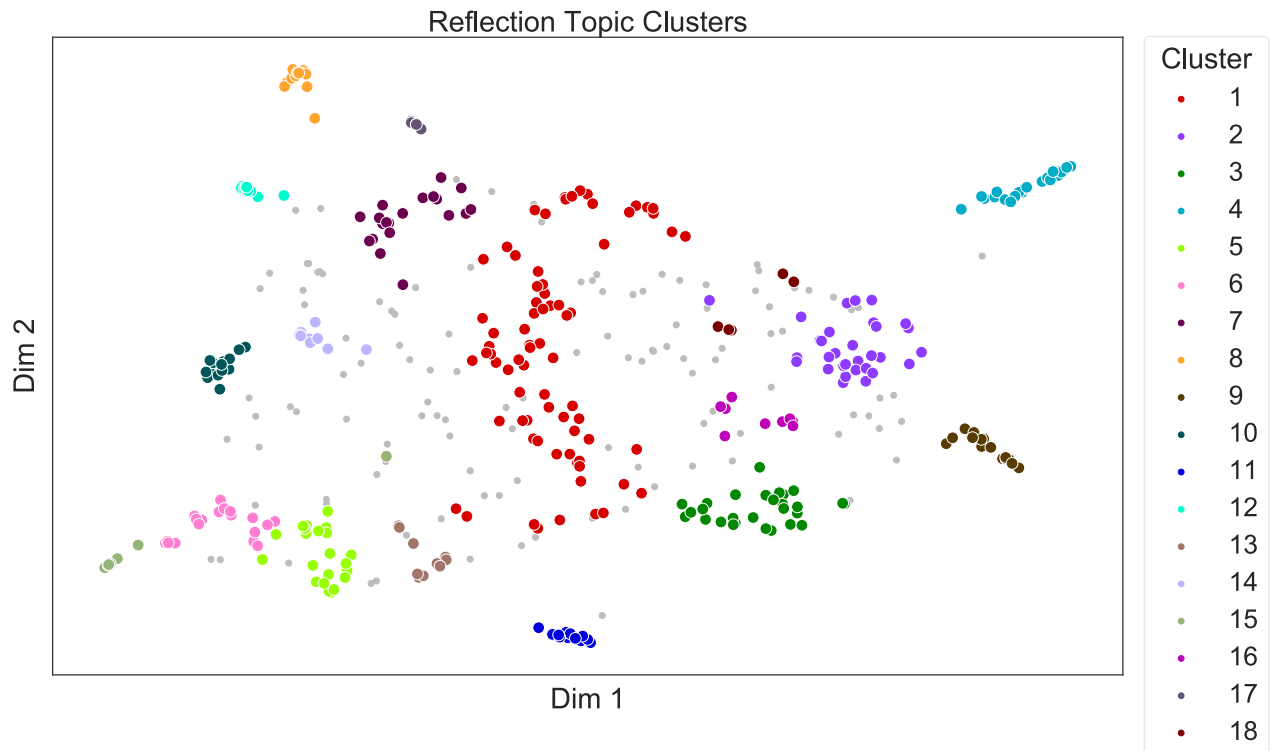


Figure 5.1 Visualization of the reflection topic clusters identified by BERTopic. Smaller greyed out data points represent reflections that were considered outliers.

Table 5.3 Descriptions of the reflection topic clusters identified by BERTopic.

Cluster #	Count	Description (generated by GPT-3.5)	Representative Reflection
Outliers	119	-	-
1	75	Understanding Microbiology and Disease Transmission	“I’ve learned about many types of microbes and diseases. I plan to keep on searching different papers to find about more diseases.”
2	31	Understanding Disease Spreading and Pathogens	“I have learned that the disease is a pathogen because the illness going around is being spread from one person to the next like a pathogen would. i plan on talking to the research team”
3	28	Diagnosis Process and Plans	“Well I’ve learned that I have to be more thorough through the rest of of the investigation because so far I’ve done well, but my diagnosis isn’t quite right, so I am check all my steps.”
4	26	Understanding of Mutagens and Carcinogens	“Carcinogens cause cancer and mutagens change DNA and can cause cancer. Neither of them can spread from person to person”
5	22	Learning through Diligence and Interactions with Characters and Clues	“the most important thing that ive learned so far is that i should look out for clues and focus on the information thats given.”
6	21	Pursuit of Clarity	“I think I have the answer but I don’t know for sure, so I’ll keep searching”
7	21	Identifying Bacterial Contamination in Food	“That E.coli might be the suspect, because the food is not properly stored,and the symptoms are similar to those of the infected.”
8	15	Identifying Pathogenic Virus in Egg	“The egg is positive it has bacteria”

Table 5.3 (continued).

9	15	Identifying Illness in Camp through Communication with Characters	“i plan to talk to people around the camp and find out”
10	14	Emphasis on Food Testing for Disease Identification	“I need to find more foods to test”
11	12	Investigation of Foodborne Illness on the Island	“I have learned that many people that have the disease have eaten the food on the island in the past few days. One item in which may be causing the sickness and the fevers would be the bananas.”
12	11	Identifying Bread as a Source of Pathogenic Virus	“I think the bread they used for toast caused Salmonella, which is a pathogenic bacteria, so it all matches up. I will look for other items to test.”
13	10	Difficulty Understanding Disease Origin	“i have leared a little bit of what is happening to those people”
14	9	Testing Consumed Food for Bacteria	“I learned about different types of bacteria and I plan to test the food they have been eating cause they got infected after they have been eating”
15	9	Unclear Writing	“that that nsndiibif[n vfhv”
16	9	Identifying Disease Symptoms and Food Transmission	“The illness seems to cause sore throat coughing and a fever”
17	8	Identifying Pathogenic Bacteria in Milk	“ive learned that not all bacteria and viruses are bad for you but in this case teres a pathogen in the milk which ma of caused the disease around camp”
18	5	Understanding Disease Transmission and Prevention	“parsites transmit diseases through many ways like from the ground, animals, and uncooked food. if we start wearing shoes, washing hands, and cooking meat before we eat it we can prevent parasite sickness.”

5.3 Identifying Links Between Reflections and In-Game Sources of Information

To investigate the relationship between the in-game source of information students interacted with and what they chose to reflect on, we needed to establish ground truth links between in-game sources of information and the content of students' reflections. However, students perform many actions in CRYSTAL ISLAND, so it is combinatorially infeasible to label each possible pair of actions and reflections based on whether the reflection is likely to have drawn information from that action. After all, there are several multi-page books in CRYSTAL ISLAND and a student's reflection might refer to any piece of information in one of those books. Thus, the full content of each book would have to be manually compared with each reflection to determine if there was a potential link. Since this is impractical, we leverage Transformer-based language models to identify a set of candidate actions that are most likely to be related to each reflection based on the semantic similarity between a text representation of the action and the text of the reflection. Then, human raters manually determine which, if any, of the candidate actions are related to the given reflection.

5.3.1 Method

For the first step in performing this reflection-interaction linking, we aligned each reflection with the corresponding set of recent gameplay interactions performed by the student in CRYSTAL ISLAND. In particular, we utilized timestamps to identify when each reflection took place in the gameplay interaction logs and extracted all interaction logs that occurred after the previous reflection and prior to the current reflection. For example, students received the first reflection prompt directly after completing the tutorial, so our approach would align the tutorial gameplay with the first reflection. Then, any new gameplay interactions occurring before the next reflection would be aligned with the second reflection. We note that a limitation of this approach is that it assumes that students will only include new information in a reflection, which may not always be the case.

Next, once the sequence of gameplay interactions was aligned with a particular reflection, we transformed each interaction into a natural language representation. In the case where an interaction log represented an interaction with one of the game's text-based educational resources (e.g., the text of a book resource or the dialogue spoken by an NPC), we replaced the

interaction log with the source text of that educational resource. However, not every important interaction in the game involves a text-based educational resource. For instance, an important problem-solving interaction in the game is scanning a food item for contaminants to determine if it is the vector through which the disease is being transmitted. In such a case where an interaction did not correspond to an educational resource, we used a template to transform the gameplay interaction. For example, interactions with the scanner were transformed to natural language using the template “Scanned *<food item>* for *<contaminant>* because *<reason>*.” A filled-out template for a scanner interaction might be “Scanned *apple* for *viruses* because *sick members ate it.*”

Next, we broke the text representation of each gameplay interaction down using a sliding window so we could embed smaller portions of a large text (e.g., books) rather than embedding the entire document. This allowed us to focus on the specific things that students may have referenced in their reflections, while also keeping enough context for human raters to validate whether this passage was a potential source of information that the student drew from when writing their reflection. In this work, we used a sliding window of three sentences. This was experimentally chosen because it was small enough to often be focused on a single concept but large enough to provide adequate context for the information. These overlapping subsequences from the text were then embedded using SBERT. Then, the reflection corresponding to these gameplay interactions was split into individual sentences. Since students may have reflected on different things across different sentences (e.g., “I’ve learned about different symptoms that someone is feeling. Also, I’ve learned about Anthrax and its symptoms.”), we wanted to be able to capture multiple different concepts being addressed within the same reflection. After being split up, these sentences were embedded using SBERT. For completeness, the whole reflection was also embedded to account for instances where multiple sentences in the reflection did actually address the same concept. Thus, all gameplay interactions and all reflections were embedded in the same vector space, so they could be compared pairwise to identify the most similar gameplay interactions for each sentence in the reflection. Since each interaction was potentially represented multiple times when being compared to the reflection as a result of the larger interactions being broken up into several passages, we extracted the five unique gameplay

Table 5.4 Rubric used to label pairs of reflections and text representations of gameplay interactions. Underlined text indicates specific parts of the text that are considered to be related.

Label	Description	Reflection Example	Gameplay Interaction Example
Unlikely Source	The gameplay interaction could not be a source of information for this reflection.	“I have just tested an egg that tested positive for pathogenic bacteria. I am going to continue to test foods.”	“Tested Milk for Viruses because Sick members ate/drank it.”
Potential Source	The reflection includes language that could be related to the information covered by the gameplay interaction, but it does not make a clear and specific reference to the information derived from the gameplay interaction.	“ <u>I've learned about certain diseases and the symptoms caused by it.</u> I've found diseases that the symptoms don't match up with. I will keep finding information on the topics.”	“ <u>Botulism is a bacterial infection. The symptoms of botulism are muscle paralysis, vomiting, nausea, and stomach cramps.</u> The way to treat or prevent botulism is with an antitoxin.”
Likely Source	The reflection uses specific language that clearly indicates that the gameplay interaction is a source for this information.	“ <u>The illness is probably a pathogen because it had spread from one person to a few others.</u> My plan is to look for different possible sources of a pathogen.”	“Pathogens are biological organisms that cause disease or illness to their host. <u>The main characteristics of pathogens are that they make people ill, they require hosts, and can spread from one host to another.</u> ”

interactions with the highest cosine similarities for each sentence. As a result, we ended up with the top-5 most relevant gameplay interactions for each sentence in the reflection, as well as the specific text subsequence for each gameplay interaction that was found to have the highest similarity.

Then, to establish a ground truth mapping of reflections to gameplay interactions, two raters manually labeled each of the top-5 candidate interactions based on relevance to the reflection sentence. The top-5 interactions were randomly reordered prior to manual labeling to ensure that annotators were not influenced by the ordering of the top-rated gameplay interactions. A three-level rubric was used to assess the reflection-interaction pairs (Table 5.4), which captures whether a gameplay interaction represented a likely source, a potential source, or an unlikely source of information for the reflection. A likely source indicates that there is specific language used in the reflection that was drawn from the interaction, a potential source indicates that the student might be referring to the interaction in their reflection but they are not specific enough to clearly point to the interaction, and an unlikely source indicates that the interaction was almost certainly not being referred to in the reflection. Raters ignored any forward-looking components of a reflection, since the goal was to understand where students draw information from when they are evaluating their knowledge and learning strategies, not when they are thinking about what they should do in the future. Between the two raters, a shared 20% of the reflection-interaction pairs were labeled in common to compute inter-rater reliability. The raters achieved a Cohen's kappa of 0.703, indicating moderate to good agreement (Azen & Walker, 2021). For any labels that the raters disagreed on, they were discussed and reconciled to agree on a single label. The remainder of the reflection-interaction pairs were labeled by one of the raters.

Based on this mapping, we analyzed the interactions that students reflected on to derive some insights that could help explain students' reflection behaviors during game-based learning. First, we analyzed the extent to which students specifically reflected on information that they gathered from in-game interactions. Then, we split students into different groups based on their learning outcomes (i.e., high versus low normalized learning gain, solved versus did not solve the mystery) to compare the types of gameplay interactions that students in each group reflected on.

5.3.2 Results

First, looking at whether students reflected on specific gameplay interactions, we found that 68% (296/434) of all reflections were directly or partially related to at least one gameplay interaction.

38% of reflections (164/434) were directly related to at least one gameplay interaction, meaning that the student specifically mentioned information from that interaction. 39% of reflections (169/434) were partially related to at least one gameplay interaction, meaning that the interaction may have been a source of information for the reflection but there was no specific language used that would definitively indicate such a relationship. We assume that this is an underestimate, since the initial approach to identifying a candidate set of interactions that may be related to each reflection may not have correctly identified every relevant interaction that a student included in their reflection. Alternatively, this could be explained by the fact that students' reflections may include metacognitive, affective, and motivational components, which would likely not directly reference specific sources of information from the learning environment.

To evaluate whether there were differences in the gameplay interactions that high-performing versus low-performing students tended to reflect on, we split students based on their learning outcomes and compared the interactions that they reflected on. The distribution of the types of interactions students reflected on can be found in Figure 5.2. Note that not every student had complete data (i.e., both pre and post test assessments), so we filtered out any students who were missing data for these analyses. First, we split students based on their normalized learning gains (NLG) between the pre and post tests. The median NLG was 0.083, which roughly split the students into those who had positive learning gains after playing CRYSTAL ISLAND (N=60) and those who did not (N=58). For these analyses, we focused exclusively on gameplay interactions that were labeled as likely sources of information for a reflection, since that indicates that there is specific information from the interaction that was used in the reflection. It appeared that students in the low NLG group tended to reflect more frequently on information from book sources (low = 38.5%, high = 21.8% of interactions) and less frequently on scanner interactions (low = 19.2%, high = 30.9% of interactions) or information obtained from posters (low = 0%, high = 7.3% of interactions). However, a chi-square test for homogeneity between these two groups indicated that there were no statistically significant differences in the types of interactions that students from either group reflected on ($X^2(9, N = 118) = 5.7, p = 0.1268$).

Next, we split students into two groups based on whether they solved the mystery (N=64) or not (N=54) and performed the same analysis. A chi-square test for homogeneity between these groups indicated that there were significant differences in which gameplay interactions students reflected on ($X^2(9, N = 118) = 8.8, p < 0.05$). However, it should be noted that this test does not tell us which gameplay interactions had significant differences between the groups. Thus, we cautiously observe that the only notable difference was that students who did not solve the

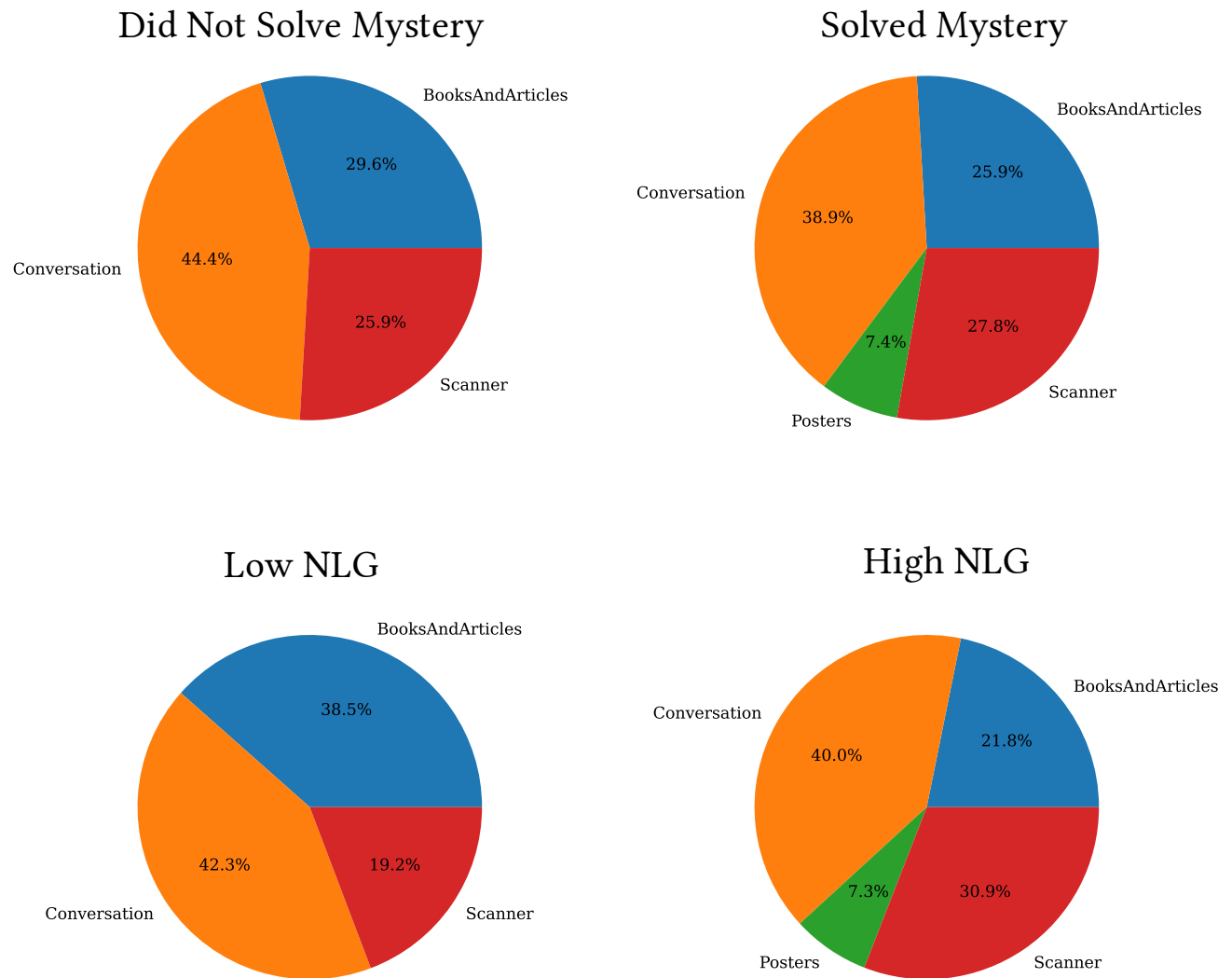


Figure 5.2 Distribution of types of gameplay interactions that students reflected on based on splits between high versus low normalized learning gain (NLG) and solving versus not solving the mystery.

mystery never directly referenced information from posters in their reflections, while students who did solve the mystery reflected on information from posters 7.4% of the time. This seems like a plausible explanation because posters are the only source of information in the game that mentions the symptoms of various diseases, which is a critical component of solving the game's mystery. It is not the case that students who did not solve the mystery never interacted with posters, as 9% of all gameplay interactions by these students were with posters, which is only slightly less than students who did solve the mystery (12%). Yet, the students who did not solve the mystery did not directly reference posters in their reflections. This suggests that these students may not have recognized the importance of the information included in the posters and thus struggled to connect that information with everything else they learned in order to reach the correct solution.

5.4 Examining Relationships Between Reflection Content and Reflection Depth

Next, we looked at relationships between the content of a reflection and the depth of the reflection to see if there was any interplay between what students decided to reflect on and how deep the resulting reflection was. In theory, these characteristics of students' reflections could be entirely independent, since it is possible to write a shallow or deep reflection on any topic. For example, given the topic "Understanding of Mutagens and Carcinogens", students can produce both deep (e.g., "carcinogens cannot spread mutagens cannot spread either meaning the disease in the camp is a pathogen because it is spreading through people.") and shallow reflections (e.g., "that pathogens need a host to spread. carngio make cancer."). However, given that our operationalization of reflection depth focuses on using reasoning to inform a hypothesis or make a plan for the future, there could be topics that were more closely associated with deep reflection.

5.4.1 Method

To investigate this research question, we split reflections into different groups based on the three phases of scientific inquiry that a reflection may focus on (i.e., Hypothesis Generation, Experimental Design, and Evidence Evaluation) and then based on the type of gameplay interaction that the student drew information from to write the reflection (i.e., Book, Conversation, Scanner, and Poster). Then, we performed one-way ANOVAs between each group.

Since there were more than two groups for each test and an ANOVA only indicates whether or not there are significant differences between the means of the groups but does not indicate which groups are significantly different, we performed Tukey post hoc tests to identify types of reflections that corresponded to different levels of reflection depth.

5.4.2 Results

First, we examined reflections across the three different scientific inquiry phases that we had mapped reflection topics to – Hypothesis Generation, Experimental Design, and Evidence Evaluation. We found that reflections in the Hypothesis Generation phase were shallower than those in Experimental Design and Evidence Evaluation (Table 5.5; $F = 11.518$, $p < 0.001$). A reflection in the Hypothesis Generation phase indicates that the student was gathering information to inform a hypothesis about what was happening to the people on the island. These information gathering actions can include learning foundational knowledge about microbiology, for instance by reading books or speaking with scientists on the island, or gathering clues about the specific problem-solving instance that was unfolding on the island by speaking with characters like the sick patients. In both cases, students who reflected on this sort of information tended to simply restate what they read without engaging in much information processing to evaluate why it was important or to synthesize multiple pieces of information to draw conclusions. As a result, reflections related to Hypothesis Generation were assigned low depth scores since, even though the information the student was reflecting on could be helpful for formulating a hypothesis, they were not yet doing so. This suggests that students should be advised to engage in more information processing in their reflections related to gathering information for creating hypotheses. For example, we could further prompt students to explain

Table 5.5 Average reflection depth scores across phases of scientific inquiry.

Scientific Inquiry Phase	Average Reflection Depth
Hypothesis Generation	2.130
Experimental Design	2.579
Evidence Evaluation	2.655

why they think that something they have learned is important, rather than simply asking them to state the most important things they have learned.

Next, we examined reflections on different types of gameplay interactions. Note that if a reflection was manually labeled as being related to more than one gameplay interaction, it was assigned to only one group based on which interaction had the highest semantic similarity score compared to the reflection. This heuristic was used because it provided a way to quantitatively approximate which interaction was most related to the reflection. We found that students' reflections on content that came from a poster were deeper than reflections on all other types of gameplay interactions (Table 5.6; $F = 8.000$, $p < 0.001$). This is likely because posters contain information that is required for students to make connections between the different pieces of information they've collected while playing CRYSTAL ISLAND. Posters often contain information about the type of contaminant that causes each disease and the symptoms associated with that disease, so students reflecting on content from a poster are likely to connect those facts with what they have observed in the game through talking with characters about their symptoms or obtaining results from the virtual scanner (e.g., "I have learned that one of the paitence has stomach cramps and a fever. I know that with an influenza has some of those symptoms so it could possible be the flu."). In comparison, reflections on books or conversations will often just restate theoretical information about microbiology (e.g., "the flu is a virus and that is why it can evolve.") and even reflections on test results from the scanner will often just restate what the scanner said without meaningfully processing what the student learned or drawing conclusions

Table 5.6 Average reflection depth scores across different in-game sources of information that were referenced in the reflection.

Gameplay Interaction Reflected On	Average Reflection Depth
Books and articles	2.143
Conversations with characters	2.444
Using the scanner	2.679
Interacting with posters	3.484

(e.g., “I have just tested an egg that tested positive for pathogenic bacteria. I am going to continue to test foods.”). This result suggests that it may be useful to guide students toward interacting with posters and then advise them to reflect on what they learned from the posters. For example, we could modify our prompting strategy to include a reflection trigger after students have interacted with several posters. We could introduce a prompt that draws attention to the importance of the content contained in the posters, noting that they are useful for making connections between all of the other information the student has collected.

5.5 Investigating Relationships Between Reflection Behaviors and Learning Outcomes

As a critical component of self-regulated learning, effective reflection should help students achieve improved learning outcomes. By reflecting effectively, students should assess the actions they have taken and what they have learned and use that to modify their learning strategies to help them achieve their goals. To empirically evaluate the relationship between students’ reflection behaviors and their learning outcomes during game-based learning with CRYSTAL ISLAND, we investigated whether science content knowledge and science problem-solving learning outcomes can be predicted by features of students’ reflections and problem-solving actions during learning. In particular, we investigated whether reflection depth and content were more predictive of learning outcomes than features that captured students’ gameplay behaviors.

These analyses only looked at students with complete data (i.e., pre-test, gameplay, and post-test) to ensure that analysis of students’ science content knowledge learning outcomes could control for prior content knowledge. In total, this used data from 105 students out of the 153 students who participated in the study. Results provide empirical support for the theoretical relationships between effective reflection and successful learning and problem solving. This motivates our work on developing systems that enable adaptive interventions aimed at improving students’ reflection skills.

5.5.1 Method

We used the same hierarchical regression technique with the same features derived from students’ problem-solving actions and reflections that were used in Chapter 5.1.1. Using this technique, we explored whether there were features of students’ reflections that were predictive of their content knowledge and problem-solving learning outcomes. We defined content

knowledge learning outcome as students' performance on the post-test that was administered after students interacted with CRYSTAL ISLAND. The problem-solving learning outcome was defined as whether or not the student solved the mystery of the disease that was spreading on the island.

In addition to the reflection features that were presented in Chapter 5.1.1, we also introduced features that captured the content of students' reflections. These were count-based features that captured the number of different types of reflections that students wrote. First, we captured the reflections that students wrote on each type of gameplay interaction - the number of reflections on information from posters, the number of reflections on information from conversations with characters, the number of reflections on information from a book, the number of reflections on information taken from the scanner. We also captured the number of reflections that students wrote that corresponded to each of the three phases of scientific inquiry - the number of reflections related to hypothesis generation, the number of reflections related to experimental design, and the number of reflections related to evidence evaluation. These features allowed us to investigate whether reflecting on a type of interaction is more predictive of learning outcomes than simply completing an interaction in the game (e.g., number of reflections on book content versus the number of books read) and if the scientific inquiry phase that a student engages in during learning can predict their learning outcomes.

Finally, we trained random forest classifiers using different sets of features to evaluate which set of features would result in the highest predictive performance when predicting students' learning outcomes. We compared feature sets using only features derived from gameplay interactions, features derived from gameplay interactions plus features that captured basic information about students' reflection behaviors, and features that additionally included information about the content of students' reflections (i.e., reflections corresponding to different scientific inquiry phases and reflections on different types of gameplay interactions). Models were trained to predict whether or not the student solved the mystery, whether they had a high or low post test score compared to all other students (determined by a median split), and whether they had a high or low normalized learning gain compared to all other students (determined by a median split). Normalized learning gain captures how much a student learned compared to how much they could have learned given their performance on the pre-test.

5.5.2 Results

Hierarchical regression analysis was conducted in four steps, with results shown in Table 5.7. First, we predicted the outcome variables using only students' pre-test scores and found this to be a significant predictor of post-test score ($\beta=0.496$, $p < .001$) and a significant predictor of successfully solving the mystery ($\beta=0.010$, $p < .05$). Next, we incorporated features related to students' problem-solving actions that were based on our reflection prompting triggers. When predicting post-test score, none of these features were found to be significant. When predicting whether the student solved the mystery, the number of plot points completed ($\beta=0.022$, $p < .001$) and a positive test result ($\beta=0.411$, $p < .001$) were both found to be significantly predictive. For the third step, we extracted features that summarize how students reflected over the course of their interactions with CRYSTAL ISLAND. For predicting post-test score, average reflection depth rating ($\beta=0.841$, $p < .05$) and the average amount of time spent on each reflection ($\beta= -0.028$, $p < .05$) were found to be significant. Looking at whether the student solved the mystery, these two features were again found to be significant ($\beta=0.058$, $p < .05$; $\beta= -0.004$, $p < .05$). We also incorporated a subset of LIWC features, which capture emotional, cognitive, and structural components of natural language. For predicting post-test score, none of the LIWC features were found to be significant. For predicting whether the student solved the mystery, *Focuspast* was nearly found to be significant ($\beta=0.042$, $p < .1$). Finally, we incorporated features that captured the content of students' reflections. These counted the number of reflections students had on different types of gameplay interactions and the number of reflections that corresponded to each of the three phases of scientific inquiry. From these features, the number of times that students reflected on content found in a book was predictive of post-test score ($\beta=0.904$, $p < .05$).

Results align with Sabourin et al. (2011), where relationships between problem-solving actions and science content knowledge learning outcomes were not observed, as well as with prior research that has demonstrated the benefits of reflection features for predictive student modeling (Geden et al., 2021). To examine the extent to which learning and performance outcomes Results suggested that features extracted from students' reflections were predictive of post-test scores, but problem-solving actions were not. Based on previous work with data from a different CRYSTAL ISLAND study (Sabourin et al., 2012), we had expected students' problem-solving actions to offer insufficient information for predicting science content knowledge learning outcomes. Yet, while students' problem-solving actions were not predictive of post-test

scores, we found that some features of their reflections were, which is consistent with prior findings by Geden et al. (2021). In particular, students' average reflection depth and the average amount of time spent on each reflection were predictive of post-test scores. These results suggest that students who are deeper or more efficient reflectors may be more successful in this learning environment. Interestingly, the average number of words per reflection and the number of domain-specific words used in students' reflections were not predictive of post-test scores. Thus, the amount of externalized reflection seems to be less important than the depth and efficiency of reflection for predicting learning outcomes. Additionally, we found that the number of reflections that students wrote on content found in books was a significant predictor of post-test score. This is an interesting finding, because the number of books that students read was not a significant predictor of their post-test score. Thus, we can see that there is a relationship between the act of reflecting on a piece of information and the extent to which a student has learned that information. The post-test primarily consists of information that is found in books, so this result suggests that students who read book content and then identified that information as important enough to reflect on integrated that information into their knowledge to a greater extent than students who may have read similar content but did not deem it worthy of reflection.

Next, we analyzed whether there were relationships between successfully solving the game's mystery and students' problem-solving actions and reflection features. Results from hierarchical regression analyses suggested that some problem-solving actions were predictive of successful problem solving. Specifically, the number of plot points completed and whether the student received a positive test result on one of their virtual tests for a pathogen were positively related to solving the mystery. Since the goal of CRYSTAL ISLAND requires students to identify the type of pathogen that is spreading and the food item through which it is spreading, it is obvious why receiving a positive test result would be predictive of solving the mystery - after getting the positive result, all students need to do to complete the game is determine a viable treatment or prevention plan and present this information to the camp nurse. As for the number of plot points completed, this feature helps to capture how widely students have explored the virtual environment. Thus, it seems reasonable that students who have explored the game more completely would be able to use the information they collected to solve the mystery. This relationship between problem-solving actions and problem-solving performance aligns with prior findings (Spires et al., 2011). As for features related to students' reflections, average

reflection depth and the average amount of time students spent writing each reflection were found to be predictive of successful problem solving, such that deeper and more efficient reflection was positively related to problem-solving success. For predicting problem-solving success, none of the features that captured the content of students' reflections were found to be predictive.

Table 5.7 Hierarchical regression for predicting post test score and whether the student solved the mystery. All regression coefficients are from the final step in the analyses.

Predictor	Post-Test Score			Mystery Solved		
	β	ΔR^2	F	β	ΔR^2	F
Step 1: Prior Content Knowledge		282.026	33.546***		0.795	4.987*
Pre-Test Score	0.496			0.010		
Step 2: Problem-Solving Actions		21.437	0.510		7.688	9.648***
Number of Plot Points Completed	0.007	1.865		0.022	2.607***	
Number of Books Read	0.000	4.288		-0.009	0.512†	
Number of Worksheet Submissions		7.073		0.038	0.031	
Number of Lab Tests	-0.011	3.401		-0.001	0.063	
Obtained a Positive Test Result	0.063	4.809		0.411	4.475***	
Step 3: Reflection Features		140.618	1.858†		2.923	2.038*
Reflection Depth Rating	0.841	46.133*		0.058	0.603*	
Number of Words	0.042	3.251		0.009	0.064	
Number of Keywords	0.183	14.711		0.057	0.262	
Time Spent Reflecting	-0.028	43.168*		-0.004	1.076*	
Focuspast	0.292	16.237		0.042	0.492†	
Focusfuture	0.343	5.980		-0.020	0.047	
Authentic	-0.011	3.193		-0.003	0.115	
I	-0.049	5.570		0.017	0.071	
Analytic	0.014	2.374		0.003	0.193	
Step 4: Reflection Content Features		51.184	0.761		0.675	0.529
Num Reflections on Posters	0.522	1.238		0.159	0.024	
Num Reflections on Conversations	0.281	0.343		-0.064	0.125	
Num Reflections on Books	0.904	34.822*		-0.015	0.005	
Num Reflections on Scanner	-0.670	5.200		-0.049	0.133	
Num Reflections Hypothesis Gen.	-0.094	0.023		0.045	0.144	
Num Reflections Experimental Des.	-0.686	8.426		0.014	0.003	
Num Reflections Evidence Eval.	-0.088	0.137		-0.046	0.062	

*** p < .001, ** p < .01, * p < .05, † p < .1

Finally, results of evaluating random forest classifiers on different sets of training data (i.e., gameplay-only, gameplay plus reflection, and gameplay plus reflection features and derived content-based reflection features) demonstrate that different learning outcomes are best predicted by different features. For predicting whether a student solved the mystery, models performed best when trained exclusively on features that captured students' gameplay interactions. To predict students' post-test scores, combining gameplay interaction features with non-content-based reflection features achieved the highest performance. Models that were trained on gameplay interaction features, non-content-based reflection features, and features derived from the content of students' reflections achieved the highest performance for predicting normalized learning gain. This suggests that the most important information for predicting if students will solve the mystery is the actions students take in the game, the most important information for predicting their overall amount of knowledge at the end of the game is how and how well students reflect (e.g., time spent reflecting and reflection depth), and to predict whether students will increase their knowledge relative to when they started the game it is important to focus on the content of students' reflections (e.g., the number of times that they reflected on book content).

These results show that students' reflections were predictive of learning outcomes more than problem-solving actions alone. This points to the importance of developing analytical

Table 5.8 Model performance for predicting students' learning outcomes using different sets of features derived from gameplay interactions and reflections.

Features	Learning Outcome			
		Solved Mystery	Post Test Score	Normalized Learning Gain
Gameplay	Acc	0.825	0.628	0.535
	F1	0.855	0.556	0.533
Gameplay + Simple Reflection	Acc	0.752	0.695	0.611
	F1	0.803	0.637	0.606
Gameplay + Simple Reflection + Content-Based Reflection	Acc	0.768	0.657	0.699
	F1	0.807	0.559	0.679

methods for automatically understanding reflections during gameplay that can be used to inform adaptive features to support students in the learning environment.

5.6 Examining Changes in Students' Reflection Behaviors Over Time

As students interact with CRYSTAL ISLAND, it is possible that their reflection behaviors may change over time. Understanding if and how students' reflections change over time may provide insight into students' reflection behaviors that could inform the design of tools that aim to support reflection during game-based learning. To explore students' reflection behaviors over time, we first explored whether students already engaged in deep reflection or learned to reflect more deeply simply through unguided practice, thereby evaluating whether there is even a need to provide support for reflection. Previous work suggests that students may require substantial guidance on how to reflect effectively (Riedinger, 2006). In the absence of such support, we expect that the depth of students' reflections will not increase over the course of their learning experience and may even decrease.

5.6.1 Method

Latent growth curve modeling provides a framework for modeling trajectories of repeated measures over time for a group and can control for factors that may account for different trajectories. Growth models estimate an initial value for the variable being modeled as well as an estimate of how that variable changes over time. We investigated the growth curves of students' reflection depth ratings over subsequent prompts for reflection during CRYSTAL ISLAND. In addition, we accounted for students' pre-test scores to determine whether students' prior content knowledge impacted their depth of reflection.

Among the 105 students who had complete data, students had different numbers of in-game reflections. Four students only completed one in-game reflection, so they were omitted from latent growth curve analysis because a single observation is insufficient to model the trajectory of reflection depth over repeated reflections. The students who completed exactly two in-game reflections (N=20), exactly three in-game reflections (N=43), and exactly four in-game reflections (N=38) were modeled by a single growth curve model, since growth curves can account for partially missing data (Curran et al., 2010).

5.6.2 Results

Results for a growth model accounting for linear growth and a no-growth model representing the null hypothesis that students' reflection depth did not change over time are presented in Table 5.9. According to a Chi-square goodness of fit test, the linear growth model very nearly fit the data ($p < 0.1$) but the no-growth model did not. We found that students' reflection depth over the course of several prompts for reflection could be described by a line with an intercept of 2.263 and a slope of -0.003. That is, for students' first reflections during their interactions with CRYSTAL ISLAND, they typically received a reflection depth rating of 2.263. The slope of -0.003 indicates that the depth of students' reflections barely changed over the course of subsequent reflections. Although the average reflection depth for students' fourth in-game reflection ($M=2.42$, $SD=0.75$) was higher than the average reflection depth for their first reflection ($M=2.30$, $SD=0.65$), reflection depth does not change significantly. To explore further, we compared the linear growth model to the baseline no-growth model, where students' reflection depth was found to remain at a constant score of 2.258, using a Chi-square difference test. We found that the two were not significantly different, indicating that we cannot reject the null hypothesis that students' reflection depth does not change over the course of several reflection responses. The no-growth trajectory compared to students' individual trajectories can be seen in Figure 5.3. The most common student trajectories are represented by darker lines in the graph.

To examine the effect that having a high or low pre-test score, as determined by a median split (median=7, $N_{low}=45$, $N_{high}=56$), has on students' initial reflection depth and changes in reflection depth over time, we included pre-test as a time-invariant covariate in our latent growth curve model and found that there were no significant effects of students' pre-test performance on reflection depth during CRYSTAL ISLAND.

Table 5.9 Growth model parameter estimates with standard error in parentheses and fit indices.

* $p < .05$

	Intercept	Slope	df	Chi-Square	AIC	BIC
Linear Growth	2.263 (0.147)*	-0.003 (0.061)	10	17.099	314.19	327.29
No Growth	2.258 (0.134)*	0	14	20.233	309.32	315.87

We found that students' reflection depth did not change significantly over the course of their learning experiences in CRYSTAL ISLAND. Based on Riedinger (2006), we had hypothesized that, in the absence of any support for reflection, reflections would not increase in depth over time and might even decrease in depth. We hypothesized that students might become less motivated to engage in deep reflection over time and would instead quickly provide a shallow reflection to satisfy the bare minimum requirement, leading to a decrease in reflection depth scores. Additionally, since we have observed that some students get stuck in their investigations and have difficulty making progress toward solving the mystery, we hypothesized that students might engage in shallower reflection over time as a combined result of increased frustration and a lack of new information to reflect on.

Using latent growth curve modeling, we found that students' reflection depth remained mostly constant over the course of their learning experience. According to the growth model of students' reflection depth, depth scores tended to stay at approximately 2.3. This roughly

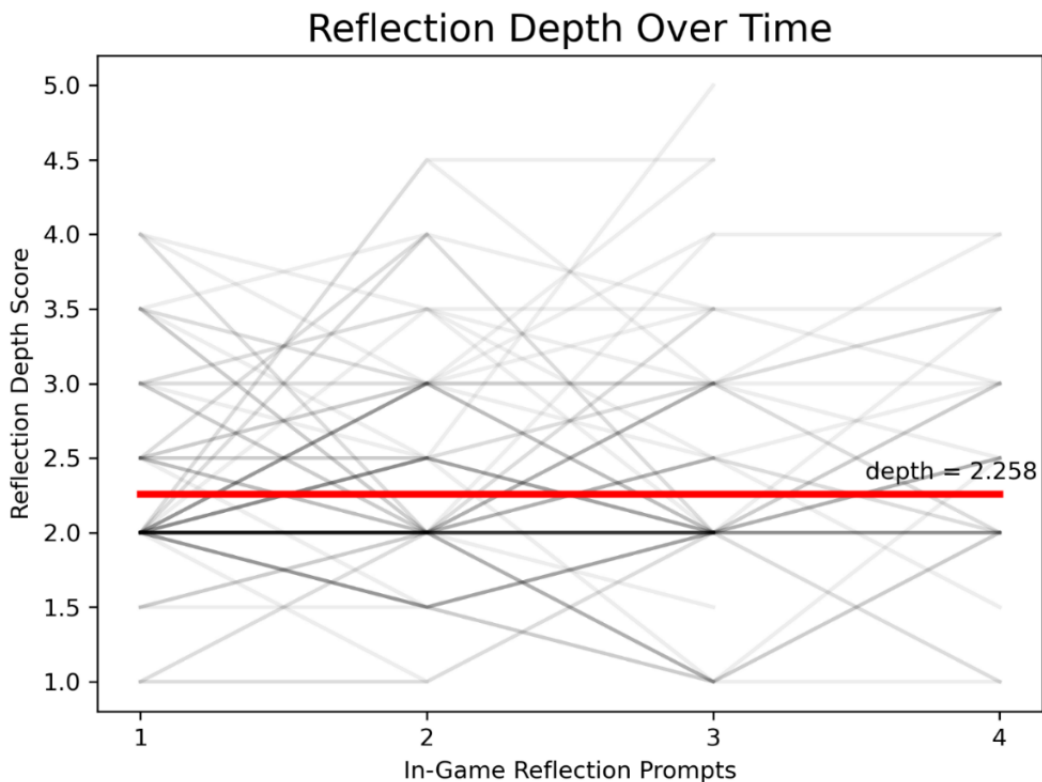


Figure 5.3 Students' reflection depth over the course of their interaction with CRYSTAL ISLAND. Darker lines represent more common trajectories, and the red line represents the model implied trajectory over all students.

corresponds to a score of 2 on the rubric, which indicates that students' reflections were either too vague to be very useful or were mostly direct restatements of information found in the game without any processing by the student. This demonstrates that students were often aware of their knowledge and learning goals, but did not evaluate this knowledge to help drive adaptations for the future. Moreover, we found that there was not a significant impact of students' prior science content knowledge, as indicated by their pre-test scores, on reflection depth. Thus, since all students' reflections were consistently shallow, these results demonstrate a need to provide support for reflection during game-based learning.

CHAPTER 6

ANALYZING STUDENTS' REFLECTIONS IN THE CONTEXT OF THEIR GAMEPLAY INTERACTIONS

This chapter describes the work done to analyze students' reflections in the context of the gameplay actions they took while interacting with CRYSTAL ISLAND. Most of the work done to analyze students' reflections focuses on reflection as a post-learning exercise, such as reviewing study behaviors at the end of a curricular unit. However, reflection is also important as a mid-process action where students think back on what they have learned and make adjustments to their immediate next actions. This is especially relevant during science problem solving, where a student's goals are likely to be constantly changing. Since reflection is goal-oriented, it is critical for students to evaluate the progress they are making toward their goals, identify whether the goals they are pursuing have changed, and reorient if necessary to ensure that their learning and problem-solving activities help them achieve their goals.

Integrating reflection prompts in a game-based learning environment presents a unique opportunity to analyze students' reflections in the context of their learning activities, since there is rich trace data that captures each action students take prior to each of their reflections. In this chapter, we first link each reflection to the student's recent gameplay interactions in CRYSTAL ISLAND that were likely to be the sources of information that the student drew from when writing their reflection. Based on this mapping between reflections and relevant gameplay interactions, we explore differences in what types of gameplay interactions students decide to reflect on across students with disparate learning outcomes. Next, we use sequential pattern mining techniques to identify gameplay patterns that commonly precede reflections on certain types of gameplay interactions or certain topics. Then, we evaluate the extent to which machine learning models that are trained on students' event sequences can predict what students will reflect on, since being able to predict this may allow an adaptive system to provide preemptive nudges toward certain topics that may be more beneficial to reflect on at a given time. Finally, we investigate whether Transformer-based machine learning models can automatically perform the linking between reflections and relevant gameplay interactions, since the manual linking process

is labor-intensive and would not be able to support real-time interventions based on this information.

6.1 Sequential Pattern Mining to Identify Associations Between Gameplay Patterns and Reflection Content

The trace data collected by CRYSTAL ISLAND provides a rich representation of every action that students take while interacting with the learning environment. By performing sequential pattern mining on students' sequences of gameplay interactions, we may be able to identify patterns in students' gameplay behaviors that commonly preceded different types of reflections (e.g., reflections related to different phases of scientific inquiry). This would help us to understand how students' reflection behaviors are foreshadowed by the actions that they have recently taken during game-based learning, which could inform insights about students' gameplay behaviors and their relationship to reflection while also forming the foundation for predictive modeling that uses gameplay interactions to predict how students will reflect.

6.1.1 Method

For these analyses, we defined a gameplay context as the sequence of gameplay interactions (e.g., reading a book, using the virtual scanner) that the student had taken since their previous reflection. We performed differential sequence mining (Kinnebrew, 2013) to identify gameplay patterns that more frequently resulted in reflections on one topic as opposed to the other. Differential sequence mining uses a sequential pattern mining algorithm to identify frequent patterns within two groups and then performs t-tests to determine whether the frequency of each identified pattern is significantly different across the two groups. In particular, the algorithm calculates i-support, which is the total number of times a pattern occurs in all sequences belonging to one group, and compares that value across groups to determine if the pattern appears significantly more in one group versus the other. It is important to note that differential sequence mining is an exploratory technique without any guarantees that the differences in the frequencies of identified patterns across the two groups are statistically significant. Nonetheless, this technique is sufficient for our purposes of exploring how different gameplay patterns may be associated with different types of reflections.

The groups that we compare across are the three phases of scientific inquiry that a reflection may focus on (i.e., Hypothesis Generation, Experimental Design, and Evidence

Evaluation) and the type of gameplay interaction that the student drew information from to write the reflection (i.e., Book, Conversation, Scanner, and Poster). Since differential sequence mining is used to compare between two different groups yet we explored differences between groups of size three and four, we ran the algorithm multiple times using a binary split for each variable. For example, to identify differences in gameplay patterns preceding a reflection that was assigned to the Hypothesis Generation phase, we compared between reflections assigned to Hypothesis Generation and those that were not assigned to Hypothesis Generation. To represent the sequences of students' gameplay interactions, we abstracted each gameplay interactions out to the event type. That is, reading the book "Viruses" in the infirmary for 32 seconds would be represented simply by the label "Book". We initially explored more detailed representations that included the target of the interaction (e.g., the title of the book that a student read), but there was too much information to be able to extract meaningful insights from the identified gameplay patterns. Since differential sequence mining is an exploratory technique and interpreting the identified patterns is the primary goal, we decided to use the abstract representation.

We used the differential sequence mining algorithm to search for patterns that occurred in at least twenty percent of all sequences that preceded a reflection. We searched for patterns with a maximum length of three interactions and allowed for a gap of size two between each interaction so two students did not have to perform the exact same sequence of interactions for the sequences to be considered the same. For example, if a student had the sequence "Book, Book, Conversation, Poster, Book" and another student had the sequence "Book, Book, Book", the algorithm would ignore up to two different interactions between common sequences (i.e., "Conversation, Poster") and would identify "Book, Book, Book" as a common sequence between these two students.

6.1.2 Results

We identified several patterns that frequently occurred prior to reflections that corresponded to different phases of scientific inquiry. As shown in Table 6.1, students typically had repeated interactions with the scanner prior to writing a reflection belonging to the Evidence Evaluation phase (e.g., Scanner → Scanner → Scanner). This is likely because reflections on the results of a scan were often mapped to Evidence Evaluation, since that is the main source of evidence that students can collect in CRYSTAL ISLAND. So, a student who is repeatedly engaging in scanning interactions will likely focus on what they found using the scanner. Conversely, reading a book,

having a conversation with a character, or interacting with the diagnosis worksheet were actions that were associated with reflections in the other phases of scientific inquiry. While reading books and speaking with characters are mostly related to gathering information, which may be more closely aligned with the Hypothesis Generation phase, the worksheet is the tool that students use to record their findings. So, we would probably expect that interacting with the worksheet would be related to Evidence Evaluation. However, since the interaction with the worksheet occurs in the context of information gathering actions (i.e., reading a book and talking with a character), it is likely that students are using the diagnosis worksheet as a place to take notes that will help them formulate a hypothesis rather than using it to combine their findings into a final diagnosis.

As shown in Table 6.2, students were also likely to interact with the scanner prior to a reflection that corresponded to the Experimental Design phase of scientific inquiry, but they were also likely to have interacted with their diagnosis worksheet (e.g., Scanner → Worksheet). This pattern may indicate that students were engaged in a deductive reasoning or trial-and-error process, where they were conducting tests with the scanner to cross potential disease-carrying foods off of the list of possibilities rather than using it to confirm that a suspected food was contaminated. This is in some ways the opposite approach to experimental design that may be indicated by interactions with the worksheet followed by conversations with character (i.e., Worksheet → Worksheet → Conversation). This pattern suggests that students were using the diagnosis worksheet to keep track of the information they had collected, such as food items that may be contaminated because sick people had eaten them. In this way, students may have been building up potential hypotheses that they intended to evaluate in the future.

These differences in the ways that students used the diagnosis worksheet are further clarified by the patterns identified when comparing interaction sequences prior to Evidence Evaluation reflections and sequences prior to Experimental Design reflections (Table 6.3). When students used the worksheet in the context of the scanner (e.g., Worksheet → Scanner) they were likely to produce a reflection that was related to Evidence Evaluation, but when they used the worksheet in the context of an information gathering interaction (e.g., Worksheet → Worksheet → Conversation; Book → Worksheet), they were more likely to write a reflection related to Experimental Design. This presents an interesting view of the role of the diagnosis worksheet as either a tool for designing experiments to narrow down the large set of possible

explanations for the mysterious disease or for building up a final diagnosis based on evidence gained from experimental results.

Table 6.1 Differentially frequent patterns - Evidence Evaluation vs Other

Pattern	I-Support Difference
Scanner	6.222
Scanner → Scanner	2.995
Scanner → Scanner → Scanner	1.988
Conversation → Scanner → Scanner	1.042
Book → Worksheet	-1.952
Conversation → Book	-1.312
Worksheet → Conversation → Conversation	-1.016

Table 6.2 Differentially frequent patterns – Experimental Design vs Other

Pattern	I-Support Difference
Scanner → Worksheet	4.107
Scanner	4.045
Scanner → Scanner	2.162
Scanner → Scanner → Scanner	1.732
Worksheet → Worksheet → Conversation	1.242

Table 6.3 Differentially frequent patterns – Evidence Evaluation vs Experimental Design

Pattern	I-Support Difference
Worksheet → Scanner	2.564
Book → Worksheet	-2.630
Worksheet → Worksheet → Conversation	-1.964
Worksheet → Worksheet	-1.655
Book → Conversation → Book	-1.533
Book → Book → Worksheet	-1.352
Worksheet → Conversation → Conversation	-1.255

Table 6.4 Differentially frequent patterns - Poster vs Other

Pattern	I-Support Difference
Conversation	4.155
Book → Worksheet → Worksheet	3.107
Book → Book → Worksheet	2.899
Worksheet → Worksheet → Book	2.572
Conversation → Conversation → Book	2.328
Conversation → Book → Conversation	1.539

When looking at patterns that commonly preceded reflections on certain types of gameplay interactions, we unsurprisingly found that gameplay interactions of a certain type tended to precede reflections on that type of interaction. For example, the pattern Scanner → Scanner often occurred before reflections related to information gathered from the scanner.

However, we also found that reflections on poster content were often preceded by other types of gameplay interactions (Table 6.4). Reflections on poster content were often preceded by interactions with the diagnosis worksheet, books, and conversations with characters (e.g., Book → Worksheet → Worksheet; Conversation → Conversation → Book). This may point to the nature of the posters in the game as the main source of information that connects other sources of information together. So, when students interact with many different sources of information or are trying to make sense of the information that they have been keeping track of in the worksheet, they are more likely to reflect on the content from a poster since it connects everything together.

6.2 Predicting Reflection Content Based on Gameplay Interactions

Based on the results from differential sequence mining, it seems that there are some relationships between the actions students take in the game and the types of reflections that they produce. Thus, it may be possible to train predictive models to predict the scientific inquiry phase or type of gameplay interaction that a student will reflect on their gameplay interaction sequences. With the ability to predict the content of students' reflections based on their gameplay behaviors, an adaptive system may be able to guide students toward certain types of reflections and away from others.

6.2.1 Method

We evaluated several machine learning approaches for predicting the content of a student's next reflection based on their recent gameplay interactions. Static models that took in a summary of the student's gameplay interactions were compared to sequential neural network-based models that took a sequence of gameplay interactions as input. We expect that sequential models would outperform the static models, since the previous chapter identified sequences of gameplay interactions that were differentially frequent across different types of reflections. Additionally, it may be the case that when a student completed a gameplay interaction would influence whether or not they would reflect on it. For example, some students may have a tendency to always reflect on the most recent thing that they did rather than being more deliberate about what they should reflect on.

For static machine learning models, we evaluated the performance of logistic regression, random forest, and feedforward neural network models. We evaluated two different feature

representations for these models – a count-based summary of a student’s 50 most recent gameplay interactions and a duration-based summary. These gameplay interaction sequences capture what students did in the game-based learning environment since their previous reflection (or the start of the game) up to before the reflection that is being used as the outcome variable. The count-based summary simply sums the number of times a student had an interaction of each type (i.e., Book, Poster, Conversation, Scanner, and Worksheet). We also wanted to account for the fact that students did not spend equal amounts of time on each interaction, since we assume that the amount of time a student spent on an interaction might indicate the extent to which they remembered that interaction or deemed it important for achieving their goals. Thus, the duration-based feature representation sums the total duration of each type of interaction a student had during the 50 most recent gameplay interactions. The logistic regression and random forest models were configured with default parameters as defined by scikit-learn and the feedforward neural network had one hidden layer with 25 neural units and twenty-five percent dropout between the hidden layer and the output layer. This model was trained for 100 epochs.

For the sequential machine learning model, we evaluated the performance of an LSTM-based neural network. To ensure that the sequential model had access to the same information as the static models, we passed in the sequence of the 50 previous interactions that a student had taken prior to the current reflection. If the student had not completed 50 interactions between the previous reflection and the current reflection, their sequence was zero-padded to get it to a length of 50. The LSTM model consisted of two hidden layers with 25 neural units in each layer and twenty-five percent dropout between each layer. These models were trained for a maximum of 500 epochs to try and allow them to sufficiently fit to the data.

The target variables that each machine learning model was trained to predict were the scientific inquiry phase of the current reflection and the type of gameplay interaction that the reflection drew information from (i.e., the same features used to split reflections into groups to perform differential sequence mining in the previous chapter). As these were classification tasks, we present accuracy and F1 scores, F1 score, which is the harmonic mean of the precision and recall metrics. These metrics indicate how frequently a classification model correctly classifies a data point belonging to a certain class (precision) and how many of the instances of that class the model correctly identified out of all the instances that it should have identified. The F1 score

represents a single metric that balances between precision and recall so we can evaluate how well our models are able to correctly classify instances of each class. Since this is a multiclass classification problem, we report an F1 score that was averaged across all classes. All machine learning models were evaluated with 10-fold student-level cross-validation. As a baseline, a majority classifier that always predicted the majority class was included.

6.2.2 Results

Results for predicting the type of gameplay interaction that a student would reflect on are shown in Table 6.5. This was a 4-class classification problem, since the potential types of interactions that students reflected on were Book, Poster, Conversation, and Scanner. Any reflections that were not labeled as related to a gameplay interaction were filtered out for this analysis. All models substantially outperformed the majority baseline, both in terms of accuracy and F1. The highest performing model by a wide margin was the feedforward neural network that was trained with static duration-based features (accuracy = 0.745, F1 = 0.718). In general, the static duration-based features outperformed the static count-based features for predicting which type of interaction a student would reflect on. This suggests that the amount of time that a student spends doing something in the game does have a substantial impact on what they reflect on. For

Table 6.5 Results for predicting the type of gameplay interaction that a student would reflect on based on their recent gameplay interactions.

Feature Representation	Metric	Model				
		Majority Baseline	Logistic Regression	Random Forest	FFNN	LSTM
Static Count-Based	Acc	0.445	0.529	0.513	0.644	-
	F1	0.184	0.411	0.443	0.557	-
Static Duration-Based	Acc	0.445	0.553	0.587	0.745	-
	F1	0.184	0.451	0.473	0.718	-
Sequential	Acc	0.445	-	-	-	0.138
	F1	0.184	-	-	-	0.058

example, if a student spends a significant amount of time speaking with characters, they may be more likely to recall those interactions when they are asked to reflect as opposed to several short interactions that they had with book content.

Results for predicting the scientific inquiry phase associated with the next reflection are shown in Table 6.6. This was a 3-class classification problem since the models were trained which of the three phases of scientific inquiry each reflection corresponded to. If a reflection was not related to a phase of scientific inquiry, it was filtered out for this analysis. All models substantially outperformed the majority baseline in terms of F1 but several did not in terms of accuracy. Again, the feedforward neural network outperformed all other models, but this time the count-based features performed slightly better than the duration-based features. This suggests that the amount of time that students spent on each type of interaction was not more important for determining the phase of scientific inquiry for their next reflection compared to the number of times that they performed each type of interaction. For example, reading five books may be a stronger indicator that a student is in the Hypothesis Generation phase since they are searching widely for new information, whereas a student who spends several minutes reading a single book may be trying to establish foundational knowledge required to formulate

Table 6.6 Results for predicting the scientific inquiry phase of a student’s next reflection based on their recent gameplay interactions.

Feature Representation	Metric	Model				
		Majority Baseline	Logistic Regression	Random Forest	FFNN	LSTM
Static Count-Based	Acc	0.707	0.711	0.680	0.742	-
	F1	0.275	0.353	0.389	0.439	-
Static Duration-Based	Acc	0.707	0.684	0.685	0.738	-
	F1	0.275	0.322	0.384	0.434	-
Sequential	Acc	0.707	-	-	-	0.166
	F1	0.275	-	-	-	0.092

a hypothesis or they may be reviewing information they had previously found to evaluate new data that they have collected.

Across both predictive tasks, sequential models significantly underperformed compared to the static models. They tended to overfit, with the models trained to predict the type of interaction that a student would reflect on only ever predicting that the focus of the reflection would be on a book. We suspect that this poor performance may be explained by the small size of this dataset and the sparse feature representation used to represent students' gameplay interaction sequences. LSTM models are fairly complex, so they tend to scale well with additional data and can learn complex representations of multi-dimensional data. Perhaps these LSTM-based models could be improved by incorporating a richer representation of students' gameplay interaction sequences, such as a representation that captures the specifics of each interaction. For example, we could include the name of a book that the student read rather than just capturing that the student read some book. We did not use such a representation in this work because we wanted to maintain a fair comparison with the feature representation used by the static models, but this is a promising direction for future work that aims to improve the performance of these models.

6.3 Automatically Linking Reflection Content to Sources of In-Game Information

In Chapter 5.3, we established a manual mapping between the gameplay interactions that students took in the game and the content of their reflections. The process for manually generating the mapping between was very time-consuming due to the combinatorial nature of having to perform a pairwise comparison between each piece of information a student came across and each of their reflections. As a result, the manual mapping is useful for informing offline analytics related to the content of students' reflections, but it cannot be used to drive an AI-enabled system that can deliver adaptive guidance and feedback based on what students decide to reflect on. In this chapter, we present an evaluation of several NLP-based machine learning models that were trained to predict the specific gameplay interaction that a student has focused on in their reflection.

In comparison to the predictive models created in Chapter 6.2, which aimed to predict the content of a student's next reflection based on their recent gameplay interaction, the predictive

models presented in this chapter aim to analyze the content of a student’s reflection in addition to a text representation of their recent gameplay interactions to identify the specific gameplay interaction that the student is referring to. For example, an interpretation of a prediction presented in this chapter would not be that the student is likely to reflect on something they recently read in a book, but that the student’s reflection is likely based on information that the student took from the book titled “How Diseases Spread”. Such predictive models would support very fine-grained feedback on the content of students’ reflections. For example, there are some educational resources in CRYSTAL ISLAND that present valid microbiology knowledge that will never be directly used to help solve the mystery of the spreading illness. Identifying that a student has reflected on such a piece of information would be useful, since an adaptive system could suggest that the student reevaluate what they have decided to reflect on based on what their current goals are. As a result, the adaptive system may be able to guide students toward more productive reflections that will address information that is more relevant to solving the mystery.

6.3.1 Method

We used FLAN-T5 (Chung et al., 2022) to automatically link students’ reflections to relevant in-game sources of information that they had recently interacted with. We used the base FLAN-T5 model, which has 250M trainable parameters. Three different modeling approaches were evaluated. First, we evaluated a zero-shot approach that only provided instructions on how to perform the task. Since FLAN-T5 is an instruction-tuned language model, it is possible that it

```
Indicate whether the provided source of information is related to the given reflection. An information source is related if the content of the reflection is at least partially based on what the information source says .
```

```
Information Source: {{ action }}
```

```
Reflection: {{ target_reflection }}
```

```
Answer (Related or Not Related):
```

Figure 6.1 Prompt used to instruct FLAN-T5 on how to link students' reflections with relevant gameplay interactions.

would be able to successfully perform this reflection-interaction mapping task without any training data simply by being instructed on what to do. This is similar to a humans' ability to flexibly apply their knowledge to new tasks and has been demonstrated as a feature of large language models (Kojima et al., 2022). To this end, we constructed a prompt for this task (Figure 6.1). The prompt was designed to instruct the model on its task, which was to indicate whether a source of information was related or not related to a reflection. We also evaluated a standard fine-tuning approach that provided instructions and also fine-tuned the model on several examples. We fine-tuned the model using LoRA, a parameter-efficient method for fine-tuning large models by only changing some of the parameters. This helps to save time and computational resources while typically achieving similar performance. Finally, we evaluated a fine-tuning approach that did not include instructions and trained the model exclusively on the text of a reflection and the text representation of a gameplay interaction.

Each of the top-5 most similar interactions identified for each reflection, as discussed in Chapter 5.3, was used as a sample for this analysis. The target variable for the prediction task was a binary variable that indicated whether or not a gameplay interaction was related or not related to a given reflection. This was derived from the existing set of labels by combining Potential Source and Likely Source into one class that indicates whether a source of information is at all related to a reflection. These models were evaluated with 10-fold student-level cross-validation and results are reported in terms of accuracy and F1 score, where the positive label for this binary classification task indicates that the gameplay interaction is related to the reflection.

6.3.2 Results

As indicated by Figure 6.2, the zero-shot model that was only provided with instructions and underwent no fine-tuning had substantially lower performance than the fine-tuned models. The average accuracy of the zero-shot model was 0.668, which was very similar to the accuracy of the majority baseline (accuracy = 0.682). While the few-shot approach was unable to beat the majority baseline in this analysis, it should be noted that prompting can be very important when working with large language models. So, the performance of the zero-shot model could be improved through prompt engineering. For example, providing a more detailed explanation of the classes the label is supposed to identify could be helpful. Alternatively, a few-shot approach could be explored by adding a handful of examples to the prompt that demonstrate how to label

the relationship between a gameplay interaction and a reflection. This has been shown to be a powerful approach for leveraging the power of large language models without needing to spend the resources to fine-tune the parameters of the model (Dong et al., 2022).

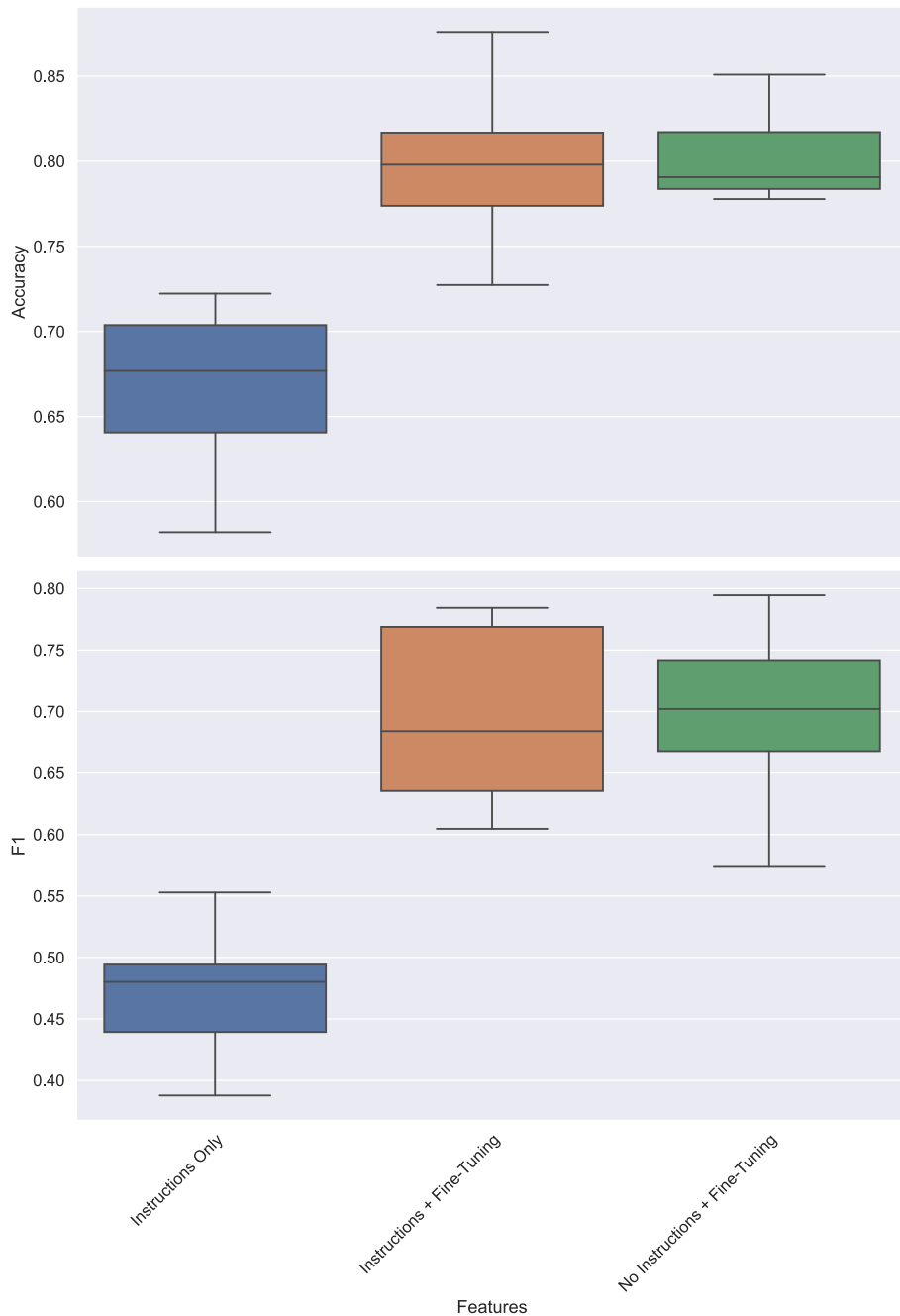


Figure 6.2 Accuracy and F1 scores for automatically linking reflections and gameplay interactions using FLAN-T5.

In comparison, the model that was fine-tuned with instructions achieved an accuracy of 0.800 and the model that was fine-tuned without instructions achieved an accuracy of 0.794. Interestingly, including instructions for this task had almost no effect on model performance. Perhaps this is because the task simply asked the model to determine if two pieces of text are related to one another, which is a fairly straightforward task that can be performed by any language model. Nonetheless, the improved performance of the fine-tuned models compared to the majority baseline and zero-shot approach demonstrates that it is possible to fine-tune large language models to perform this task with a reasonably high level of accuracy. Thus, these models may be able to support adaptive feedback and guidance on the specific gameplay interactions that students decide to reflect on during their interactions with CRYSTAL ISLAND.

CHAPTER 7

AUTOMATED ASSESSMENT OF REFLECTION DEPTH USING DATA AUGMENTATION AND TRANSFORMER-BASED LANGUAGE MODELS

This chapter describes the work done to construct predictive models for automatically assessing the depth of students' reflections during game-based learning. Prior research has explored automated techniques for predicting measures of reflection depth, but they often deal with longer forms of reflective writing produced by post-secondary students, such as reflective essays written as part of an undergraduate class (Gibson et al., 2017; Jung & Wise, 2020). The natural language processing techniques used to analyze students' reflections have included simple count-based representations that capture the words used in a reflection (Ullmann, 2015), rule-based features that capture aspects of language that are theorized to be related to effective reflection (Gibson et al., 2017), and distributed embedding-based representations of language (Kovanović et al., 2018). Recently, Transformer-based language models have been used to assess reflection depth and have demonstrated promising performance (Nehyba & Štefánik, 2022). However, there are still open questions about how to most effectively leverage modern Transformer-based language models for constructing predictive models of reflection depth.

To that end, we evaluated several natural language representations of students' reflections across a range of different machine learning algorithms (i.e., logistic regression, random forest, support vector machine, and gradient boosting trees) to investigate whether Transformer-based language models (i.e., BERT, DistilBERT, and RoBERTa) can outperform baseline natural language representations (i.e., unigrams, tf-idf, GloVe, and ELMo) for automatically assessing reflection depth and identifying components of reflection depth. Then, we investigated whether fine-tuning pre-trained language models on text extracted from educational resources in CRYSTAL ISLAND could improve the predictive performance of these models. Next, we evaluated the performance of these models on predicting several components that are related to the depth of a student's reflection. Finally, we investigated data augmentation techniques for generating a larger amount of training data than the small set of student reflections that we currently have access to. Many machine learning models require a substantial amount of training data to

achieve high predictive performance, but collecting new reflection data from students is expensive. So, automated approaches for generating new data based on our small set of reflections may help improve the predictive performance of these models relatively inexpensively.

Results demonstrate that Transformer-based approaches for automatically assessing reflection depth achieve high predictive performance and that training machine learning models on synthetic reflections generated by large language models significantly improves model performance. These results indicate that the framework presented in this dissertation can be leveraged to provide adaptive support for reflection during game-based learning. Moreover, the performance of these predictive models when trained on synthetic reflections points toward the potential for this framework to readily generalize to new settings without requiring a substantial amount of new reflection data to be used for model training.

7.1 Comparison of Natural Language Representations for Predicting Reflection Depth

Many different ways to represent text have been developed over the years. These range from simple approaches that count the number of occurrences of different words in a piece of text to distributed embedding representations that use neural networks to create a high-dimensional representation of language that can be used to quantitatively measure the similarity of different words. New natural language processing techniques are constantly being introduced, often demonstrating state-of-the-art performance on a wide range on a wide range of tasks. However, it is not always the case that more complex representations of natural language will lead to improved predictive performance, especially on niche tasks in very specific domains (Kovanović et al., 2018). Thus, there is a need to evaluate the performance of several competing natural language representations for the task of reflection depth assessment using our small reflection dataset to determine which approach can achieve the highest performance in this setting.

7.1.1 Method

We evaluated the performance of several different machine learning algorithms for predicting reflection depth scores using different natural language representations of reflection as input. Prior to transforming student reflections into several baseline natural language representations (i.e., binary unigram, tf-idf, GloVe, and ELMo), the text responses were normalized using

tokenization, conversion to lowercase, and removal of non-grammatical characters. When generating binary unigram vectors, tokens that appeared fewer than ten times throughout the corpus were removed. Similarly, any words that were not found in the GloVe embeddings were ignored when calculating average GloVe and ELMo word embeddings, effectively removing misspelled words from the data. For BERT, DistilBERT, and RoBERTa, the reflection text was not transformed prior to being passed to the language models. This is because these Transformer-based language models are robust to character-level perturbations and may actually be able to use character-level information to produce more accurate predictions.

As a baseline, we encoded each natural language reflection as a binary vector representing the unique unigrams that occurred in that reflection (i.e., a one-hot encoding). This was a 220-dimension vector, where each index represents the presence of a specific word in the corpus vocabulary after infrequent words were removed. We also encoded the student reflections as tf-idf vectors, which are sparse real-valued vectors that represent documents based on the frequency of each term in the corpus, weighted by the uniqueness of that term in the corpus. Since tf-idf accounts for the frequency of each word, unlike the binary unigram representation, infrequent words were not removed. Finally, we examined two different word embedding techniques: GloVe (Pennington et al., 2015) and ELMo (Peters et al., 2018). GloVe embeddings are word-based, so it is possible to use pre-trained GloVe embeddings that have been trained on other corpora (i.e., Wikipedia and Gigaword), and simply look up embeddings by word. ELMo, which was also trained on large corpora but uses character-based methods to represent text, is built with the intention that sentences, and not individual words, are used to create embeddings (Peters et al., 2018). To maintain a fair comparison between the various representations of students' written reflections, we first embedded entire written reflection responses with ELMo and then extracted individual word embeddings. This allows the embeddings to capture information related to the specific context in which each word was used. The ELMo word embeddings were 256-dimension real-valued vectors. For GloVe and ELMo, we represented the reflection text as the average embedding across all words in the reflection.

We also evaluated several Transformer-based language models from the BERT family – BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), and RoBERTa (Y. Liu et al., 2019). These models were selected because they represent variations of the original BERT model that may demonstrate different performance on the reflection depth assessment task. All three models

utilize the Transformer architecture and were pre-trained in a self-supervised fashion on a masked language modeling task. This means that the data the models used for training was not manually labeled and instead tokens in the text used for training were randomly masked at a rate of 15% and the model was trained to reconstruct the masked tokens. DistilBERT is a smaller and faster version of BERT that was trained via a distillation process where the original BERT model was used as a teacher and the distilled model was trained to produce the same language probabilities as the original model but with a much smaller neural architecture. RoBERTa is an optimized version of the BERT architecture that was trained on a different dataset and with modified hyperparameters. It has been shown to achieve improved performance compared to the original BERT model. Each of the models from the BERT family produced 768-dimension embeddings. As with GloVe and ELMo, a single embedding representing each reflection was created by taking the average of the individual embeddings for each word in the reflection.

Additionally, we investigated whether fine-tuning the BERT family embeddings on text extracted from CRYSTAL ISLAND's educational resources would improve model performance for evaluating the depth of reflections that students wrote while interacting with CRYSTAL ISLAND. The idea is that, by adapting the language models to the specific language of the target domain, they may be able to perform better on tasks in the domain compared to models that were exclusively pre-trained on general text sources. To perform this fine-tuning, we extracted all of the text from CRYSTAL ISLAND (i.e., book content, the lines spoken by in-game characters during conversations, and poster content) and used it for fine-tuning via masked language modeling. This is identical to the method that was used to pre-train BERT, just at a much smaller scale and in a more specific domain. Since the amount of text found in CRYSTAL ISLAND is still relatively small compared to the amount of text originally used to pre-train BERT, we copied the text twenty times to increase the amount of text used for fine-tuning. We used the same approach as BERT's original fine-tuning, randomly masking 15% of all tokens and training the model to reconstruct those tokens.

Finally, each natural language representation was used to train four different scikit-learn (Pedregosa et al., 2011) classification models (i.e., logistic regression, random forests, gradient boosting, and support vector machines) to predict the average of the reflection depth scores assigned by the raters, rounded down. Reflection assessment models were trained using 5-fold student-level cross-validation at the student level. Across all feature representations, standard

scaling and dimension reduction to 10 dimensions via primary component analysis was performed. This was done to reduce the number of features used as input to the machine learning models, since these classical models tend to perform better with a small set of features as opposed to neural networks which are better able to handle high-dimensional input. Also, this ensured that the models were all receiving the same number of features as input, regardless of the natural language representation being evaluated. Models were trained using 5-fold student-level cross-validation to classify reflections in terms of their assigned depth scores, which ranged from 1 to 5.

7.1.2 Results

Each machine learning model explored in this work was evaluated in terms of accuracy (Table 7.1), average F1 score across all classes (Table 7.2), and quadratic weighted kappa (QWK). Accuracy is important because we want to create machine learning models that can reliably assess students’ reflection depth, so we do not create adaptive systems that are confusing or frustrating for students. However, accuracy alone is insufficient as an evaluation metric since an unsophisticated baseline such as a majority classifier can achieve relatively high accuracy for

Table 7.1 Accuracy scores for reflection depth assessment models.

Feature Representation	Majority Baseline	Logistic Regression	Random Forest	Gradient Boosting	SVM	Average
Binary Unigram	0.651	0.648	0.659	0.627	0.658	0.648
tf-idf	0.651	0.648	0.639	0.622	0.651	0.640
GloVe	0.651	0.644	0.666	0.651	0.675	0.659
ELMo	0.651	0.681	0.677	0.658	0.671	0.672
BERT	0.651	0.701	0.661	0.675	0.653	0.672
BERT-FT	0.651	0.655	0.655	0.660	0.651	0.655
DistilBERT	0.651	0.687	0.666	0.657	0.656	0.667
DistilBERT-FT	0.651	0.693	0.662	0.661	0.653	0.667
RoBERTa	0.651	0.693	0.667	0.637	0.651	0.662
RoBERTa-FT	0.651	0.603	0.653	0.609	0.651	0.629
Average	0.651	0.665	0.661	0.646	0.657	

some datasets. For example, for this dataset, a majority classifier can achieve an accuracy of 0.651 by always predicting that a reflection’s assigned depth score is 2 since that is the most common score. Thus, we also report F1 score, which is the harmonic mean of the precision and recall metrics. Finally, we also evaluate models in terms of quadratic weighted kappa (QWK). This metric is commonly used to evaluate the classification performance of NLP models (Magooda et al., 2022) that are trained to predict classification labels that can be ordered. Since our reflection depth scores go from 1 to 5, with higher values representing deeper reflections, we can evaluate our models in terms of QWK. QWK accounts for the difference between a predicted class and the true class, so a model that makes predictions that are close to the true label but not exactly correct will be evaluated more favorably than a model that predicts an entirely wrong class. The metric produces a score between -1 and 1, with -1 indicating total disagreement, 0 representing random agreement, and 1 representing total agreement. We report QWK for this predictive task because assigning a score of 3 to a reflection that should have been assigned a score of 2 is far better than assigning the same reflection a score of 5. Much of the feedback provided to a student in such a case would be the same, and the potential for causing confusion or frustration due to mismatched feedback would be lower.

Table 7.2 F1 scores for reflection depth assessment models.

Feature Representation	Majority Baseline	Logistic Regression	Random Forest	Gradient Boosting	SVM	Average
Binary Unigram	0.181	0.333	0.305	0.337	0.205	0.295
tf-idf	0.181	0.181	0.204	0.250	0.181	0.204
GloVe	0.181	0.292	0.268	0.297	0.274	0.283
ELMo	0.181	0.452	0.281	0.288	0.256	0.319
BERT	0.181	0.433	0.219	0.320	0.190	0.291
BERT-FT	0.181	0.368	0.243	0.287	0.181	0.270
DistilBERT	0.181	0.458	0.240	0.316	0.202	0.304
DistilBERT-FT	0.181	0.416	0.224	0.281	0.190	0.278
RoBERTa	0.181	0.413	0.277	0.290	0.181	0.290
RoBERTa-FT	0.181	0.311	0.216	0.226	0.181	0.234
Average	0.181	0.366	0.248	0.289	0.204	

We primarily compare model results across different natural language representations and not machine learning algorithms, since the algorithms were chosen simply to represent a range of classical techniques (i.e., linear modeling, an ensemble technique, boosting, and a kernel method). First, we can see that many of the models evaluated in this work did not substantially outperform the majority baseline in terms of accuracy, but did achieve improved performance in terms of F1 score. We note that QWK evaluates to zero when the same class is always predicted, so there is not a meaningful comparison to the QWK score of the majority baseline. These results indicate that, while the trained models are not producing many more correct predictions than can what could be achieved by always assuming that a reflection has a depth of 2 (the majority class), the models are better at capturing the nuances of the data. Looking at average performance across machine learning algorithms, the baseline models tend to achieve similar performance to the BERT models. In fact, ELMo and BERT achieve the same accuracy (0.672) and ELMo achieves the highest F1 score (0.319) and QWK (0.265). However, the BERT language representations may scale better than the baseline techniques with additional training data or may perform better when used as input to more complex machine learning models, such as deep neural networks. Nonetheless, in educational settings it is often a good idea to favor the simplest method that achieves good performance because they are often more explainable than more complex methods and easier to scale into the classroom, where computational resources are severely limited.

Next, looking at the performance of pre-trained versus fine-tuned models from the BERT family, we found that fine-tuning did not tend to improve model performance. In fact, the average F1 score across all machine learning algorithms fell (DistilBERT: F1 = 0.304 \rightarrow 0.278; BERT: F1 = 0.291 \rightarrow 0.270; RoBERTa: F1 = 0.290 \rightarrow 0.234). Fine-tuning may not have been helpful because we were only trying to fine-tune to the domain and not to the task. Rather, fine-tuning the BERT models for the reflection depth assessment task, as opposed to only fine-tuning on a masked language modeling task, would have likely resulted in improved performance. However, we explored this technique because of the lack of reflection data that is available for model training and the desire to investigate whether leveraging additional sources of relevant text data could improve model performance. It may be beneficial to investigate a combined fine-tuning approach, where models are first fine-tuned to the language of the domain (i.e., fine-tuned via

Table 7.3 Quadratic weighted kappa scores for reflection depth assessment models.

Feature Representation	Majority Baseline	Logistic Regression	Random Forest	Gradient Boosting	SVM	Average
Binary Unigram	0.000	0.357	0.248	0.333	0.103	0.260
tf-idf	0.000	0.010	0.036	0.183	0.000	0.057
GloVe	0.000	0.084	0.119	0.180	0.107	0.123
ELMo	0.000	0.508	0.167	0.272	0.112	0.265
BERT	0.000	0.484	0.066	0.238	0.009	0.199
BERT-FT	0.000	0.377	0.186	0.261	0.000	0.206
DistilBERT	0.000	0.445	0.072	0.188	0.015	0.180
DistilBERT-FT	0.000	0.405	0.048	0.172	0.002	0.157
RoBERTa	0.000	0.398	0.182	0.289	0.000	0.218
RoBERTa-FT	0.000	0.303	0.035	0.149	0.000	0.122
Average	0.000	0.337	0.116	0.226	0.035	

masked language modeling on the text extracted from CRYSTAL ISLAND educational resources) and then fine-tuned to the task of reflection depth assessment.

7.2 Predicting Components of Reflection Depth

To enable the creation of systems that adaptively provide support for reflection during game-based learning in CRYSTAL ISLAND, more fine-grained evaluations of students' reflections could allow for more actionable feedback. The reflection depth rubric introduced in Chapter 5.1 can be used to provide a measure of the quality of a student's reflection, but feedback based on the depth score alone has limited ability to prescribe ways to improve a reflection. The levels of the rubric describe several components that should be included in effective reflections, such as an evaluation of what the student previously did in the game as well as what they plan to do moving forward. By training machine learning models to individually detect the presence of important components of reflection depth, we can provide fine-grained information that can be presented to students alongside a numerical evaluation of their reflection depth as guidance for how to improve their reflection.

7.2.1 Method

Five components of reflection depth were identified based on the established reflection depth rubric: Relevant (the content of the reflection was related to the learning scenario), Backward (the presence of a backward-looking component), Forward (the presence of a forward-looking component), Hypothesis (the presence of a hypothesis related to the science problem solving scenario), and Processing (evidence of information processing rather than simply restating information, such as synthesizing information from multiple sources or evaluating the importance of some information). Each component was considered to be binary (i.e., it either was or was not present in each reflection) and a rubric describing the criteria for identifying each component was created (Table 7.4).

Using the rubric, two researchers labeled a shared twenty percent of the reflections to establish inter-rater reliability. We targeted a Cohen's kappa of 0.8 for each component, which would indicate substantial agreement between the annotators. For the components Forward ($K=0.878$) and Hypothesis ($K=0.821$), acceptable agreement was achieved in the first round of annotation. For each of the other components, the rubric items needed to be revised. For the Relevant component, the rubric item needed to be revised so that vague reflections (e.g., "I'm solving the mystery") were considered not to be relevant. Labeling a new subset of the data with this revision resulted in substantial agreement ($K=0.920$). For Backward, the raters agreed that reflections that implicitly referenced past events by stating information that they had learned (e.g., "pathogens can transfer from people to another") should not be considered to have a backward-looking component. Upon making this change to the rubric item, substantial agreement was achieved ($K=0.938$). Finally, for Processing, the annotators initially disagreed about whether simply including the phrase "the most important thing" in a reflection counted as evaluating the subsequent information as important for solving the mystery. This was a point of confusion because the reflection prompt directly asks students to state what the most important thing they learned was, so it appeared that many students blindly copied this language when writing their reflections. The annotators agreed to give students the benefit of the doubt that they actually did evaluate the knowledge in their reflection as important, but only when the information presented was specific. For example, "the most important thing is figuring out the symptoms" was considered relevant while "the most important thing is finding clues" was

considered too vague to demonstrate information processing. Labeling a new subset of the data with this revised understanding of the rubric item led to substantial agreement ($K=0.836$).

Separate binary classifiers were trained to label each of the reflection depth components. The same set of natural language feature representations used in Chapter 7.1 (i.e., binary

Table 7.4 Rubric for components of reflection depth.

Category	Description	Labels	Explanation
Relevant	Does the reflection address information that is relevant to this problem or learning scenario?	Yes	The reflection mentions something that is relevant to the science content or the mystery.
		No	The reflection is unintelligible or does not address anything related to the game's learning content, even if it mentions the game's mechanics.
Backward	Does the reflection contain a backward-looking component?	Yes	There is a backward-looking component where students discuss something that they did or learned in the past, even if it's unrelated to the game.
		No	There is no backward-looking component.
Forward	Does the reflection contain a forward-looking component?	Yes	There is a forward-looking component that presents an action or set of actions the student wants to take or at least includes vague information about what the student plans to do.
		No	There is no forward-looking component.
Processing	Does the reflection process information by explicitly evaluating the importance of what the student has learned or synthesizing information from multiple sources to make inferences? For this label, students have to be explicit with their information processing. Even though we may be able to imply that they included certain information because they thought it was important, they need to explicitly state that.	Yes	The reflection processes information by explicitly evaluating the importance of information that they have learned or the importance of certain actions they have taken OR by synthesizing information from multiple sources to make an inference, make comparisons, or draw conclusions.
		No	There is no processing of information. Importantly, the presence of information from multiple sources does not necessarily indicate processing.
Hypothesis	Does the reflection present a hypothesis about the mystery (e.g., the source, illness, or contaminant)?	Yes	Even if not in a formalized structure, the reflection presents some sort of hypothesis about some component of the mystery (e.g., the source, illness, or contaminant).
		No	There is no hypothesis or a hypothesis about something irrelevant to solving the mystery.

unigram, tf-idf, GloVe, ELMo, BERT, DistilBERT, and RoBERTa) were used to train the same set of classification algorithms (i.e., logistic regression, random forest, support vector machines, and gradient boosting trees). In addition, FLAN-T5 was introduced. In this analysis, FLAN-T5 was fine-tuned using a typical supervised learning approach, where data from the training set was used to update model parameters and then the fine-tuned model was evaluated on the test set. All models were trained using 10-fold student-level cross-validation and results are reported in terms of accuracy and F1 score.

7.2.2 Results

Many of the labels for depth components had very skewed distributions, with 87% of reflections being relevant to the learning scenario, 83.3% including a backward-looking component, 82% not demonstrating evidence of information processing, and 88% not including a hypothesis related to the mystery. Thus, there was little room for improvement over the majority baseline. Nonetheless, the models did largely outperform the majority baseline. Across all predictive tasks, FLAN-T5 achieve the highest performance in terms of accuracy ($\text{accuracy}_{\text{avg}} = 0.920$). FLAN-T5 also achieved a higher F1 score than all other models for predicting Relevant, Backward, and Forward, but performed worse for predicting Processing and Hypothesis. In terms of average F1

Table 7.5 Accuracy for reflection depth component detection models.

Feature Representation	Relevant	Backward	Forward	Proc.	Hyp.	Average
Maj. Baseline	0.870	0.833	0.591	0.820	0.880	0.799
Binary Unigram	0.868	0.855	0.857	0.819	0.882	0.857
tf-idf	0.873	0.842	0.787	0.815	0.884	0.840
GloVe	0.921	0.868	0.788	0.810	0.880	0.853
ELMo	0.931	0.870	0.781	0.823	0.886	0.858
BERT	0.903	0.844	0.762	0.826	0.881	0.843
DistilBERT	0.909	0.840	0.706	0.811	0.880	0.829
RoBERTa	0.869	0.832	0.670	0.808	0.877	0.811
FLAN-T5	0.968	0.898	0.951	0.846	0.938	0.920
Average	0.905	0.856	0.788	0.820	0.888	

Table 7.6 F1 score for reflection depth component detection models.

Feature Representation	Relevant	Backward	Forward	Proc.	Hyp.	Average
Maj. Baseline	0.930	0.909	0.743	0.901	0.936	0.884
Binary Unigram	0.928	0.917	0.854	0.837	0.849	0.877
tf-idf	0.931	0.911	0.777	0.820	0.844	0.856
GloVe	0.955	0.923	0.779	0.817	0.854	0.866
ELMo	0.960	0.925	0.778	0.840	0.869	0.874
BERT	0.945	0.911	0.757	0.851	0.854	0.864
DistilBERT	0.948	0.909	0.691	0.821	0.841	0.842
RoBERTa	0.927	0.904	0.659	0.823	0.852	0.833
FLAN-T5	0.981	0.940	0.938	0.463	0.677	0.800
Average	0.947	0.917	0.779	0.784	0.830	

score across all tasks, none of the models were able to beat the majority baseline and the top performing model was trained on the binary unigram representation ($F1_{avg} = 0.877$).

These results demonstrate the benefits of using FLAN-T5, which has achieved state-of-the-art performance on several NLP tasks (Chung et al., 2022), for automatically detecting different components related to the depth of students’ reflections. Leveraging these models, we can reliably provide fine-grained feedback to students during game-based learning based on whether their reflections are relevant to the learning scenario, whether they have included a backward-looking component that indicates what they have previously done, and whether they have included a forward-looking component that describes what they will do moving forward. Feedback on these components can help students write effective reflections that bridge the gap between their past and future learning activities, which is a major goal of reflection. However, the more complex components of reflection depth (i.e., Processing and Hypothesis) presented a challenge for these models. In particular, being able to provide feedback on whether or not a student has demonstrated information processing in their reflection would likely be very beneficial for supporting effective reflection, since this could help students move away from reflections that simply restate information to those that synthesize information from multiple

sources or evaluate what they have learned relative to their goals. This is one of the most important components of reflection, since it is through this processing that students can make changes to their learning strategies and goals.

The evaluated models do not demonstrate adequate performance for detecting Processing and Hypothesis, so there are opportunities for additional research to investigate ways for improving these models. One possible direction for improving model performance on these tasks would be to explore multi-task learning. In multi-task learning, a single shared representation is learned for all of the tasks, and since these tasks are likely related to one another, this shared representation could help improve predictive performance across all depth components.

7.3 Improving Automated Reflection Depth Assessment Models using Data Augmentation Techniques

As demonstrated in Chapter 7.1, machine learning models trained to automatically assess the depth of students' reflections were able to outperform a majority baseline in terms of F1 score and quadratic weighted kappa, but model performance overall was not very impressive. However, we expect that model performance could be improved with access to more training data. The typical option for collecting additional training data would be to run additional classroom studies with students, but it is very expensive to collect additional reflection data from students. Collecting additional data would involve planning studies with teachers, conducting the study with several classes of students, cleaning the data to account for missing data, and data loss due to low consent/assent rates. To collect data at a scale that even approaches what is typical in a modern natural language processing setting presents a significant challenge.

An alternative approach to acquiring more training data is through the use of AI-based methods for augmenting the existing reflection dataset. With data augmentation techniques, our small reflection dataset may be used as the core of a new, larger dataset that can be artificially expanded to a significant degree. While there may be a limit to the benefits of generating additional data from a small set of genuine data, there is significant potential to improve the performance of automated reflection depth assessment models by training on additional data generated by perturbing existing reflections or synthesizing entirely new reflections. In this chapter, we investigate several techniques for generating synthetic reflection data and

demonstrate that the performance of Transformer-based machine learning models trained on this data significantly outperforms models trained exclusively on genuine student reflections.

7.3.1 Baseline Data Augmentation Techniques

We explored several baseline data augmentation techniques: character-level noise injection, masked language modeling, and back-translation. Noise injection involves adding, removing, or modifying some of the characters in a string to produce a slightly perturbed version of the string (e.g., “The egg has bacteria” → “Th egx has bactrtia”). The ratio at which characters were perturbed in this work is 15%, which allows for some variance while maintaining a low likelihood of significantly changing the meaning of the string. Masked language modeling is similar to noise injection but on a word level. With this approach, a certain percentage of the words in a string (again, 15% in this work) are removed and replaced with a mask token. Then they are passed to a language model that was trained on a masked language modeling task, in this case the base BERT model, and the model is asked to replace the mask tokens with the most likely words at each location. This can produce strings that are more different than the strings created via noise injection (e.g., “The egg has bacteria” → “The egg has viruses”), but once again we assume that a replacement rate of 15% is low enough that the meaning of each reflection is unlikely to be substantially altered by this transformation. Back-translation transforms strings by using machine translation models to translate a string from English to another and then back to English. This can result in significant changes to the string (e.g., “The egg has bacteria” → “There is bacteria in the egg”) but preserves most of the meaning. These techniques modify existing data points (i.e., reflections) to create additional training data with roughly the same characteristics as the original data. We use the same labels for each new reflection since we expect that these baseline transformations do not change the content of the reflection enough to change their labels.

Additionally, we used oversampling as a naïve data augmentation baseline, where each reflection was copied multiple times to create additional training data. This baseline was included to evaluate whether simply having more training data improved model performance or if transforming the reflections actually had an impact.

7.3.2 Reflection Synthesis using Large Language Models

We also explored the benefits of generating entirely new reflections using large language models (LLMs), which can potentially allow us to augment our dataset with a diverse set of entirely new reflections. For this work, we used GPT-3.5 (OpenAI, 2022) to synthesize new reflections. Synthesis of new reflections via GPT-3.5 required a sophisticated prompt to “explain” the reflection-writing task to the LLM. The prompt template can be seen in Figures 7.1 and 7.2. First, the system is presented with the overall task and provided with background information on the learning setting and the specific definition of reflection that we use in this work. This includes a list of the actions students can take in CRYSTAL ISLAND, since without this the system would have an unconstrained view of the types of actions that it could generate as potential future steps to include in a reflection.

```
Please generate possible reflections that a middle school student might
create based on their experiences in a game-based learning environment for
microbiology.
```

```
Here is my definition of reflection: “A reflection is deliberate
contemplation of one’s own thoughts and actions with the intent to assess or
evaluate one’s knowledge or learning processes for the purpose of driving
adaptation in the future.”
```

```
The setting for the learning environment is a remote island called Crystal
Island where some members of a research team have recently started to become
sick. The player must determine what disease is spreading amongst the
researchers, what food item it is spreading though, and what the correct
treatment or prevention plan is based on the disease that’s spreading.
```

```
Players can take the following actions to solve the mystery:
```

- ```
1) speak to virtual characters to learn about their symptoms , the mystery,
or microbiology knowledge
2) read informational texts about types of diseases or microbiology
knowledge
3) collect food items in the world
4) scan food items to see if they contain bacteria , viruses, or carcinogens
5) view posters to learn about the symptoms and treatments of various
diseases
```

**Figure 7.1** Background information about CRYSTAL ISLAND that was included in the reflection synthesis prompt.

In addition to the general information, we provided specific examples of past students' reflections and the depth score that those reflections had received. This is an example of in-context learning (Dong et al., 2022), where previous examples are used to condition the model to produce a particular output without actually modifying the model's parameters via fine-tuning. To ensure that there was no data leakage and that the system was not learning to generate synthetic reflections based on the test data, only reflections from the training dataset were sampled to use as examples of genuine reflections. Next, a sequence of gameplay interactions that had been converted into natural language, using the same approach as described in Chapter 5.3, was presented and the system was notified that it would be generating a reflection based on these events. Similarly, gameplay interaction sequences were sampled from a separate CRYSTAL ISLAND dataset to make sure that the synthetic reflections would not be informed by any data that the machine learning system being evaluated should not have access to.

Then, specific criteria for the reflections that we wanted the system to generate were described. These included the minimum and maximum number of actions that the system should draw from when writing the reflection, a reminder that the reflection should use language that was similar to what a middle school student would use, and the minimum and maximum desired length of the reflections, in sentences. This length requirement was important, as it helped the system generate short reflections as opposed to the very long ones that it generated when it was not provided with a desired reflection length.

Finally, the system was instructed to generate five reflections based on the provided context – one for each depth score – and to produce its output in a JSON format to allow for easy data extraction. Since initial explorations indicated that the system had a tendency to produce only high-quality reflections, which would not be representative of the types of reflections that real students produce, it was reminded one last time that the reflections would be used as both good and bad examples to help students learn how to reflect and that the reflections should seem like they were generated by middle school students, which means that they may include misspellings.

We are assessing reflection depth on a scale from 1 to 5 using this rubric:

- 1: The text is not reflective at all. It is brief and does not provide any details about what the student did or learned. It is either entirely off task or vaguely mentions a random action without reflecting on it.
- 2: Presents a vague hypothesis or plan, often directly restating information that was presented in the game AND the student demonstrates awareness of their own knowledge and goals but does not show that they are evaluating their knowledge to inform their future actions .
- 3: Presents a clear hypothesis or plan without any reasoning, demonstrating that the student has evaluated their knowledge and made connections to their goals . However, they have not articulated the reasoning behind the importance of this knowledge or its benefit toward achieving their goals .
- 4: Presents a clear hypothesis or plan with reasoning, demonstrating that the student has evaluated their knowledge and made connections to their goals . However, they have only provided reasoning for the importance of this knowledge or its benefit toward achieving their goals , not both.
- 5: Presents both a clear hypothesis and plan with reasoning, demonstrating that the student has evaluated what they have learned and made connections to their goals . Furthermore, they have provided reasoning for the importance of this knowledge , and they have indicated how it will help them achieve their goals .

Here are some examples of players' reflections and the depth scores they were given. IMPORTANT: Only use these reflections as an example, but don't use the information from these reflections when writing new reflections. Only use the information from the gameplay sequence provided below.

```
{% for reflection in list_of_sample_reflections -%}
- "{{ reflection.text }}" Score: {{ reflection.depth }}
{% endfor %}
```

Here is a gameplay sequence that you should draw on when creating a reflection :

```
{% for action in list_of_gameplay_actions -%}
{{ loop.index }}) {{ action }}
{% endfor %}
```

The reflections you generate should meet the following criteria :

- It should include between {{ min\_actions }} and {{ max\_actions }} of the actions the student took. Ignore all other actions.
- It should use language that makes it sound like a middle schooler wrote it .
- It should be between {{ min\_sentences }} and {{ max\_sentences }} sentences long.

Please generate 5 reflections in total - one for each of the 5 reflection depth scores in the rubric (i.e., 1, 2, 3, 4, and 5). Don't just use the same actions for every reflection. Instead, try to use different actions for each reflection.

Write them in JSON format like this:

```
[{
 "text": REFLECTION TEXT GOES HERE,
 "score": REFLECTION DEPTH SCORE GOES HERE
}]
```

Keep in mind that the reflections do not have to be good - a low score indicates that a reflection needs improvement, which is fine since we'll be using these reflections as both good and bad examples to help students learn how to reflect. In fact, a reflection with a score of 1 should be a BAD reflection. It should be minimal and maybe not even related to the learning environment . Reflections should also contain a few misspellings for challenging words , since they're supposed to seem like they were written by middle schoolers .

**Figure 7.2** Prompt instructions for synthesizing new reflections based on the reflection depth rubric, examples of previous students' reflections, and a sample gameplay interaction sequence.

### 7.3.3 Exploring Synthetic Reflections

Overall, the synthetic reflections generated by GPT-3.5 were very well-formed and highly related to the reflection task that was presented in the prompt. However, they demonstrate some idiosyncrasies that distinguish them from genuine student reflections. First, they tend to include more detail in their reflections than real students do. For example, a synthetic reflection might reference a conversation with “Quentin, the camp cook” while real students would likely refer to the same character as simply “the cook”. Another example is that synthetic reflections often used the “behind-the-scenes” names for food items, such as “ChickenLeg” because that is the information that was extracted from the game logs and passed to the system. These discrepancies could easily be addressed by modifying the information that is provided to the LLM. Another idiosyncrasy is that the synthetic reflections often included plans to make real life changes based on what was learned in the game, such as making sure to always wash their hands before eating or seeking out vaccinations as a means of staying healthy. The reflection prompt never explicitly asks students to focus their reflections only on what they intend to do in CRYSTAL ISLAND, but that much was understood by the students. However, the LLM did not have the same frame of reference when writing reflections, so it often looked beyond the constraints of the learning environment and thought about the problem being solved with a wider, more expansive view.

There were many things that the LLM did that made the synthetic reflections seem convincingly like the real reflections written by students. For example, the system occasionally generated reflections with misspellings, such as “im not sure wha I wlearned from this. it was kind of boring”. This was a very common feature of the genuine reflections, since new terms like “carcinogens” and “salmonellosis” are challenging for middle school students. Also, many of the reflections demonstrated confusion, frustration, or boredom (e.g., “I don't get this game at all. Why do I have to learn about bacteria and viruses? It's boring.”), which was common to see among the genuine student reflections. Finally, the level of sophistication of the language used in the synthetic reflections was generally consistent with what real students produced, such as the use of simple terms like “germs” as opposed to “bacteria” when describing a contaminant that was found on a food item (e.g., “the milk has germs but idk if it's bad bc not all germs are bad and robert told me there are 25 different germs in my mouth”).

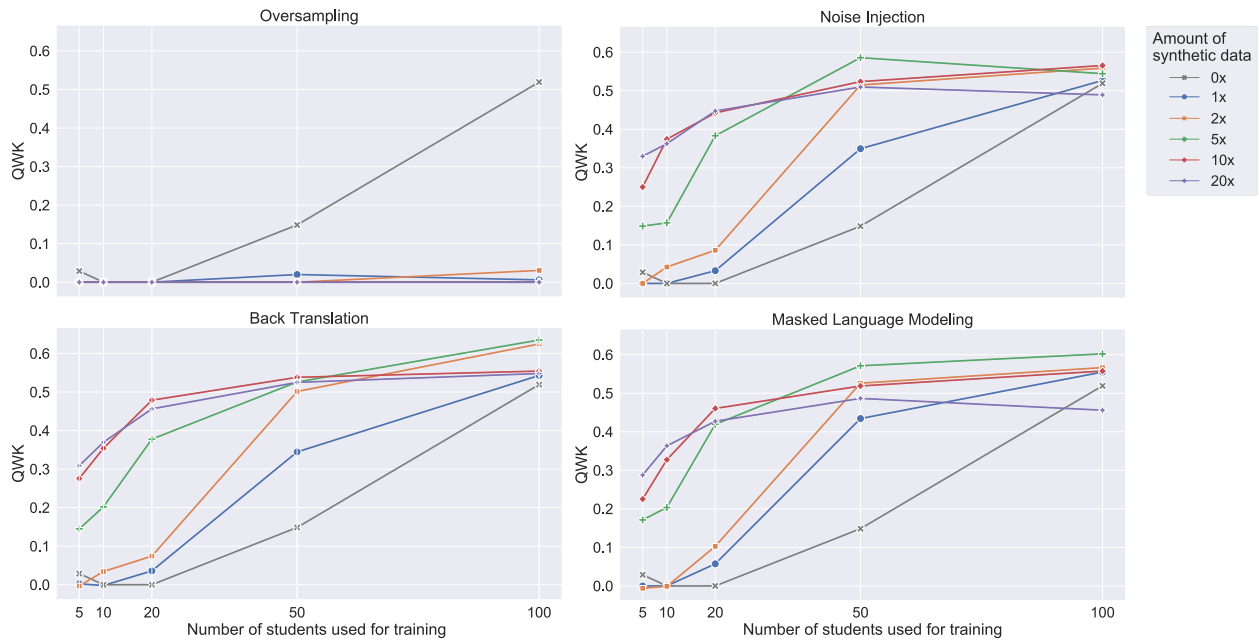
### 7.3.4 Predictive Modeling with Augmented Reflection Data

First, we evaluated the performance of each language model when trained on all of the available reflection data that was collected from students. In this experiment, each model was trained using 5-fold student-level cross-validation to classify the depth of a given reflection from 1 to 5. Just like the analyses presented in Chapter 6.3, we fine-tuned the FLAN-T5 base model (250M trainable parameters) using LoRA parameter-efficient fine-tuning. The BERT-family models were also fine-tuned using LoRA. Average accuracy, F1, and quadratic weighted kappa (QWK) results across cross-validation folds are shown in Table 7.. FLAN-T5 outperformed each BERT-based model across all metrics. All remaining experiments were run with both BERT and FLAN-T5, but we report only FLAN-T5 results to save space since FLAN-T5 continued to demonstrate higher performance across all experiments.

Next, we designed an experimental setup to evaluate the impact of synthetic data on model performance across several settings representing different levels of data scarcity. We performed 5-fold student-level cross-validation with our set of genuine reflections. Within each fold, a group of around 30 students was held out as the test set and was used to evaluate the performance of the trained models. The data that was included in the training set for each fold was further divided to evaluate model performance when the model had access to different amounts of training data – 5, 10, 20, 50, or 100 students. For each training data configuration, we investigated different multiples of synthetic data relative to the genuine data – 0x, 1x, 2x, 5x, 10x, and 20x. That is, in the 0x setting there was no additional synthetic data and in the 20x setting there was 20 times as much synthetic data as there was genuine data. This allowed us to investigate whether there was a baseline amount of genuine data that resulted in the best model

**Table 7.7** Depth assessment results for FLAN-T5 and BERT-family models.

| Language Model | Accuracy     | F1           | QWK          |
|----------------|--------------|--------------|--------------|
| BERT           | 0.658        | 0.299        | 0.210        |
| DistilBERT     | 0.671        | 0.304        | 0.197        |
| RoBERTa        | 0.656        | 0.301        | 0.219        |
| FLAN-T5        | <b>0.701</b> | <b>0.416</b> | <b>0.519</b> |



**Figure 7.3** QWK for baseline data augmentation techniques across different amounts of synthetic data.

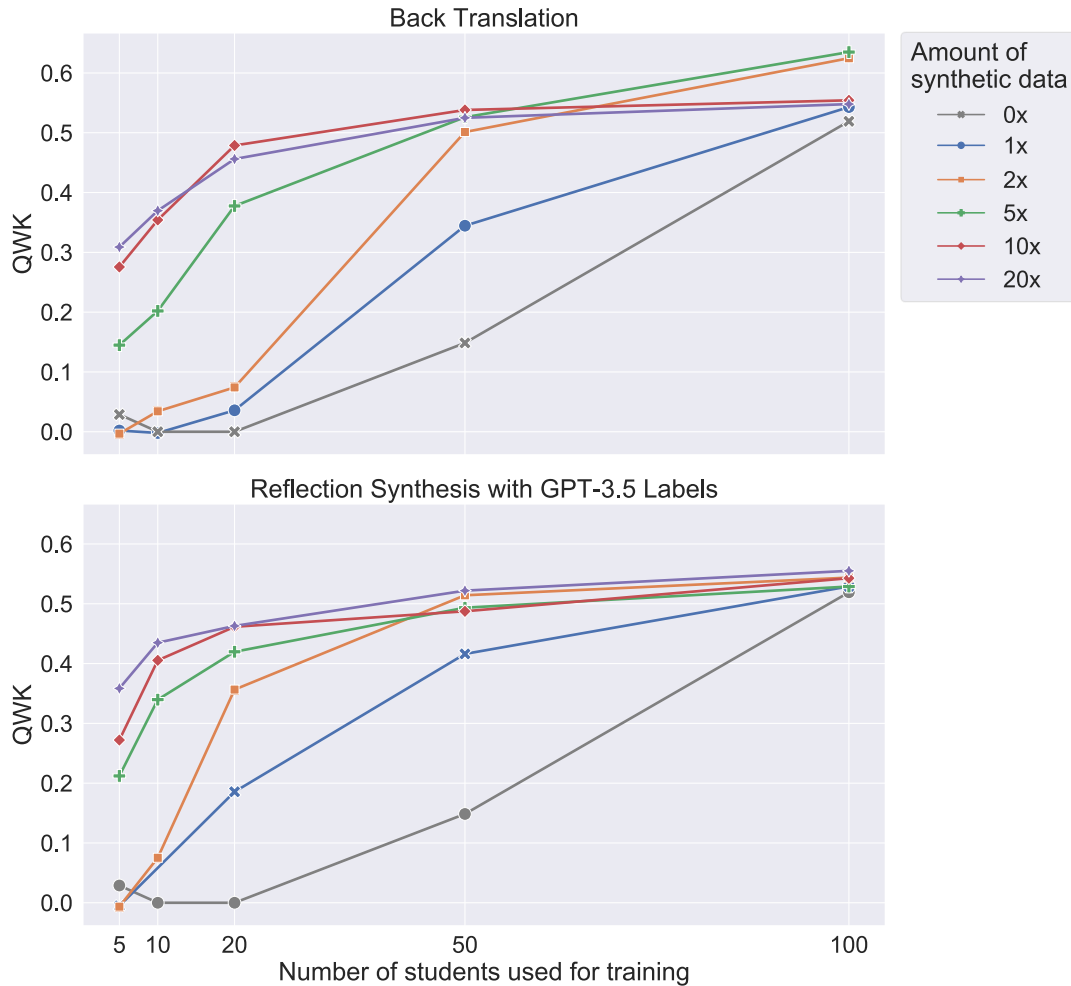
performance or if there was always the same benefit for introducing additional synthetic training data. Also, this allowed us to determine whether any performance change due to additional synthetic training data would scale consistently or if there was a saturation point at which additional synthetic data was no longer useful. Models were evaluated in terms of accuracy, F1 score, and quadratic weighted kappa for this 5-class classification task. To save space, we focus on QWK as the primary evaluation metric since it captures how close each prediction was to being correct rather than just evaluating whether or not it was exactly correct.

We begin by evaluating the performance of each baseline data augmentation technique as a function of the amount of real and synthetic data available (Figure 7.). Oversampling was ineffective, as the models consistently overfit to the data and only ever predicted reflection depth scores of 2. As a result, the QWK for those models was very near to zero across all data configurations. Noise injection, back translation, and masked language modeling all outperformed models that were trained exclusively on genuine student reflections. Back translation achieved the highest overall performance with a QWK of 0.635 when trained on reflections from 100 students augmented with 5x synthetic data. This was a significant improvement over the model trained on 100 students with no synthetic data, which achieved a QWK of 0.519. Since back translation achieved the highest performance across these data

augmentation techniques, it was used as the baseline data augmentation technique when compared to models trained on reflections synthesized by GPT-3.5 in the next experiment.

Results in Figure 7. demonstrate that back translation also outperforms models that were trained with entirely new reflections generated by GPT-3.5. Again, a QWK of 0.635 when the model was trained on reflections from 100 students augmented with 5x synthetic data was significantly higher than the highest performance achieved by training FLAN-T5 on synthesized reflections, which achieved a QWK of 0.555 when trained on reflections from 100 students augmented with 20x synthetic data. However, models trained with reflections synthesized by GPT-3.5 did outperform models trained with data augmented via back translation when data was very scarce (i.e., there were only 5 or 10 students worth of genuine reflection data). The top-performing model trained with 10 students and 20x data generated with back translation achieved a QWK of 0.370 while the model trained with 10 students and 20x GPT-3.5-generated reflections achieved a QWK of 0.435. These results may be due to the fact that the synthesized reflections provide more diversity in the training set, which can be helpful when the training set is relatively small. However, while inspecting the synthetic reflections, we noticed that the labels that GPT-3.5 assigned to the reflections it generated were often not correct. For instance, a reflection would very clearly restate information that was learned in the game (matching the criteria for a depth score of 2), but the assigned label would be 4 or 5. So, as the amount of data available increased, it is possible that the less reliable labels from the reflections generated by GPT-3.5 may have decreased performance compared to the labels associated with reflections generated via back translation, which were likely more accurate since back translation does not significantly alter the meaning of the reflection.

These results demonstrate several benefits of training reflection depth assessment models on an augmented reflection dataset. Using back translation, models not only achieved higher maximum performance, but also achieved performance on par with models trained on a larger set of genuine student reflections. That is, models trained with only 50 students worth of genuine data and 10x synthetic data (QWK = X) outperformed models trained with 100 students worth of genuine data (QWK = X). In fact, models trained with only 20 students worth of genuine data and 10x synthetic data (QWK = X) achieved comparable performance. This shows that data augmentation via back translation can be used to substantially bolster model performance in settings with small amounts of reflections collected from students. Thus, there is significant



**Figure 7.4** QWK for FLAN-T5 models trained with reflections generated via back translation and synthetic reflections created by GPT-3.5.

potential to leverage this approach to deploy AI-enabled adaptive support for reflection in new settings without the need to spend time collecting a large set of training data that is specific to a new reflection prompt or learning environment.

### 7.3.5 Improving Predictive Models with Manually Labeled Synthetic Reflections

Since we noticed that several of the labels assigned to the reflections generated by GPT-3.5 were not accurate, we investigated whether having correct labels would improve model performance. To this end, two researchers manually labeled the synthetic reflections using the established reflection depth rubric. First, 10 students worth of reflections were held out and used exclusively as examples provided to GPT-3.5 to demonstrate what students' reflections looked like. This allowed us to only label two thousand reflections while still running experiments for most of the

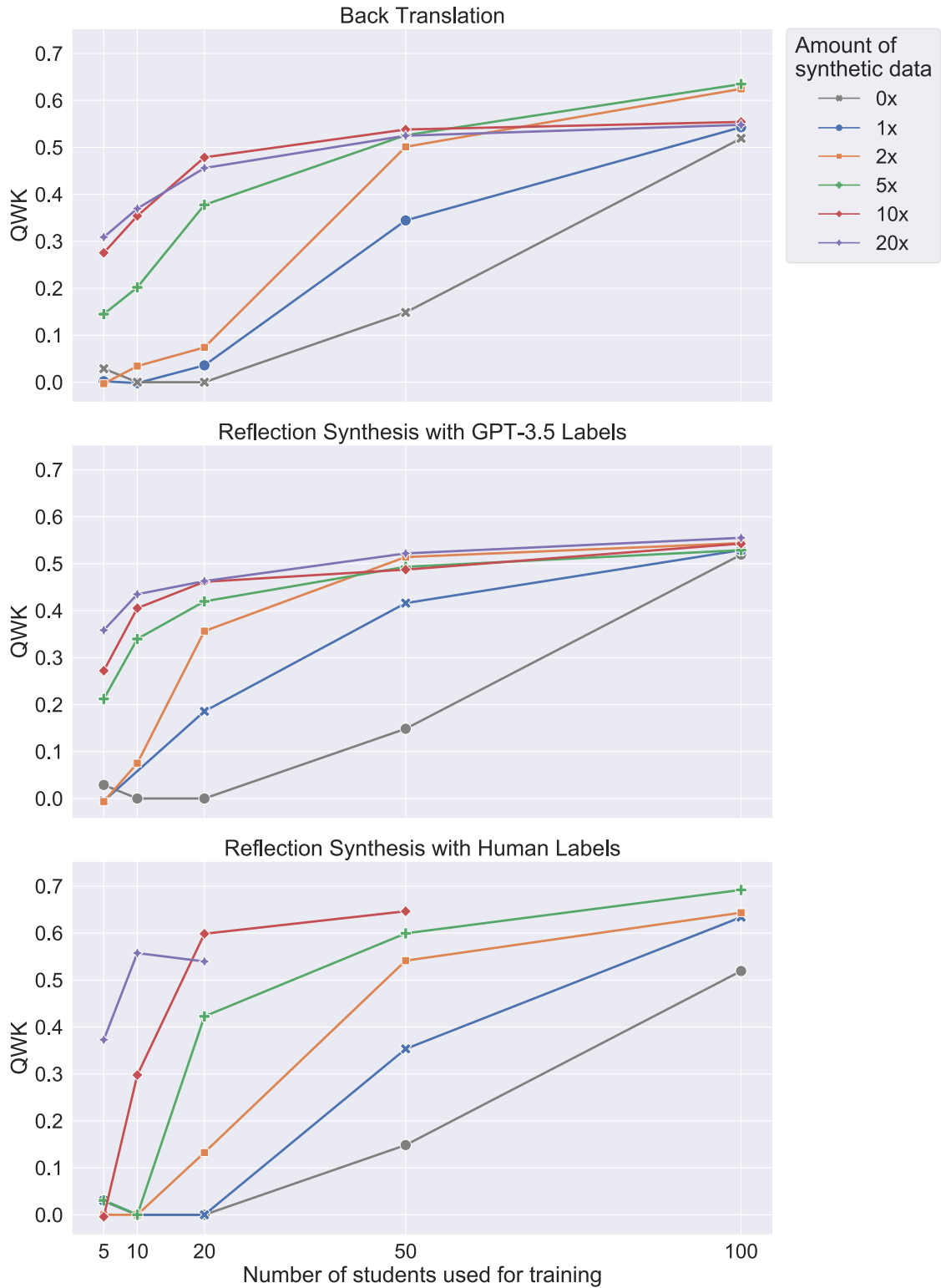
data configurations. We started by labeling 250 synthetic reflections and calculated inter-rater reliability between the two researchers. We achieved an intraclass correlation of 0.866, which indicates substantial agreement. Comparing the manual labels to the labels provided by GPT-3.5 yielded an intraclass correlation of 0.365, indicating poor agreement. While this manual labeling approach still takes a lot of effort and requires the time of an expert annotator, it is an ideal tradeoff in exchange for additional training data. The cost to collect genuine reflection data from students would be massive in comparison, with the need to recruit participants for a new study, facilitate a multi-day study, and receive consent and assent from participants to use their data.

Using the set of two thousand synthetic reflections with manually assigned depth scores, FLAN-T5 models were trained using 5-fold student-level cross-validation. Since there were only two thousand synthetic reflections, the training data for some configurations (e.g., 50 students with 5x synthetic data) consisted of 50 students worth of genuine reflection data plus the same 2000 synthetic reflections in each split. This is because there were roughly four genuine reflections per student, so there were not enough synthetic reflections to use a different subset for each split. However, even though the synthetic reflections used in each split was not always unique, each split still evaluated the model on a different test set. The limited set of synthetic reflections allowed us to run all experiments that evaluated models with 1x, 2x, and 5x synthetic data, but some of the experiments requiring 10x and 20x synthetic data had to be discarded.

Results of these experiments are presented in Figure 7.. First, we note that models trained on manually labeled synthetic reflections outperformed models trained on synthetic reflections using depth scores assigned by GPT-3.5, except when there was very little training data available (i.e., 20 or fewer students worth of genuine data and less than 5x synthetic data). The top-performing data configuration was 100 students with 5x synthetic reflections that were manually assigned new depth scores (QWK = 0.692), which outperformed both the model trained with 100 students worth of genuine reflections and no synthetic data (QWK = 0.519) and the top-performing model trained with reflections generated via back translation (QWK = 0.635). Moreover, models trained with only 10 students worth of data and 20x synthetic data (QWK = 0.599) significantly outperformed the top model with only genuine reflection data (QWK = 0.519). Finally, models trained with only 50 students worth of data and 10x synthetic data (QWK = 0.646) outperformed models trained with 100 students worth of data and 5x synthetic data generated via backtranslation (QWK = 0.635). These results demonstrate that manually assigning depth

scores to GPT-3.5-generated synthetic reflections can lead to substantial improvements in model performance. By generating entirely new reflections, models can learn from a wider range of more diverse data compared to the small set of genuine reflections that we had access to.

The results presented in this work demonstrate that improvements due to additional synthetic data begin to diminish after 5x or 10x synthetic data is introduced, but results still show some performance improvements with 20x synthetic data. Thus, additional testing with even greater amounts of synthetic data could be investigated to identify whether there is a clear point where adding synthetic data will no longer improve model performance. Additionally, it would be interesting to evaluate models that are trained exclusively on synthetic data. For instance, example reflections could be generated for a new learning environment and they could be used to generate synthetic reflections. Then, these fully synthetic reflections could be used to train models to predict the score of reflections students generate in the new learning environment to evaluate whether a fully synthetic training set can be useful for predicting reflection depth. If so, such a workflow would allow this reflection analysis framework to be readily used in any environment that wishes to support student reflection, since there would no longer be a requirement to collect reflection data in the new environment before deploying AI-augmented adaptive reflection support.



**Figure 7.5** QWK for FLAN-T5 models trained with reflections generated via back translation, synthetic reflections created by GPT-3.5, and manually relabeled synthetic reflections.

## CHAPTER 8

### CONCLUSION

Self-regulated learning skills are critical for 21<sup>st</sup> century students, since the increase in student autonomy enabled by digital learning technologies requires students to monitor and manage their own learning processes to be successful. However, despite the importance of SRL skills, they are often not explicitly taught to students. As a result, many students do not employ them effectively, and thus struggle to get the maximum benefit from complex learning environments such as game-based learning. Thus, there is a clear need for systems and tools that can help students develop and deploy self-regulated learning skills. With recent advances in artificial intelligence, and especially natural language processing, there is significant potential to build systems to automatically analyze students' self-regulated learning processes as they interact with a digital learning environment and deliver real-time adaptive support for these skills. With a particular focus on the self-regulated learning skill of reflection, this dissertation has investigated the use of natural language processing techniques for automatically assessing students' reflection. Transformer-based approaches for automatically assessing reflection depth during game-based learning have provided automated insights into students' reflection behaviors. Taken together, the work represents a unified framework that enables contextualized feedback for reflection during game-based learning.

#### 8.1 Hypotheses Revisited

The work presented in this dissertation has evaluated the following thesis statement:

*A natural language processing framework leveraging the Transformer architecture can be used to automatically analyze the depth and content of students' written reflections during game-based learning to provide insights into students' reflection behaviors and enable adaptive intervention.*

The following hypotheses have been tested to evaluate the performance of the framework with respect to the ability to automatically analyze students' reflections. Machine learning models

have been evaluated in terms of accuracy, F1 score, and quadratic weighted kappa via 5- or 10-fold student-level cross-validation.

H.1. Reflection depth assessment models that incorporate Transformer-based natural language representations will outperform models that do not. This will be the case for predicting a single measure of reflection depth as well as for identifying individual components that make up a deep reflection.

H.1.1. The use of pre-trained Transformer-based embeddings to represent students' reflections will improve performance compared to binary unigrams, tf-idf, GloVe, and ELMo.

*When comparing the performance of these natural language representations across the same set of machine learning algorithms, Transformer-based representations (i.e., BERT-family models) did not consistently outperform competing baseline representations for automatically assessing reflection depth. However, FLAN-T5 demonstrated improved performance over other representations for automatically detecting several components of reflection depth, but did not achieve the highest performance across all tasks.*

As a result, this hypothesis is partially accepted.

H.1.2. Fine-tuning Transformer-based embeddings on text extracted from CRYSTAL ISLAND will improve predictive performance compared to off-the-shelf pre-trained embeddings.

*Fine-tuning models from the BERT family via masked language modeling on text extracted from CRYSTAL ISLAND did not result in improved performance for automatically assessing the depth of students' reflections. Moreover, this fine-tuning often led to decreased performance compared to models that were trained on language representations created by versions of the language models that had only undergone pre-training.*

This hypothesis is rejected.

H.2. Data augmentation and reflection synthesis techniques will improve performance of reflection depth assessment models compared to models trained solely on genuine student reflections.

H.2.1. The use of transformation-based data augmentation techniques (i.e., mask-filling using BERT, noise injection, and back-translation) to provide more data for model training will improve predictive performance.

*Predictive performance for automatically assessing the depth of students' reflections was substantially improved by transformation-based data augmentation techniques, with back translation achieving the highest performance. Models trained on this data were able to achieve higher overall performance than models trained solely on genuine student reflections, and they were able to achieve higher performance using a smaller number of genuine student reflections for training.*

This hypothesis is accepted.

H.2.2. The use of synthetic reflections generated by large language models based on sample gameplay interaction sequences and desired reflection characteristics will significantly improve predictive performance over transformation-based augmentation techniques.

*Predictive performance for automatically assessing the depth of students' reflections was substantially improved by the incorporation of LLM-generated synthetic reflections compared to models that were trained exclusively on genuine student reflections, but did not outperform models that leveraged baseline data augmentation techniques. However, manually assigning depth scores to the synthetic reflections and then using those for model training resulted in substantial performance improvements compared to all competing approaches. Models trained on synthetic data with manually assigned depth scores achieved higher overall performance than all other approaches and were able to achieve higher performance using a smaller number of genuine student reflections for training.*

This hypothesis is accepted.

H.3. Zero-shot Transformer-based approaches for automatically identifying which gameplay interactions students referred to in their reflections will achieve equivalent performance compared to supervised Transformer-based approaches.

*Using FLAN-T5 with a few-shot approach that included instructions about the reflection-interaction linking task achieved worse performance compared to FLAN-T5 models that were*

*fine-tuned on examples that demonstrated the link between information contained in gameplay interactions and the content of a reflection.*

This hypothesis is rejected.

H.4. Predictions of students' learning outcomes based on features that capture the content of students' reflections will outperform predictive models based on gameplay trace data or non-content-based reflection features.

*Random forest classifiers trained to predict students' learning outcomes (i.e., successfully solving the mystery, post-test score, and normalized learning gain) achieved different performance when trained on different features. Models that were trained using features that capture the content of students' reflections in addition to features capturing gameplay behaviors and other reflection behaviors were shown to achieve improved performance when predicting normalized learning gain but not for predicting whether a student would solve the mystery or their post-test score.*

As a result, this hypothesis is partially accepted.

## **8.2 Summary**

This dissertation has presented research that focuses on the reflection phase of SRL, in which students evaluate what they have learned and the strategies they have enacted during a learning experience to drive adaptations that may be necessary for achieving their learning or problem-solving goals. We have investigated reflection during game-based learning by using machine learning and natural language processing techniques to analyze students' written reflections collected during interactions with the CRYSTAL ISLAND game-based learning environment. Leveraging recent advances in natural language processing based on the Transformer architecture, this dissertation has introduced a natural language processing-based framework for automatically analyzing students' reflections during game-based learning in terms of how effectively they reflect and what they choose to reflect on. We have presented several ways to characterize the reflections that students produce, such as the depth and topic of the reflection, and explored relationships between students' reflection behaviors and their learning outcomes. We have also introduced contextual analyses that provide insights into the relationship between students' learning activities during game-based learning and their reflection behaviors, which addresses a significant gap in the current research on reflection. Finally, we have leveraged data

augmentation techniques and Transformer-based language models to train robust machine learning models for automatically assessing the depth of students' reflections, demonstrating that models trained on synthetic reflections with human-assigned labels significantly improve predictive performance compared to models trained exclusively on genuine student reflections. This work advances research on reflective writing analytics by creating a unified framework for automatically analyzing students' reflections during game-based learning.

### **8.3 Future Work**

There are several promising directions for future research based on the work presented in this dissertation. First, the automated reflection analysis framework presented in this work can be evaluated in a new setting to investigate its generalizability. For example, student reflections could be collected in a different game-based learning environment or a different prompt could be used to capture different types of reflections in CRYSTAL ISLAND. Then, we could use the same approach to train machine learning models using that data and evaluate whether performance is similar to what we presented in this work. Additionally, we could investigate transfer learning between environments or prompts to determine whether models that are trained on the CRYSTAL ISLAND reflections that we explored in this work can readily be used to assess reflection depth for reflections that share either the same learning environment or prompting language.

Another promising direction is to investigate more generalizable ways to label students' reflections. The rubric that we developed in this work focuses on deep versus shallow behaviors in the context of a science problem solving activity. This is not specific only to CRYSTAL ISLAND, but it is still not general enough to be applied to any setting in which we might wish to support students' reflection skills. For example, a common time to ask students to reflect on their learning behaviors is after the completion of a curricular unit in a class. In such a setting, students would probably not be expected to include a hypothesis in their reflections (although, a broad hypothesis such as "I think I did well on this test because I studied flashcards every night" would not be out of place), so the rubric used in this work would not be applicable. The more general reflection components that were analyzed in Chapter 7.2, such as the presence of forward and backward-looking thoughts and evidence of information processing, could inform a measure of reflection depth that could be applied more broadly.

There is also an opportunity to investigate how adaptive guidance and feedback driven by the automated reflection analysis framework impacts students' reflection behaviors and learning outcomes. Interventions based on the analytics provided by the framework can be deployed in an AI-augmented version of CRYSTAL ISLAND and students' reflection depth scores, learning gains, and ability to solve the game's mystery can be compared to students who interact with a control version of the game without any AI-driven adaptivity. A specific intervention that may be interesting to investigate would be to use LLMs to create contextualized reflections in real-time that could be presented to students as examples of possible reflections. This could be a good scaffolding tool to help students see, for instance, what a good versus poor reflection might look like at a given point in the learning scenario. Also, it would be beneficial to explore approaches for more deeply integrating reflection into the game-based learning environment in such a way that how students reflect actually has an impact on what happens in the game. This could provide stronger motivation to engage in reflection and could also help students see how reflection can be important in science problem solving, which are both important aspects of supporting the development of reflection skills.

Finally, given that reflection is just one component of self-regulated learning and that it has such strong ties with goal setting and planning, another promising direction for future research is to identify a way to determine what a student should reflect on at a given point in the learning scenario to help nudge them toward productive reflection topics. With enough data, we may be able to use data mining techniques to identify reflection topics that have proven to be beneficial for previous students in a similar gameplay context, or we could work to develop a theory-driven approach to ranking potential reflection topics. An alternative approach would be to directly ask students to state their goals as they interact with CRYSTAL ISLAND. Then we could investigate the relationship between goal setting and reflection, for instance by examining whether effective reflection corresponds to effective goal setting. We could use students' declared goals as preferred topics for reflection since reflection is goal-oriented and should align with what students said they want to do in the game. Thus, if our system detects that students are not reflecting on the goals that they have set for themselves, we could intervene and suggest that they review their goals. However, we need to be cautious about how we design such a system, since students often get fatigued when explicitly reporting their self-regulated learning

processes with any of the tools that we have integrated into our game-based learning environments.

## REFERENCES

- Ahmed, H., Hina, S., & Asif, R. (2021). Evaluation of descriptive answers of open ended questions using NLP techniques. *Proceedings - 2021 IEEE 4th International Conference on Computing and Information Sciences, ICCIS 2021*.  
<https://doi.org/10.1109/ICCIS54243.2021.9676405>
- Akbari, R. (2007). Reflections on reflection: A critical appraisal of reflective practices in L2 teacher education. *System, 35*(2), 192–207. <https://doi.org/10.1016/j.system.2006.12.008>
- Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique : A Comparative Study. *Image and Signal Processing, 1*, 317–325. <https://doi.org/10.1007/978-3-030-51935-3>
- Azen, R., & Walker, C. M. (2021). *Categorical data analysis for the behavioral and social sciences*. Routledge.
- Azevedo, R., & Gašević, D. (2019). Analyzing Multimodal Multichannel Data about Self-Regulated Learning with Advanced Learning Technologies: Issues and Challenges. In *Computers in Human Behavior* (Vol. 96, pp. 207–210). Elsevier Ltd.  
<https://doi.org/10.1016/j.chb.2019.03.025>
- Barthakur, A., Joksimovic, S., Kovanovic, V., Mello, R. F., Taylor, M., Richey, M., & Pardo, A. (2022). Understanding Depth of Reflective Writing in Workplace Learning Assessments Using Machine Learning Classification. *IEEE Transactions on Learning Technologies, 15*(5), 567–578. <https://doi.org/10.1109/TLT.2022.3162546>
- Bernacki, M. L., Vosicka, L., Utz, J. C., & Warren, C. B. (2021). Effects of Digital Learning Skill Training on the Academic Performance of Undergraduates in Science and Mathematics. *Journal of Educational Psychology, 113*(6), 1107–1125. <https://doi.org/10.1037/edu0000485>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. [https://doi.org/10.1162/tacl\\_a\\_00051/1567442/tacl\\_a\\_00051.pdf](https://doi.org/10.1162/tacl_a_00051/1567442/tacl_a_00051.pdf)

- Boulanger, D., & Kumar, V. (2020). SHAPed automated essay scoring: explaining writing features' contributions to english writing organization. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12149 LNCS, 68–78. [https://doi.org/10.1007/978-3-030-49663-0\\_10](https://doi.org/10.1007/978-3-030-49663-0_10)
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *ArXiv Preprint ArXiv:1508.05326*. <http://arxiv.org/abs/1508.05326>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners*. <http://arxiv.org/abs/2005.14165>
- Carpenter, D., Cloude, E., Rowe, J., Azevedo, R., & Lester, J. (2021). Investigating Student Reflection during Game-Based Learning in Middle Grades Science. *LAK21: 11th International Learning Analytics and Knowledge Conference*, 280–291.
- Carpenter, D., Emerson, A., Mott, B. W., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., & Lester, J. C. (2020). Detecting off-task behavior from student dialogue in game-based collaborative learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12163 LNAI, 55–66. [https://doi.org/10.1007/978-3-030-52237-7\\_5](https://doi.org/10.1007/978-3-030-52237-7_5)
- Carpenter, D., Geden, M., Rowe, J., Azevedo, R., & Lester, J. (2020). Automated analysis of middle school students' written reflections during game-based learning. *Proceedings of the 21st International Conference on Artificial Intelligence in Education*, 67–78. [https://doi.org/10.1007/978-3-030-52237-7\\_6](https://doi.org/10.1007/978-3-030-52237-7_6)
- Carter, R. A., Rice, M., Yang, S., & Jackson, H. A. (2020). Self-regulated learning in online learning environments: strategies for remote learning. *Information and Learning Science*, 121(5–6), 311–319. <https://doi.org/10.1108/ILS-04-2020-0114>
- Cavilla, D. (2017). The Effects of Student Reflection on Academic Performance and Motivation. *SAGE Open*, 7(3), 1–13. <https://doi.org/10.1177/2158244017733790>

- Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: A case study of pre-Service teachers' journals. *ACM International Conference Proceeding Series, 25-29-April*, 1–5. <https://doi.org/10.1145/2883851.2883951>
- Chronopoulou, A., Baziotis, C., & Potamianos, A. (2019). An Embarrassingly Simple Approach for Transfer Learning from Pretrained Language Models. *ArXiv Preprint ArXiv:1902.10547*. <http://arxiv.org/abs/1902.10547>
- Chung, H. W., Longpre, S., Zoph, B., Castro-ros, A., Yu, A., Dai, A., Chi, E. H., & Le, Q. V. (2022). Scaling Instruction-Finetuned Language Models. *ArXiv Preprint ArXiv:2210.11416*.
- Coates, A., Huval, B., Wang, T., Wu, D. J., Ng, A. Y., & Catanzaro, B. (2013). Deep learning with COTS HPC systems. *International Conference on Machine Learning*, 1337–1345.
- Cochran, K., Cohn, C., Hutchins, N., Biswas, G., & Hastings, P. (2022). Improving Automated Evaluation of Formative Assessments with Text Data Augmentation. *International Conference on Artificial Intelligence in Education*, 390–401. [https://doi.org/10.1007/978-3-031-11644-5\\_32](https://doi.org/10.1007/978-3-031-11644-5_32)
- Condor, A., Litster, M., & Pardos, Z. (2021). Automatic short answer grading with SBERT on out-of-sample questions. *International Educational Data Mining Society*. <https://educationaldatamining.org/edm2021/>
- Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research*, 11(2), 251–270. <https://doi.org/10.17239/jowr-2019.11.02.01>
- Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, 11(2), 121–136. <https://doi.org/10.1080/15248371003699969>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv Preprint ArXiv:1810.04805*. <http://arxiv.org/abs/1810.04805>

- Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. (2020). Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *ArXiv Preprint ArXiv:2002.06305*. <http://arxiv.org/abs/2002.06305>
- Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., & Sui, Z. (2022). A Survey on In-context Learning. *ArXiv Preprint ArXiv:2301.00234*. <http://arxiv.org/abs/2301.00234>
- Engelmann, K., & Bannert, M. (2021). Analyzing temporal data for understanding the learning process induced by metacognitive prompts. *Learning and Instruction, 72*. <https://doi.org/10.1016/j.learninstruc.2019.05.002>
- Fan, Y., Saint, J., Singh, S., Jovanovic, J., & Gašević, D. (2021). A learning analytic approach to unveiling self-regulatory processes in learning tactics. *ACM International Conference Proceeding Series*, 184–195. <https://doi.org/10.1145/3448139.3448211>
- Fincham, E., Gašević, D., Jovanović, J., & Pardo, A. (2019). From Study Tactics to Learning Strategies: An Analytical Method for Extracting Interpretable Representations. *IEEE Transactions on Learning Technologies, 12*(1), 59–72. <https://doi.org/10.1109/TLT.2018.2823317>
- Flavell, J. H. (1979). Metacognition and Cognitive Monitoring A New Area of Cognitive-Developmental Inquiry A Model of Cognitive Monitoring. *American Psychologist, 34*(10), 906–911.
- Gao, Z., Lynch, C., Heckman, S., & Barnes, T. (2022). Automatically classifying student help requests: a multi-year analysis. *Proceedings of the 15th International Conference on Educational Data Mining*. <https://educationaldatamining.org/edm2021/>
- Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive Student Modeling in Game-Based Learning Environments with Word Embedding Representations of Reflection. *International Journal of Artificial Intelligence in Education, 31*(1), 1–21. <https://doi.org/10.1007/s40593-020-00220-4>

- Gibson, A., Aitken, A., Agnes Sándor, A., Buckingham Shum, S., Tsingos-Lucas, C., Knight, S., & Sándor, A. (2017). Reflective Writing Analytics for Actionable Feedback. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 153–162.  
<https://doi.org/10.1145/3027385.3027436>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology*, 34(1), 18–29. <https://doi.org/10.1016/j.cedpsych.2008.05.006>
- Greene, J. A., & Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist*, 4(45), 203–209.
- Grivokostopoulou, F., Paraskevas, M., Perikos, I., Nikolic, S., Kovas, K., & Hatzilygeroudis, I. (2018). Examining the Impact of Pedagogical Agents on Students Learning Experience in Virtual Worlds. *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 602–607.  
<https://ro.uow.edu.au/eispapers1><https://ro.uow.edu.au/eispapers1/2439>
- Grootendorst, M. (2020). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv Preprint ArXiv:2203.05794*.
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks*.  
<http://arxiv.org/abs/2004.10964>
- Hasanah, U., Permanasari, A. E., Kusumawardani, S. S., & Pribadi, F. S. (2019). A scoring rubric for automatic short answer grading system. *Telkomnika (Telecommunication Computing Electronics and Control)*, 17(2), 763–770.  
<https://doi.org/10.12928/TELKOMNIKA.V17I2.11785>

- Hastings, P., Hughes, S., & Britt, M. A. (2018). Active Learning for Improving Machine Learning of Student Explanatory Essays. *Proceedings of the International Conference on Artificial Intelligence in Education*, 140–153.  
<http://reed.cs.depaul.edu/peterh/papers/HastingsAIED2018.pdf>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. [http://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf?casa\\_token=UIArJxWxv8AAAAAA:Oy-NpjsCKFVK4lQyTv9Y8qkPbw3k0AbSmo-KCKA94GvJOuMTSs24mHMnMd2iB6Tzdelifn2D](http://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf?casa_token=UIArJxWxv8AAAAAA:Oy-NpjsCKFVK4lQyTv9Y8qkPbw3k0AbSmo-KCKA94GvJOuMTSs24mHMnMd2iB6Tzdelifn2D)
- Hossen, M. S., Chowdhury, M. N. A., Sristy, A. M., & Jahan, N. (2022). Sentiment Analysis using Machine Learning and NLP for Digital Education. *Proceedings - 6th International Conference on Computing Methodologies and Communication, ICCMC 2022*, 902–908.  
<https://doi.org/10.1109/ICCMC53470.2022.9754065>
- Hsiao, I.-H., Huang, P.-K., Murphy, H., & Carey, W. P. (2017). Uncovering Reviewing and Reflecting Behaviors From Paper-based Formal Assessment. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 319–328.  
<https://doi.org/10.1145/3027385.3027415>
- Hu, Y., Jing, X., Ko, Y., & Rayz, J. T. (2020). Misspelling Correction with Pre-trained Contextual Language Model. *IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing*, 144–149. <http://arxiv.org/abs/2101.03204>
- Jensen, E., Umada, T., Hunkins, N. C., Hutt, S., Huggins-Manley, A. C., & D’mello, S. K. (2021). What you do predicts how you do: Prospectively modeling student quiz performance using activity features in an online learning environment. *ACM International Conference Proceeding Series*, 121–131. <https://doi.org/10.1145/3448139.3448151>
- Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., & Gehringer, E. (2022). Insta-Reviewer: A Data-Driven Approach for Generating Instant Feedback on Students’ Project Reports. *International Educational Data Mining Society*. <https://doi.org/10.5281/zenodo.6853099>

- Jivet, I., Wong, J., Scheffel, M., Valle Torre, M., Specht, M., & Drachsler, H. (2021). Quantum of choice: How learners' feedback monitoring decisions, goals and self-regulated learning skills are related. *ACM International Conference Proceeding Series*, 416–427.  
<https://doi.org/10.1145/3448139.3448179>
- Jung, Y., & Wise, A. F. (2020). How and how well do students reflect?: Multi-dimensional automated reflection assessment in health professions education. *ACM International Conference Proceeding Series*, 595–604. <https://doi.org/10.1145/3375462.3375528>
- Kastrati, Z., Dalipi, F., Imran, A. S., Nuci, K. P., & Wani, M. A. (2021). Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. In *Applied Sciences (Switzerland)* (Vol. 11, Issue 9). MDPI AG. <https://doi.org/10.3390/app11093986>
- Keerthi Kumar, H. M., & Harish, B. S. (2018). Classification of short text using various preprocessing techniques: An empirical evaluation. *Recent Findings in Intelligent Computing Techniques: Proceedings of the 5th ICACNI 2017*, 709, 19–30.  
[https://doi.org/10.1007/978-981-10-8633-5\\_3](https://doi.org/10.1007/978-981-10-8633-5_3)
- Kinnebrew, J. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *JEDM-Journal of ...*, 5(1), 190–219.  
<http://www.educationaldatamining.org/JEDM13/index.php/JEDM/article/view/34>
- Klahr, D., & Dunbar, K. (1988). Dual Space Search During Scientific Reasoning. *Cognitive Science*, 12, 1–48.
- Kojima, T., Reid, M., & Gu, S. S. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 22199–22213.
- Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-Reflections through learning analytics. *ACM International Conference Proceeding Series, March*, 389–398.  
<https://doi.org/10.1145/3170358.3170374>
- Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *ArXiv Preprint ArXiv:1808.06226*.  
<http://arxiv.org/abs/1808.06226>

- Laban, P., Wu, C.-S., Murakhovs'ka, L., Liu, W., & Xiong, C. (2022). Quiz Design Task: Helping Teachers Create Quizzes with Automated Question Generation. *ArXiv Preprint ArXiv:2205.01730*. <http://arxiv.org/abs/2205.01730>
- Law, C.-Y., Grundy, J., Vasa, R., Cain, A., & Cummaudo, A. (2016). User Perceptions of Using an Open Learner Model Visualisation Tool for Facilitating Self-regulated Learning. *PervasiveHealth: Pervasive Computing Technologies for Healthcare, 2016-January*, 278–282. <https://doi.org/10.1145/12345.67890>
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lim, L. A., Gasevic, D., Matcha, W., Ahmad Uzir, N., & Dawson, S. (2021). Impact of learning analytics feedback on self-regulated learning: Triangulating behavioural logs with students' recall. *ACM International Conference Proceeding Series*, 364–374. <https://doi.org/10.1145/3448139.3448174>
- Lim, L. A., Gentili, S., Pardo, A., Kovanović, V., Whitelock-Wainwright, A., Gašević, D., & Dawson, S. (2021). What changes, and for whom? A study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction*, 72. <https://doi.org/10.1016/j.learninstruc.2019.04.003>
- Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021). A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing. *International Conference on Artificial Intelligence in Education*, 255–267. [https://doi.org/10.1007/978-3-030-78292-4\\_21](https://doi.org/10.1007/978-3-030-78292-4_21)
- Liu, M., Shum, S. B., Mantzourani, E., & Lucas, C. (2019). Evaluating machine learning approaches to classify pharmacy students' reflective statements. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11625 LNAI, 220–230. [https://doi.org/10.1007/978-3-030-23204-7\\_19](https://doi.org/10.1007/978-3-030-23204-7_19)

- Liu, W., Lin, S., Gao, B., Huang, K., Liu, W., Huang, Z., Feng, J., Chen, X., & Huang, F. (2022). BERT-POS: Sentiment Analysis of MOOC Reviews Based on BERT with Part-of-Speech Information. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13356 LNCS, 371–374. [https://doi.org/10.1007/978-3-031-11647-6\\_72](https://doi.org/10.1007/978-3-031-11647-6_72)
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv Preprint ArXiv:1907.11692*. <http://arxiv.org/abs/1907.11692>
- Luo, T., & Baaki, J. (2019). Scaffolding problem-solving and instructional design processes: Engaging students in reflection-in-action and external representations in three online courses. In *Student-Centered Virtual Learning Environments in Higher Education* (pp. 40–69). IGI Global.
- Lust, G., Elen, J., & Clarebout, G. (2013). Regulation of tool-use within a blended course: Student differences and performance effects. *Computers and Education*, 60(1), 385–395. <https://doi.org/10.1016/j.compedu.2012.09.001>
- Magooda, A., Litman, D., Ashraf, A., & Menekse, M. (2022). Improving the Quality of Students' Written Reflections Using Natural Language Processing: Model Design and Classroom Evaluation. *International Conference on Artificial Intelligence in Education*, 519–525. [https://doi.org/10.1007/978-3-031-11644-5\\_43](https://doi.org/10.1007/978-3-031-11644-5_43)
- Mavrikis, M., Cukurova, M., Di Mitri, D., Schneider, J., & Draschler, H. (2021). A short history, emerging challenges and co-operation structures for Artificial Intelligence in Education. *Bildung Und Erziehung*, 74(3), 249–263.
- Mayfield, E., & Black, A. W. (2020). Should You Fine-Tune BERT for Automated Essay Scoring? *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 151–162. <https://course.fast.ai/>
- McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11).

- McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv Preprint ArXiv:1802.03426*.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv Preprint ArXiv:1301.3781*, 1–12. <https://doi.org/10.1162/153244303322533223>
- Min, B., Ross, H., Sulem, E., Veyseh, A. P. Ben, Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2021). Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *ACM Computing Surveys*, *56*(2), 1–40.
- Molenaar, I., Horvers, A., & Baker, R. S. (2019). Towards Hybrid Human-System Regulation. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 471–480. <https://doi.org/10.1145/3303772.3303780>
- Molenaar, I., Horvers, A., Dijkstra, R., & Baker, R. S. (2020). Personalized visualizations to promote young learners' SRL: The learning path app. *ACM International Conference Proceeding Series*, 330–339. <https://doi.org/10.1145/3375462.3375465>
- Moradi, M., & Samwald, M. (2021). Evaluating the robustness of neural language models to input perturbations. *ArXiv Preprint*.
- Nehyba, J., & Štefánik, M. (2022). Applications of deep language models for reflective writings. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-022-11254-7>
- Nikiforos, S., Tzanavaris, S., & Kermanidis, K. L. (2020). Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing. *Journal of Computers in Education*, *7*(4), 531–551. <https://doi.org/10.1007/s40692-020-00166-5>
- OpenAI. (2022). *Introducing ChatGPT*. <https://openai.com/blog/chatgpt>
- Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604–624. <https://doi.org/10.1109/TNNLS.2020.2979670>
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*(APR), 1–28. <https://doi.org/10.3389/fpsyg.2017.00422>

- Park, K., Sohn, H., Min, W., Mott, B., Glazewski, K., Hmelo-Silver, C. E., & Lester, J. (2022). Disruptive Talk Detection in Multi-Party Dialogue within Collaborative Learning Environments. *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 490–499.
- Patel, N., Baker, S. G., & Scherer, L. D. (2019). Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs. *Journal of Experimental Psychology*, 148(12), 2129–2153. <https://doi.org/10.31234/osf.io/xeyj8>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot Matthieu, & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <http://scikit-learn.sourceforge.net>.
- Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Franci, M. E. (2014). *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates.
- Pennington, J., Socher, R., & Manning, C. (2015). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543. <https://doi.org/10.3115/v1/d14-1162>
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv Preprint ArXiv:1802.05365*. <http://arxiv.org/abs/1802.05365>
- Pintrich, P. R. (2000). The Role of Goal Orientation in Self-Regulated Learning. In *Handbook of Self-Regulation* (pp. 451–502). Academic Press.
- Poldner, E., Van der Schaaf, M., Simons, P. R. J., Van Tartwijk, J., & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teacher Education*, 37(3), 348–373. <https://doi.org/10.1080/02619768.2014.892479>

- Pugh, S. L., Krishna Subburaj, S., Ramesh Rao, A., Stewart, A. E., Andrews-Todd, J., & D, S. K. (2021). Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. *International Educational Data Mining Society*.  
<https://educationaldatamining.org/edm2021/>
- Puustinen, M., & Pulkkinen, L. (2001). Models of Self-regulated Learning: A review. In *Scandinavian Journal of Educational Research* (Vol. 45, Issue 3, pp. 269–286).  
<https://doi.org/10.1080/00313830120074206>
- Qiao, Y., Xiong, C., Liu, Z., & Liu, Z. (2019). Understanding the Behaviors of BERT in Ranking. *ArXiv Preprint ArXiv:1904.07531*. <http://arxiv.org/abs/1904.07531>
- Qiu, X. P., Sun, T. X., Xu, Y. G., Shao, Y. F., Dai, N., & Huang, X. J. (2020). Pre-trained models for natural language processing: A survey. In *Science China Technological Sciences* (Vol. 63, Issue 10, pp. 1872–1897). Springer Verlag. <https://doi.org/10.1007/s11431-020-1647-3>
- Rahimi, Z., Litman, D., Correnti, R., Matsumura, L. C., Wang, E., & Kisa, Z. (2014). Automatic Scoring of an Analytical Response-To-Text Assessment. *International Conference on Intelligent Tutoring Systems*, 601–610.
- Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *ArXiv Preprint ArXiv:1606.05250*.  
<http://arxiv.org/abs/1606.05250>
- Raković, M., Kilgour, J., Lim, L., Moore, J., Bannert, M., & Molenaar, I. (2022). Using Learner Trace Data to Understand Metacognitive Processes in Writing from Multiple Sources. *12th International Learning Analytics and Knowledge Conference*, 130–141.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv Preprint*. <http://arxiv.org/abs/1908.10084>
- Riedinger, B. (2006). Mining for meaning: Teaching students how to reflect. In *Handbook of Research on ePortfolios* (pp. 90–101). IGI Global.

- Riordan, B., Bichler, S., Bradford, A., King Chen, J., Wiley, K., Gerard, L., & C. Linn, M. (2020). An empirical investigation of neural methods for content scoring of science explanations. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 135–144. <https://doi.org/10.18653/v1/2020.bea-1.13>
- Riordan, B., Flor, M., & Pugh, R. (2019). How to account for misspellings: Quantifying the benefit of character representations in neural content scoring models. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 116–126. <https://doi.org/10.18653/v1/w19-4411>
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3). <https://doi.org/10.1002/widm.1355>
- Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2012). Exploring Inquiry-based Problem-Solving Strategies in Game-based Learning Environments. *Intelligent Tutoring Systems: 11th International Conference*, 470–475.
- Saint, J., Whitelock-Wainwright, A., Gasevic, D., & Pardo, A. (2020). Trace-SRL: A Framework for Analysis of Microlevel Processes of Self-Regulated Learning from Trace Data. *IEEE Transactions on Learning Technologies*, 13(4), 861–877. <https://doi.org/10.1109/TLT.2020.3027496>
- Saldaña, J. (2021). *The Coding Manual for Qualitative Researchers*. Sage.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv Preprint ArXiv:1910.01108*. <http://arxiv.org/abs/1910.01108>
- Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1848–1858. <https://www.duolingo.com>

- Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gasevic, D., & Chen, G. (2021). Assessing Algorithmic Fairness in Automatic Classifiers of Educational Forum Posts. *Proceedings of the International Conference on Artificial Intelligence in Education*, 381–394.
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76.  
<https://doi.org/10.1016/j.asw.2013.04.001>
- Siadaty, M., Gasevic, D., & Hatala, M. (2016). Trace-based Micro-analytic Measurement of Self-Regulated Learning Processes. *Journal of Learning Analytics*, 3(1).  
<https://doi.org/10.18608/jla.2016.31.11>
- Society for Learning Analytics Research. (2022). *What is learning analytics?* Society for Learning Analytics Research.
- Sonnenberg, C., & Bannert, M. (2016). Evaluating the Impact of Instructional Support Using Data Mining and Process Mining: A Micro-Level Analysis of the Effectiveness of Metacognitive Prompts. *Journal of Educational Data Mining*, 8(2), 51–83.
- Spires, H., Rowe, J., Mott, B., & Lester, J. (2011). Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes. In *Journal of Educational Computing Research* (Vol. 44, Issue 4, pp. 453–472).  
<https://doi.org/10.2190/EC.44.4.e>
- Srivastava, N., Fan, Y., Rakovic, M., Singh, S., Jovanovic, J., Van Der Graaf, J., Lim, L., Surendrannair, S., Kilgour, J., Molenaar, I., Bannert, M., Moore, J., & Gasevic, D. (2022). Effects of Internal and External Conditions on Strategies of Self-regulated Learning: A Learning Analytics Study. *ACM International Conference Proceeding Series*, 392–403.  
<https://doi.org/10.1145/3506860.3506972>
- Stasaski, K., Rathod, M., Tu, T., Xiao, Y., & Hearst, M. A. (2021). Automatically Generating Cause-and-Effect Questions from Passages. *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 158–170.

- Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., & Xiong, C. (2020). Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT. *ArXiv Preprint ArXiv:2003.04985*. <http://arxiv.org/abs/2003.04985>
- Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J. H., & Sumner, T. (2022). The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves. *ArXiv Preprint*. <http://arxiv.org/abs/2204.09652>
- Tarricone, P. (2011). *The taxonomy of metacognition*. Psychology Press.
- Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., & Lester, J. (2018). Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. *Learning and Instruction, 54*, 93–103. <https://doi.org/10.1016/j.learninstruc.2017.08.005>
- Timmers, C. F., Walraven, A., & Veldkamp, B. P. (2015). The effect of regulation feedback in a computer-based formative assessment on information problem solving. *Computers and Education, 87*, 1–9. <https://doi.org/10.1016/j.compedu.2015.03.012>
- Ullmann, T. D. (2015). *Automated detection of reflection in texts - A machine learning based approach*. The Open University.
- Ullmann, T. D. (2019). Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches. *International Journal of Artificial Intelligence in Education, 29*(2), 217–257. <https://doi.org/10.1007/s40593-019-00174-2>
- Van Manen, M. (1977). Linking ways of knowing with ways of being practical. *Curriculum Inquiry, 6*(3), 205–228.
- van Velzen, J. H. (2016). Eleventh-Grade High School Students' Accounts of Mathematical Metacognitive Knowledge: Explicitness and Systematicity. *International Journal of Science and Mathematics Education, 14*(2), 319–333. <https://doi.org/10.1007/s10763-015-9689-3>
- Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*.

- Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online learning environments: A review of empirical research. *ACM International Conference Proceeding Series*, 524–533. <https://doi.org/10.1145/3375462.3375483>
- Vrzakova, H., Amon, M. J., Stewart, A., Duran, N. D., & D’Mello, S. K. (2020). Focused or stuck together: Multimodal patterns reveal triads’ performance in collaborative problem solving. *ACM International Conference Proceeding Series*, 295–304. <https://doi.org/10.1145/3375462.3375467>
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *ArXiv Preprint ArXiv:1804.07461*. <http://arxiv.org/abs/1804.07461>
- Wang, D., & Han, H. (2021). Applying learning analytics dashboards based on process-oriented feedback to improve students’ learning effectiveness. *Journal of Computer Assisted Learning*, 37(2), 487–499. <https://doi.org/10.1111/jcal.12502>
- Wang, Z., Lan, A. S., Nie, W., Waters, A. E., Grimaldi, P. J., & Baraniuk, R. G. (2018, June 26). QG-Net: A Data-Driven question generation model for educational content. *Proceedings of the 5th Annual ACM Conference on Learning at Scale, L at S 2018*. <https://doi.org/10.1145/3231644.3231654>
- Winne, P. H. (2017). Learning Analytics for Self-Regulated Learning. In *Handbook of Learning Analytics* (pp. 241–249). Society for Learning Analytics Research (SoLAR). <https://doi.org/10.18608/hla17.021>
- Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Lawrence Erlbaum Associates Publishers.
- Winne, P. H., Teng, K., Chang, D., Lin, M. P. C., Marzouk, Z., Nesbit, J. C., Patzak, A., Rakovic, M., Samadi, D., & Vytasek, J. (2019). NStudy: Software for learning analytics about learning processes and self-regulated learning. *Journal of Learning Analytics*, 6(2), 95–106. <https://doi.org/10.18608/jla.2019.62.7>

- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. <https://github.com/huggingface/>
- Wong, F. K. Y., Kember, D., Chung, L. Y. F., & Yan, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing*, 22(1), 48–57.
- Xue, K., Yaneva, V., Runyon, C., & Baldwin, P. (2020). Predicting the Difficulty and Response Time of Multiple Choice Questions Using Transfer Learning. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 193–197. <https://doi.org/10.18653/v1/2020.bea-1.20>
- Zhang, H., & Litman, D. (2021). Essay Quality Signals as Weak Supervision for Source-based Essay Scoring. *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 85–96. <https://aclanthology.org/2021.bea-1.9>
- Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2022). An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, 30(1), 177–190. <https://doi.org/10.1080/10494820.2019.1648300>
- Zhang, Y., Shah, R., & Chi, M. (2016). Deep Learning + Student Modeling + Clustering: a Recipe for Effective Automatic Short Answer Grading. *International Educational Data Mining Society*. [http://www.educationaldatamining.org/EDM2016/proceedings/paper\\_61.pdf](http://www.educationaldatamining.org/EDM2016/proceedings/paper_61.pdf)
- Zimmerman, B. J. (2000). Attaining self-regulation: a social cognitive perspective. In *Handbook of Self-Regulation* (pp. 13–39). Academic Press.