

ABSTRACT

FARHANA, EFFAT. Science Reading Behavior of Middle School Students within a Digital Literacy Platform. (Under the direction of Dr. Collin F. Lynch).

Reading is an integral part of learning. The purpose of *reading to learn* is to comprehend meaning from informational texts. Reading comprehension tasks require self-regulated learning (SRL) behaviors—to plan, monitor, and evaluate one’s reading strategies. Students without SRL skills may struggle in reading which in turn may inhibit them to acquire domain-specific knowledge. Thus, understanding students reading behavior and SRL usage is important for intervention.

In this thesis, we aim to study students’ science reading and SRL and connect those activities with performance. In doing so, we apply theory and methodology from learning science to analyze students’ activity data within a digital literacy platform, *Actively Learn*. The work consists of four studies: (i) analyzing relationship between SRL and question feature, (ii) identifying patterns those differ between productive and unproductive students, (iii) assessing reading text and question difficulty, and (iv) analyzing the association of students’ SRL usage after receiving teachers’ feedback on questions.

© Copyright 2021 by Effat Farhana

All Rights Reserved

Science Reading Behavior of Middle School Students within a Digital Literacy Platform

by
Effat Farhana

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
2021

APPROVED BY:

Dr. James Lester

Dr. Noboru Matsuda

Dr. Teomara Rutherford

Dr. Collin F. Lynch
Chair of Advisory Committee

DEDICATION

To my parents, Dr. Begum Rahila Akhtar and Dr. Muhammad Elias Ali, who always encouraged me in all of my pursuits.

BIOGRAPHY

Effat Farhana was born in Bangladesh. She became interested in engineering during her high school years. She then joined Bangladesh University of Engineering and Technology (BUET) to study Computer Science and Engineering. After getting her bachelor's degree in 2011, she worked as a Lecturer in the Computer Science Department at a private university.

She started her PhD in Computer Science at North Carolina in Fall 2015. She joined ArgLab and started working with Dr. Lynch on an educational data mining project. On April 28, 2021, she defended her thesis titled "Science Reading Behavior of Middle School Students within a Digital Literacy Platform". She will join as a Postdoctoral Fellow in the Electrical Engineering & Computer Science (EECS) department at Vanderbilt University.

Aside from studies, she enjoys hiking, traveling, and window shopping.

ACKNOWLEDGEMENTS

I have received a great deal of support and assistance throughout the writing of this dissertation.

First and foremost, I would like to thank my advisor, Dr. Collin F. Lynch. I was his student in a special topic course in Fall 2017. This was the first time I came to know about “*Educational*” Data Mining. Later, I joined his lab, ArgLab, to continue my PhD. I had great flexibility in formulating research questions and methodology of my dissertation. I really enjoyed the years I spent at ArgLab.

I would also like to thank my committee member, Dr. Rutherford, who guided me in all my studies. I would like to thank all whom I have worked with during my years at NC State. My labmates at ArgLab, including Shreeya, Jen, Adam, and Niki were all cooperative and helped in my draft proofreading and providing suggestions. The Faculty and Staff of the Department of Computer Science, including Dr. Rouskas, Carol Allen, Linda Honeycutt, Kathy, and Ken Tate were always supportive of my academics.

Lastly, I would like to thank my parents, who always encourage me to excel academically. I am thankful to my husband, Akond, who has been a great support throughout my PhD journey.

TABLE OF CONTENTS

List of Tables	viii
List of Figures	x
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Thesis Roadmap	5
Chapter 2 Background	8
2.1 Self-Regulated Learning	8
2.1.1 SRL: Definition and SRL Models	8
2.1.2 Reading and SRL	12
2.1.3 SRL and CBLE	13
2.2 SRL in Learning Management Systems	16
2.3 Reading Platforms	18
2.4 Reading Literature in Educational Data Mining	21
2.4.1 Reading Speed and Student Modelling	21
Chapter 3 Actively Learn Platform	23
3.1 The Actively Learn Platform	23
3.1.1 The Actively Learn Platform	23
3.1.2 SRL in AL	24
3.1.3 Comparison with Other Reading Platforms	26
3.2 Dataset Construction	27
Chapter 4 SRL and Question Features	28
4.1 Background	29
4.1.1 Question Format	29
4.1.2 Predictive Student Modelling	30
4.2 SRL and Question Features	31
4.2.1 Methodology of RQ1.1	32
4.2.2 Results of RQ1.1	33
4.2.3 Discussion of RQ1.1	34
4.3 SA Response Analysis	35
4.3.1 Methodology of RQ1.2	35
4.3.2 Results of RQ1.2	36
4.3.3 Results of Statistical Analysis	36
4.3.4 Results of Score and Cosine Similarity	36
4.4 Predictive Student Modelling & Temporal SRL Behavior	37
4.4.1 Methodology of RQ1.3	38
4.4.2 Experiments of RQ1.3	42

4.4.3	Results of RQ1.3	43
4.4.4	Discussion of RQ1.3	45
4.5	Conclusions	47
Chapter 5	SRL Patterns and Performance	48
5.1	Background	49
5.1.1	Sequential Pattern Mining	49
5.1.2	SRL in Cross Domain	50
5.2	SRL Patterns and Performance	51
5.2.1	Methodology of RQ2	52
5.2.2	Results: RQ2.1	55
5.2.3	Results: RQ2.2	57
5.3	Discussion of RQ2	63
5.4	Conclusion and Future Work	64
Chapter 6	SRL and Content Difficulty	65
6.1	Background	66
6.1.1	Background: Question Difficulty	66
6.1.2	Background: Text Complexity and Question Positioning	67
6.2	SRL and Question Difficulty	69
6.2.1	Dataset	70
6.2.2	Methodology of RQ3.1	70
6.2.3	Result of RQ3.1	72
6.2.4	Discussion of RQ3.1	76
6.3	SRL and Text-Complexity	77
6.3.1	Methodology of RQ3.2.1	77
6.3.2	Methodology of RQ3.2.2	81
6.3.3	Results of RQ3.2.1	82
6.3.4	Results of RQ3.2.2	85
6.3.5	Discussion of RQ3.2	87
6.4	Conclusion	89
Chapter 7	Feedback And Students' SRL	90
7.1	Background	90
7.2	Dataset Preparation	92
7.3	Methodology	93
7.3.1	RQ4.1 Methodology	93
7.3.2	RQ4.2 Methodology	93
7.4	Results	94
7.4.1	RQ4.1 Result	94
7.4.2	RQ4.2 Results	96
7.5	Discussion	97
7.6	Conclusion	97

Chapter 8	Conclusions and Future Work	99
8.1	Contributions	100
8.2	Design Implications	101
8.2.1	Design Implicaion: EDM Community	101
8.2.2	Design Implication: Learning Science	103
8.2.3	Design Implication: The AL Platform	103
8.3	Future Work	104
References		106

LIST OF TABLES

Table 1.1	Thesis Roadmap. Each Chapter presents one broad RQ, usage of AL interaction data, and task associated with the RQ.	6
Table 3.1	AL Dataset Construction: Science	27
Table 4.1	Mean and (Standard Deviation) of Descriptive Variables	33
Table 4.2	Results from Spearman correlation measuring association between SRL and science score	34
Table 4.3	Results from HLM measuring association between SRL and science score	35
Table 4.4	Results from HLM measuring association between cosine similarity (SA question text, response) and score	37
Table 4.5	A hypothetical Student Action Log in AL	37
Table 4.6	Experimental Results. Mean (Standard Deviation) of 5-fold cross validation on test dataset. w = window size.	43
Table 5.1	Mean and (Standard Deviation) of Performance feature for Science and Social Studies	53
Table 5.2	Differential patterns :Science. $*** = p < 0.001$, $** = p < 0.05$	58
Table 5.3	Differential patterns :Social Study. $*** = p < 0.001$, $** = p < 0.05$	60
Table 5.4	Differential patterns :Science vs Social Study. $*** = p < 0.001$, $** = p < 0.05$	62
Table 6.1	Question category by difficulty ratio	71
Table 6.2	Mean with (Standard Deviation), and p value from KW = Kruskal-Wallis test for student behavior features on Easy, Medium, and Hard questions.	74
Table 6.3	Mean with (Standard Deviation), and p value from KW = Kruskal-Wallis test for class-level features on Easy, Medium, and Hard questions.	75
Table 6.4	RQ3.2 Methodology Overview	77
Table 6.5	Descriptives of Article Category by FRES. Mean (SD) of features. n = No. of articles in each category.	80
Table 6.6	Descriptives of Multiple Articles per Assignment. Mean (SD) and Median of features. n = No. of articles in each category.	80
Table 6.7	Vertical Positioning of Questions in Assignments Grouped by FRES.	81
Table 6.8	Question Readability in Assignments Grouped by FRES (Lower value \sim Difficult Readability)	82
Table 6.9	Spearman's Correlation between Question Position and SRL by FRES. Bold = Statistically Significant. R =No. read, A=No. Annotation, H= No. Highlight, V= No. Vocab. lookup	86

Table 6.10	Spearman's Correlation between Question Readability and SRL by FRES. Bold = Statistically Significant. R =No. read, A=No. Notes, H= No. Highlight, V= No. Vocab. lookup	87
Table 7.1	Results from HLM measuring association between SRL and science score	96

LIST OF FIGURES

Figure 1.1	Thesis Overview	3
Figure 2.1	Winne and Hadwin’s Model (COPES Model) (Pieschl et al. (2008)) . .	11
Figure 2.2	MetaTutor platform. (A: time remaining for task completion, B: table of contents, C: pallet for subgoals and progression, D: pedagogical agent, E: SRL and monitoring actions palette (Bouchet et al., 2012))	16
Figure 2.3	An interface of the <i>nstudy</i> platform for defining Learning Objectives (Beaudoin & Winne, 2009)	19
Figure 3.1	A reading text and embedded questions. Question 1 is an MCQ and question 3 is an SA.	25
Figure 4.1	Proposed model. Q = Question Attempt, H = Highlighting, A = Annotation.	39
Figure 4.2	LSTM with Attention	40
Figure 4.3	Case (b) Attention Visualization at test time. True score: 1.00. Predicted score: 0.84. Scores are scaled in [0-1] range. Q = question attempt, A = Annotation. For annotations, <i>Blue</i> font = Student’s note text and <i>black</i> font = selected text.	45
Figure 4.4	Case (a) Attention Visualization at test time. True score = 0.5. Predicted score = 0.59. Scores are scaled in [0-1] range. Q = question attempt, H = Highlighted text.	46
Figure 5.1	Methodology Overview of RQ2	52
Figure 5.2	Clustering Students by Assignment Performance	56
Figure 6.1	Student performance distribution by question difficulty	72
Figure 6.2	ICC and IIC plots from 1PL model	73
Figure 6.3	Effect Size Comparison of SRL among three pairs at student-level. A: Annotating, H: Highlighting, and V: Vocabulary lookup	74
Figure 6.4	Article Category by Flesch Readability: Single-Article Per Assignment	79
Figure 6.5	Formality and FK Grade Levels in RQ3.2.1	82
Figure 6.6	SRL and Article Text Complexity	83
Figure 6.7	Score, Reading and Article Text Complexity	85
Figure 7.1	Avg. cosine similarities of students’ responses to a question	94

CHAPTER

1

INTRODUCTION

1.1 Motivation

Reading to learn is an integral part of learning across domains. Reading to learn involves three types of activities: decoding, comprehension, and mature reading strategies (Forrest-Pressley & Waller, 2013). One such example of a mature reading strategy is re-reading and paraphrasing while trying to understand a set of concepts presented by a writer. Further, mature reading strategies require metacognitive skills, which are a key component of self-regulated learning (SRL). This includes goal setting, self-controlling, self-monitoring, and self-evaluation (Zimmerman, 2000*b*; Zimmerman & Bandura, 1994). Metacognition during reading-to-learn activities help a reader to monitor their reading comprehension, analyze the effectiveness of their strategy, and adapt their strategies when necessary (Forrest-Pressley & Waller, 2013).

Science reading is different from other domain-specific readings. First, science reading involves discipline-specific scientific terminologies and multi-modal knowledge representations including diagrams, charts, symbols, and equations (Yore, 2012). Readers have to navigate through the text as well as interact with above-mentioned knowledge represen-

tations to comprehend the message. Second, science reading has a high vocabulary load (Buehl, 2017; Yore, 2012). For example, *work, force, matter, mass, cell*—these words have different meaning in science and daily use. Researchers have shown that students' reading comprehension and motivation decreases while processing science concept words (Yore & Tippett, 2014). Third, the academic text of science is difficult to comprehend (Yore & Tippett, 2014; Buehl, 2017). Science text has a high lexical density (i.e., the ratio of content words compared to overall words) (Yore & Tippett, 2014). The high lexical density and academic language of science texts make reading comprehension more challenging. Thus, science reading requires SRL skills (Yore, 2012).

The demands on students' reading abilities increase as students enter into middle school (Salinger, 2003). Unfortunately, many students lack adequate reading proficiency as they move from elementary school to middle school. Students in the US, in particular, have lower performance in domain-specific reading comprehension compared to other countries (Snow, 2002*a*). Students get few opportunities to learn reading strategies and SRLs from school, especially in domain-specific areas (Louis Gomez & Gomez, 2007; Paris & Paris, 2001). Also, they are poor at judging and monitoring their reading comprehension levels (Baker, 1985; Garner & Kraus, 1981; Garner & Reis, 1981). Prior researchers have identified significant gaps between the complexity of texts that students encounter in K-12 learning and those they encounter at the college level or in professional settings (ACT, 2006; Williamson, 2006). To mitigate such gap, updated standards, such as the Common Core (CCS, 2010), have focused on increasing expectations for reading levels and text complexity. The new science education framework in the USA proposed by National Research Council emphasized reading comprehension by stating

“ Being literate in science and engineering requires the ability to read and understand their literatures. ... Reading, interpreting, and producing text are fundamental practices of science in particular, and they constitute at least half of engineers' and scientists' total working time.” – National Research Council, 2012, page 74.(NRC, 2012)

Concerns about reading interventions combined with recent advancements in the computer-based learning environment (CBLE) have motivated researchers to study fine-grained student behavior within learning environments. Examples of reading interventions within CBLE include iSTART (McNamara et al., 2007), nStudy (Beaudoin & Winne, 2009), and Rederbench (Dascalu et al., 2015).

In this thesis, we aim to understand students' interaction within a CBLE, Actively Learn (AL), and link those behaviors to their academic performance. One key difference with AL and other reading platforms is AL facilitates curriculum integrated reading articles with *embedded questions*. Embedded questions in AL chunk a long reading text into smaller parts. Unlike reading a long reading text, chunking helps students to organize information and get the main idea from the text (Keenan, 1984). As students proceed with reading in the AL, they can *see* embedded questions, and adapt their reading and SRL strategies, such as highlight, annotate, and look up unknown words.

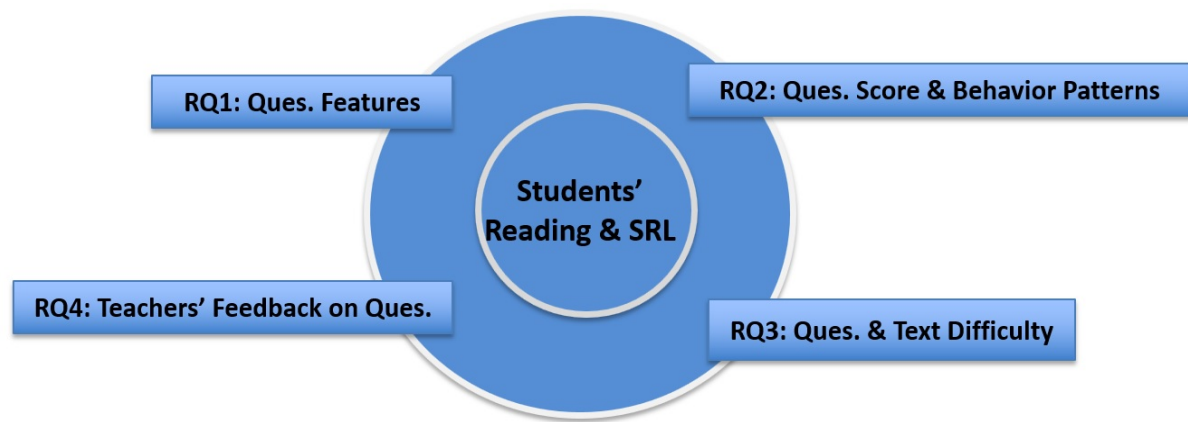


Figure 1.1: Thesis Overview

Thesis Goal The goal of my thesis is to understand students' SRL actions and how those SRL actions are linked with science performance situated in the AL platform—a digital reading platform with *embedded* questions. Figure 1.1 presents the high-level goal of my thesis. We hypothesize, to understand students' reading and SRL behavior within the AL platform, it is necessary to understand how they interact with questions. Towards that end, we will examine four broad research questions (RQs)

RQ1. How do students' reading and SRL strategies vary based upon the features of comprehension questions?

RQ2. *How do students' question scores connect with their reading and SRL patterns?*

RQ3. *How do students' SRL strategies vary with question and text difficulty?*

RQ4. *How do teachers' feedback on questions connect with students' reading and SRL strate-*

gies?

In RQ1, we focus on predictive student modelling. Prior research on predictive student modelling has primarily focused on students' prior-question answers, such as Learning Factors Analysis (LFA) (Chi et al., 2011) and other factor analysis approaches based upon Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994), Q-matrix (Barnes, 2005). These statistical models rely on time-intensive expert annotated skills and represent students' understanding as binary variables. To overcome the limitations of knowledge tracing models, Piech et. al (Piech et al., 2015) introduced the deep knowledge tracing (DKT)—taking into account the temporal ordering and correctness of prior question attempts resulting in a 25% gain in AUC compared to statistical models. Since then, researchers proposed several enhancements (Nagatani et al., 2019; Zhang et al., 2017; Liu et al., 2019). However, all these enhancements focused on primarily students' question solving behavior. These methods do not utilize the temporal ordering of students' interaction behavior within systems, such as reading, video watching, and others. This lacking of combining both question solving and other learning interaction behavior in predictive student modelling is emphasized in a recent study of 2020, Choi et. al (Choi et al., 2020). In RQ1, we aim to address this research gap by combining the temporal ordering of question attempts and SRL in predictive student modelling. We argue that to predict students' question performance, it is necessary to (i) consider other learning activities beyond question-answering and (ii) understand how these activities are related to question-solving behaviors.

The motivation of RQ2 is to identify *domain-specific* reading and SRL behaviors for productive and unproductive students. As stated by Greene et al. (Greene et al., 2013), CBLEs will have limited effectiveness unless learners possess domain-specific SRLs. Although there is a large body of literature in SRL, there is a lack of research on which SRLs are domain-specific (Greene et al., 2015). In another survey of 77 SRL literature conducted by Alexander and colleagues (Alexander et al., 2011) found that 37% (n = 24) did not consider any domain and only 3.90% (n = 3) included cross-domain comparison studies. In RQ2, we examine domain-specific reading and SRL for two subject domains: science and social studies and connect those to academic performance.

In RQ3, we will assess student' SRL behaviors considering question and text difficulty. We hypothesize that in question-embedded AL reading assignments, students first interpret the questions and then adapt their reading and SRL strategies accordingly. We will assess question difficulty from student interaction data at the *class-level* context. SRL researchers emphasized that analyzing SRL requires understanding students' learning contexts (Butler

& Cartier, 2005). At the classroom level, students' engagement in learning is shaped by teachers' instructional approaches and by interactions with the teacher and peers (Butler & Cartier, 2005). Additionally, we hypothesize students may perceive a question as difficult if the text is difficult to read and comprehend. Also, the question positioning in the text may affect performance, as previous research indicated decreasing in human attention span for longer text (Zheng et al., 2019). Towards that end, we examine reading and SRL behavior considering text-readability of text and question placement in the text.

In RQ4, we will examine students' SRL and reading behavior upon receiving feedback on questions. Prior research on teaching and learning has identified a problem of the *feedback gap* (Evans, 2013) — students are unlikely to act upon feedback. However, feedback is effective only if students act upon it (Winstone et al., 2017). In other words, to understand how feedback can help in student learning, it is necessary to understand how they are acting upon feedback. Our RQ4 will focus on analyzing students' reading, SRL, and adopting question responses after receiving teachers' feedback.

1.2 Thesis Roadmap

This thesis is organized as follows.

First, we present background and related literature on reading research, SRL, and reading platforms in Chapter 2. We introduce the AL platform and SRL coding in Chapter 3. Next, we present studies related to four RQs in Chapters 4 to Chapter 7. Table 1.1 presents tasks associated with each RQ and interaction data used. Finally, we conclude and present future research directions in Chapter 8.

In Chapter 4, we first discuss results of a preliminary experiment to understand if the association of SRL varies depending on two question formats. We counted the frequency of SRL and measured association with two question formats in the AL: multiple-choice questions (MCQ) and short answer (SA) questions including fill-in-the-blanks and free text questions. Next, we report *what* SRL actions students performed and *how* those actions predicted question scores. We took the textual content of highlights, annotation, and vocabularies to measure *what* SRL action student performed. To predict the score on a question, we designed experiments taking temporal ordering of SRL contents and questions texts before the question attempt. Additionally, we assessed how the predictive model is assigning weights to its inputs in making predictions.

In Chapter 5, we describe methodology to identify reading and SRL behavioral patterns

Table 1.1: Thesis Roadmap. Each Chapter presents one broad RQ, usage of AL interaction data, and task associated with the RQ.

Chapter	Interaction Data	Task
4	Student	<ul style="list-style-type: none"> • Question format and SRL • Question and SRL's temporal ordering to predict score
5	Student	<ul style="list-style-type: none"> • SRL patterns and clustering students • Cross domain SRL patterns
6	Student	<ul style="list-style-type: none"> • Question difficulty and SRL • Text complexity, question position, and SRL
7	Student Teacher	<ul style="list-style-type: none"> • SRL after receiving feedback on questions

for productive and unproductive students. We also report how patterns vary across two subject domains, namely science and social studies. First, we describe grouping students using question performance features within two subject domains. Next, we describe identifying frequent patterns within each group and tested whether these patterns vary at a statistically significant level.

In Chapter 6, we describe students' SRL behaviors as a proxy for their perception of question difficulty. Our experiments focused on how students may perceive question difficulty at the *class-level*, and how students vary their reading and self-regulated learning activities in response to it. In the later part of the Chapter, we discuss text complexity and its association with SRL. We hypothesize students may perceive a question as difficult if the text is difficult to read and comprehend. Additionally, we discuss our findings of how the length of text prior to a question attempt impacts score.

In Chapter 7, we discuss our results of how students' adapt their SRL behavior upon receiving feedbacks on questions. Contrary to the previous three studies, this study utilizes both teachers' and students' interaction data within the AL system. We report how students modified their question response in subsequent submissions after receiving feedback and exhibited SRL behavior.

We finally conclude in Chapter 8 outlining contributions of this thesis and possible future directions.

The contributions of this thesis are two-fold: for researchers and for student support. From a research perspective, we will describe predictive student modelling combining tem-

poral ordering of question attempts and learning activities in predictive student modelling—an underexplored area identified by Choi et. al (Choi et al., 2020). From the student support perspective, we will present experiments to understand students' reading and SRL behavior as they interact with questions in the AL platform.

CHAPTER

2

BACKGROUND

2.1 Self-Regulated Learning

2.1.1 SRL: Definition and SRL Models

Self-regulated learning (SRL), as described by Zimmermann, involves four regulatory components during learning: goal-setting, self-monitoring, self-evaluating, and using strategies (Zimmerman, 2000*b*). Self-regulated learners tend to set more challenging goals for academic achievements (Zimmerman et al., 1992). They use self-monitoring strategies to monitor their working time and solve tasks (Bouffard-Bouchard et al., 1991). Self-evaluation refers to being able to judge the outcomes of self-monitoring process (Zimmerman & Bandura, 1994). In the process of self-evaluating, students change learning strategies to achieve their learning goals (Zimmerman et al., 1992).

SRL Models

SRL researchers have provided various SRL models, for example, Pintrich's SRL framework (Pintrich, 2000), Zimmerman's Cyclic model (Zimmerman, 2000*a*) and Winne and Hadwin's

model (Winne & Hadwin, 1998). All models rely on different assumptions. However, all models view learning as an active process where learners set goals by understanding topics or domains, regulate their cognition process, and modify behaviors to achieve goals by self-evaluation (Weinstein et al., 2000; Pintrich, 2000). We discuss three SRL models below.

Zimmerman's Cyclic Model and Pintrich's Framework. SRL researchers Zimmerman (Zimmerman, 1989, 2000*b*; Schunk & Zimmerman, 2012) and Pintrich (Pintrich, 2000) developed their models based on social-cognitive theoretical assumption. Social-cognitive theorists view SRL process as being influenced by self-regulated processes and environmental factors (Zimmerman, 1989). Zimmerman's model has three phases: forethought, performance, and self-reflection (Zimmerman, 2000*b*). According to this model, students plan and set goals in the forethought phase, execute and monitor planned tasks in the performance phase, and assess their self-performance in the self-reflection phase. The assessment of the self-reflection phase influences the forethought phase of the next task; implying the cyclic nature of the model. For example, if a student's self-evaluation is unfavorable, then they may not try to learn the material (Zimmerman, 1989). Pintrich's SRL model (Pintrich, 2000) has four phases: (i) forethought, planning, and activation, (ii) monitoring, (iii) control, and (iv) reaction and reflection. Pintrich's SRL model differs from Zimmerman's model by introducing the third phase, control (Panadero, 2017). The model incorporates aspects of SRL by which students "*monitor, control, and regulate the (learning) context*" (Pintrich (2000), p. 469).

Winne and Hadwin's Model. Zimmerman and Pintrich's models have a socio-cognitive theoretical assumption., i.e., a students' SRL is shaped by environmental factors (Zimmerman, 1989). In contrast, the Winne and Hadwin's model (Winne & Hadwin, 1998; Winne, 2011) is described to measure SRL considering *events*. As our study is a retrospective analysis of student interaction data and we do not have demographic details, we chose Winne and Hadwin's model. We discuss this model in detail below.

Winne and Hadwin characterized SRL as four recursive phases of learning (Winne & Hadwin, 1998; Winne, 2011). Each phase is embedded in a cognitive architecture model with the acronym COPES (see Figure 2.1). First, we describe four phases of the model, namely task definition, goal setting and planning, enacting study tactics, and adapting metacognition. In **Phase 1**, a learner creates a task definition by identifying resources and constraints necessary to complete the task. Based on this task definition, in the second **Phase**, the learner sets goals for working on the task and drafts plans to execute goals. In **Phase 3**, the learner applies tactics and strategies to reach goals as planned in **Phase 2**. In **Phase 4**, the learner changes strategies based on task performance after applying strategies.

The cognitive architecture describes each SRL phase within a student's condition, operation, product, evaluation, and standard (COPES). Conditions include resources and constraints those can impact the task. Conditions can be external (i.e., resources, time, instructional cues) or internal (i.e., beliefs, motivational factors). Operation include information manipulations during learning, such as searching, monitoring, assembling, rehearsing, and translating. Each operation results in products. For example, a product in **Phase 1** may be the definition of a task whereas a product in **Phase 3** may be understanding a science terminology after searching vocabulary. Standards include different criteria a student sets for learning goals (i.e. targeted time to complete a task). Through monitoring, the learner evaluates the standard of the product of the phase. In summary, Winne and Hadwin's model complements Pintrich's and others SRL models by specifying cognitive processes during learning (Greene & Azevedo, 2007)

To better understand SRL phases of Winne and Hadwin's model, we present an example. Suppose a student is working on a programming assignment that is due in a week. Here, the student defines the task at hand, **Phase 1**—a plan to submit a programming assignment by the due date. Based on their task definition, they set their goals **Phase 2**, i.e, reading lecture slides, writing pseudocode, selecting the algorithm to implement. In the **Phase 3**, the student executes their plans required for their task definition. This includes coding, debugging, testing, and submitting the assignment. Finally, the student may change their working style in the next assignment based on their assignment score and self-evaluation. This self-monitoring step is the **Phase 4**.

Winne and Hadwin's model describes SRL can be measured as *events* performed by students in **Phase 3** (Winne & Perry, 2000). Examples of measuring SRL as events are think aloud protocol, trace methodology in CBLEs, and performance observation of students. As CBLEs log students' trace data, researchers have widely used Winne and Hadwin's model to measure SRL in CBLEs (Panadero et al., 2016). As our study relies on AL log trace data, we adopt Winne and Hadwin's model to identify SRL.

Measuring SRL

SRL can be measured as an aptitude or as an event (Winne & Perry, 2000). An aptitude describes a student's behavior to predict future behavior. For example, if we know a student adopts study behavior depending on the material, then we can predict the student will study differently for a quiz and a term assignment. Common approaches to measuring SRL as an aptitude are self-report, questionnaires, structured interviews, and teachers' ratings.

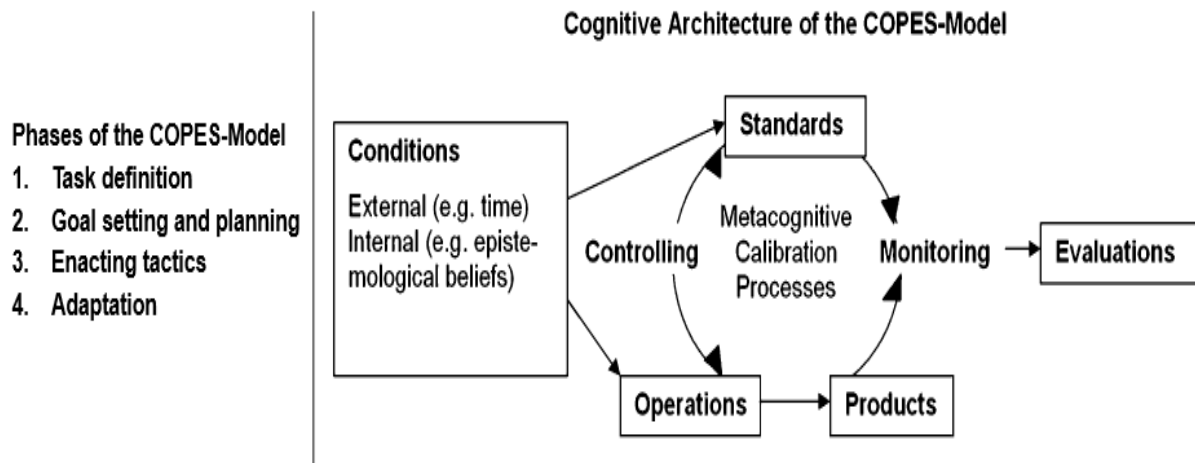


Figure 2.1: Winne and Hadwin's Model (COPES Model) (Pieschl et al. (2008))

Components used to measure SRL as an aptitude include knowledge about applying tactics, knowledge of self-parameters (e.g., interest, effort), understanding task difficulty (ease of learning, EOL), and others. Table 1 in (Winne & Perry, 2000) lists components of SRL as an aptitude. An event spans time with preceding and following events. For example, a student may take a note, highlight a text, and attempt an SA. Here, the highlighting event is preceded by a note-taking event and followed by an SA attempt. Common approaches to measuring SRL as events are think-aloud protocol (TAP), trace data, and performance observation.

Although self-report data are easy to administer, students can be inaccurate in self-report their SRL (Rovers et al., 2019; Zhou & Winne, 2012). In addition, self-reports restrict students to choose from limited options whereas the SRL process is dynamic in nature (Winne & Perry, 2000). In contrast, SRL measured by events, such as TAP, provide a much accurate measure of SRL (Bannert & Mengelkamp, 2008; Veenman, 2007). Researchers can code the frequency and task-specific strategy usage of an SRL (Azevedo, Moos, Johnson & Chauncey, 2010) from students' verbalization. Similarly, the (log) trace methodology captures fine-grained students' interaction within a CBLE. Thus, the trace methodology provides more accurate measures of SRL compared to questionnaires and self-reported SRL.

2.1.2 Reading and SRL

In this section, we discuss literature examining how reading and SRL strategy usage connect with academic performance.

Türkben investigated reading and metacognitive strategies in two public middle schools with one test (N =27) and one control (N =26) group (Türkben, 2019). Both groups completed a pre-test and a post-test. Only test group received SRL based strategic reading education interventions. The test group showed a significant difference in the post-test scores. Mason et al. studied low-performers' perception of reading (Mason et al., 2013) using the *Think before reading, think While reading, think After reading* (TWA) strategy as an SRL instruction strategy. The authors assigned 59 students randomly to two instructional groups (29 students to TWA, 30 students to (TWA + writing instruction), and 18 students to one no-treatment control group. They found that students who were assigned to instructional groups. 56 out of 58 students agreed that TWA helped them become better readers. Among them, 33 mentioned specific steps of the TWA strategy that they used before, during, and after reading texts, and 23 stated that TWA as a strategy helped them to become better readers. In both of these studies, researchers trained reading and SRL strategies to students. In our analysis, we utilize students' log trace data from AL as a proxy for reading and SRL usage.

Within the context of computer-based learning environments (CBLE), Yen et al. reviewed 55 studies from 1990 to 2016 to summarize the literature of metacognitive strategies in science reading. They performed content analysis to identify metacognitive knowledge (MK), metacognitive skills (MS), and metacognitive experience (ME) in science reading. Their analysis showed researchers assessed MK through questionnaires and mainly examined it with paper-based material. ME was assessed mostly by questionnaires and occasionally by interviews. Researchers assessed MS by tracing student behavioral data. This research shows that judgments of learning and comprehension induce metacognition—which enhances performance. This suggests an interaction between the measurement of ME and science performance.

Overall, previous research findings show that reading performance increases when adopting SRL strategies and contribute to higher learning gain and academic performance. Our study falls in SRL usage of the CBLE context. We do not have questionnaires or interview data. Our retrospective analysis falls within the category of MS, i.e., tracing students' log data to assess reading and SRL behaviors. In our study, we examine students' reading and reading-related SRL usage within the AL platform and how these strategies connect with

their academic performance.

2.1.3 SRL and CBLE

CBLEs have integrated features to help students fostering SRL skills. These systems vary in tracing and fostering SRL supports to students. In this section and Section 2.3, we focus on SRL strategies in CBLEs.

Crystal Island. *Crystal Island* is a game-based microbiology learning environment for middle school students. In the *Crystal Island* learning environment, a student attempts to diagnose and treat an infectious disease spread among a research team in a desert island. Students work with books and texts, talk to virtual characters, and test hypotheses. As students gather clues and talks with game characters, they need to organize thoughts, identify relevant information from conversation and revise hypotheses about possible explanations of the disease. Once they have identified the correct type of illness and propose a diagnosis, the game ends.

Within the *Crystal Island* platform, strategies taken by students scaffold SRL behaviors in *Crystal Island*, such as taking notes while reading and revising their hypothesis of the disease (Lester et al., 2013). Sabourin and colleagues measured real-time SRL behaviors considering three criteria (i) identification of SRL component, (ii) the cyclic nature, and (iii) instantiation of SRL (Sabourin, 2013). The authors adopted Zimmermann's model (Zimmerman, 2008a) to meet the first two criteria (p. 48, Sabourin (2013)). For the third criteria, authors relied on Winne's model (Winne & Perry, 2000) to identify event features. Event features are frequency and patterns of an SRL behavior, such as taking a note after taking to a medical representative (p. 64, Sabourin (2013)). The authors identified event features as a feature construction step for building prediction models and differential sequence mining. In one study, Sabourin and colleagues measured students' goal-setting and monitoring SRL strategies in *Crystal Island* with 260 eighth grade students' text-based responses. During interacting with the environment, students' affect data was collected by self-reported prompts from seven options: *anxious, bored, confused, curious, excited, focused, and frustrated*. After reporting students' affect state, the system prompted students to briefly describe their action plans. The texts were tagged for SRL (Sabourin, Shores, Mott & Lester, 2012) and students were categorized as High, Medium, and Low as per SRL strategy usage. The analysis showed High and Medium SRL student groups experience significantly better learning gain—measured by post-test score than Low group. High SRL students used more in-game resources, such as reading books and posters, and also took

more notes than the low-SRL student group. A similar work by the same author aimed at early prediction of students' SRL usage, particularly for Low-SRL group students to provide necessary scaffolding. They combined multiple prediction models to improve classification accuracy and prediction capability (Sabourin, Mott & Lester, 2012). Prediction results demonstrated significant improvements in both model's overall accuracy and recall for the low-SRL group compared to single prediction model. Our study is similar to Sabourin et al. considering middle school science SRL. Sabourin et al. identified real-time SRL in a game-based learning environment, *Crystal Island*, whereas our analysis relies on historical trace data within the AL system.

Betty's Brain. The *Betty's Brain* system helps middle school students develop SRL skills through learning-by-teaching model. Students apply SRL as they learn science and mathematics concepts by teaching a virtual agent, Betty. (Kinnebrew & Biswas, 2012; Kinnebrew et al., 2013). Activities within the system include READ: to locating required teaching materials, EDIT: to create links between science concepts, QUERY: to check their teaching by asking Betty questions, and EXPLAIN and QUIZ: to monitor Betty's progress by asking her to explain and take a quiz.

The system does not explicitly teach SRL but Betty's persona incorporates SRL strategies (Kinnebrew et al., 2010; Leelawong & Biswas, 2008). As the platform is open-ended in nature, students have to decide the best strategies to achieve their goal i.e., teach Betty. Two monitoring SRL strategies are present in the system: *checking*: when the student uses QUERY or QUIZ features to check and *probing*: when students ask Betty to EXPLAIN. Additionally, Betty can suggest refusing to take a quiz (Leelawong & Biswas, 2008).

In the *Betty's Brain* system, researchers have identified behavioral patterns (Kinnebrew & Biswas, 2012) and affective states (Munshi et al., 2018) that are indicative of high and low-performing students' SRL usage. Our study is similar to Kinnebrew et al. as we are focusing on SRL on middle school science. However, our scope is reading-related SRL whereas *Betty's Brain* fosters SRL through learning-by-teaching. *Betty's Brain* is an open-ended learning environment with choice-rich features. Students develop SRL skills by teaching Betty and observing her responses (i.e., refusing to take a quiz, analyzing Bettys' explanation) In AL, reading support features can act as proxies for students SRL usage.

MetaTutor. Another example of an SRL-fostering tutoring system is *MetaTutor*, a hyper-media environment for learning biology by Azevedo and colleagues (Bouchet et al., 2012; Azevedo et al., 2009). Azevedo and colleagues used a combination of cognitive SRL models (Azevedo, Johnson, Chauncey & Burkett, 2010) including Pintrich, Winne and Hadwin, and Zimmerman's in *MetaTutor* to analyze SRL. (Pintrich, 2000; Winne & Hadwin, 1998;

Zimmerman, 2008b). In the *MetaTutor* environment, the teacher or experimenter set up learning goals for students. An example of a learning goal is "Your task is to learn all you can about the circulatory system. Make sure you know about its components, how they work together, and how they support the healthy functioning of the human body (Azevedo, Johnson, Chauncey & Burkett, 2010)." Figure 2.2 presents a snapshot of MetaTutor's platform and SRL processes. A learning goal can have several subgoals (marked within C in Figure 2.2). Students can select processes within the system to monitor their SRL. Occasionally, a pedagogical agent may also prompt students to select question-specific SRL process. SRL processes in the system are: regulating reading behavior including selecting relevant materials to achieve subgoals, allocating time to read pages, re-reading and co-ordinating information, and annotating (Bouchet et al., 2012). The system also facilitates students to monitor their SRL behaviors. For example, the system prompts students to self-evaluate their judgement of learning (JOL), content evaluation of a page (CE), and feeling of knowing (FOK) (Bouchet et al., 2012; Azevedo, Johnson, Chauncey & Burkett, 2010). Additionally, students can also type-in to express their SRL and metacognitive process (i.e., summarize a paragraph they do not understand). The SRL and monitoring actions pallet of the system is shown in Figure 2.2 (marked with E).

Azevedo and colleagues have tested 38 different regulatory processes on *MetaTutor* including planning, monitoring, task defining, and others (Azevedo, Johnson, Chauncey & Burkett, 2010). Their findings show that scaffolding SRL can improve students' using these strategies and increase learning gains.

APLUS. Matsuda and colleagues investigated metacognition scaffolding in the context of the learning-by-teaching agent, *APLUS* (Matsuda et al., 2014). *APLUS* is an online learning environment where students teach an agent, SimStudent. Two types of metacognitive supports were implemented in *APLUS* to examine metacongitive scaffolding: the *quiz help* suggesting students when to quiz SimStudent and the *problem help* suggesting students which problems to select next. In a classroom study settings with 173 seventh-grade to ninth-grade students, their findings showed that students with metacognitive scaffolding performed better on the post-test score than the no-scaffolding group.

AutoTutor. Graesser, McNamara, and VanLehn developed *AutoTutor*, a conversational agent to support metacognition and explanation-centered learning (VanLehn et al., 1992). Explanation-centered learning is defined as learners' attempts to explain learning material and apply it to problem-solving (VanLehn et al., 1992). In *AutoTutor*, the agent assists students in constructing explanations of questions and clarifies misconceptions. *AutoTutor* simulates the human tutor by responding to students' question in a dialogue fashion (e.g.,

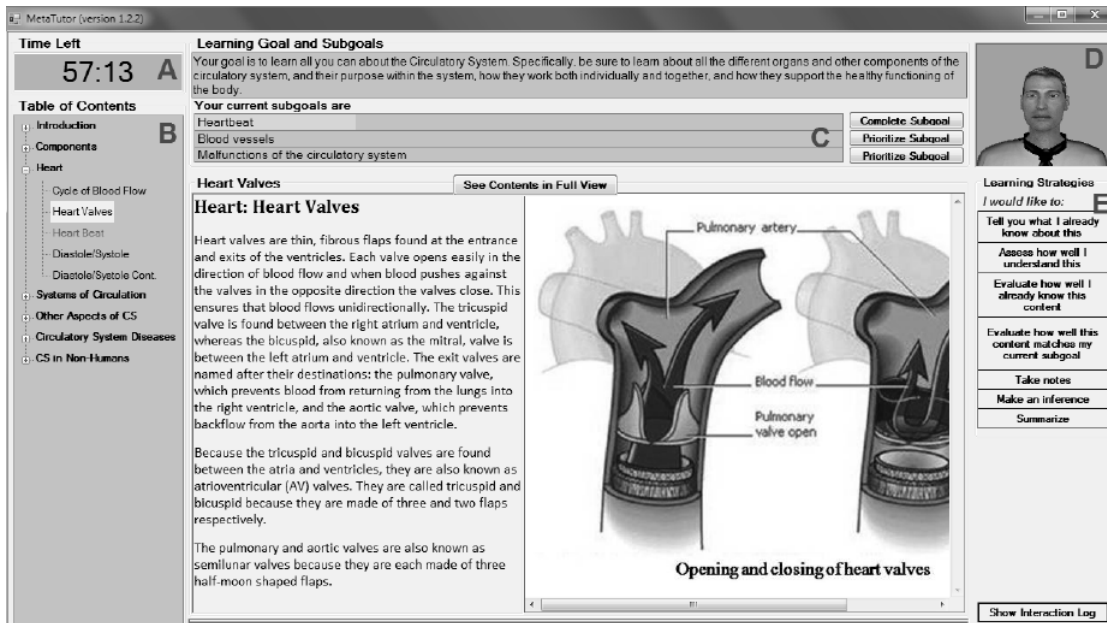


Figure 2.2: MetaTutor platform. (A: time remaining for task completion, B: table of contents, C: pallet for subgoals and progression, D: pedagogical agent, E: SRL and monitoring actions palette (Bouchet et al., 2012))

Tell me more, What else?). Studies have shown using *AutoTutor* yielded important learning gains (Graesser, Lu, Jackson, Mitchell, Ventura, Olney & Louwerse, 2004; Graesser et al., 2003).

To summarize, Crystal Island, Betty's Brain, MetaTutor, and Autotutor environments scaffold SRL by features available within systems. These systems do not train students SRL strategies but facilitate features fostering SRL. All four systems focus on science learning. The AL system is similar to these systems with respect to indirect SRL facilities. AL is a K-12 reading platform cataloging readings for Science, English Language Arts (ELA), and Social Studies. We focus on middle school physical science readings for our SRL analysis.

2.2 SRL in Learning Management Systems

In this section, we present literature review about SRL usage in Massive open online courses (MOOCs) and learning management systems (LMS). **MOOCs**. Lee and colleagues (Lee et al., 2019) conducted a systematic literature review on SRL strategies in MOOC with 21 papers published from 2014 to 2017. SRL methodologies analyzed fell into three categories: (i) motivational regulation strategies, (ii) cognitive and metacognitive monitoring strategies,

and (iii) behavioral and contextual regulation strategies.

Within motivational regulation strategies, five studies analyzed self-efficacy as an SRL measurement strategy. Self-efficacy is the measure of one's belief about completing an academic task (Pintrich, 1999). Littlejohn et al. (Littlejohn et al., 2016), revealed that MOOC participants who worked as data professionals, both high and low SRL behavior had high self-efficacy in the SRL questionnaire. Two studies found that self-efficacy was correlated with task familiarity. Within cognitive and metacognitive monitoring categories, five studies reported goal setting as an SRL measurement strategy. The study by Littlejohn et al. (Littlejohn et al., 2016), showed data scientists with high SRL had specific goals when enrolling in MOOC courses, such as improving knowledge and skills for professional development. In contrast, learners with low SRL levels mentioned general goals, such as getting a certificate. Two studies showed learners with specific goals participated longer in MOOC courses and reviewed course materials more. Within the behavioral and contextual regulation category, studies analyzed help-seeking, time management, and effort regulation as SRL strategies. Help-seeking by participating in discussion forums had mixed reports in MOOCs settings.

Poor time management and time conflict were causes for student dropout in MOOCs. In contrast, students who completed MOOCs, had good time management, reported by the questionnaire.

LMS. Araka et. al (Araka et al., 2020) analyzed 30 papers from 2008 to 2018 related to SRL learning strategies in e-learning platforms. Out of 30 studies, 10 were performed on LMS, 8 studies on MOOCs, and the rest 12 on personal learning environments (PLE). Studies used LMS include different types of courses, such as college level courses (Arnold & Pistilli, 2012; Cicchinelli et al., 2018; Hashemyolia et al., 2014), distant education class using Blackboard (Cho & Shen, 2013), and third-year medicine studies (Gaupp et al., 2018). To measure SRL, 16 studies used questionnaires, and 9 used log data-trace data from systems. The authors found 8 studies describing SRL measurement as well as SRL intervention strategies. Authors categorized into three groups: (i) learning analytics and dashboard visualization, (ii) animated agents, and (iii) prompts.

Their analysis showed that LA had been used to measure SRL as well as feedback to students by a dashboard. Learners showed better self-regulation when they followed the agents' suggestions.

2.3 Reading Platforms

nStudy. *nStudy* is a web-based reading platform designed by Beaudoin and Winne (Beaudoin & Winne, 2009). The platform is designed primarily for personalized learning. However, students can also interact with peers using the platform's collaborative features, such as workspaces for group study of reading or projects. Figure 2.3 presents an interface of the *nStudy* platform. The platform's reading support features help students to develop SRL skills. One such feature allows them to organize readings by defining *Learning Objectives* (Beaudoin & Winne, 2009). A student can link a text from the web and use annotation to define *Learning Objectives*. After linking the text to the interface, there are several fields to fill up, including *Importance*, *Difficulty*, and *Due Date*. *Importance* describes the importance of *Learning Objectives*, *Difficulty* describes perceived difficulty by the learner, and the *Due Date* field contains a timeline to achieve the goal (i.e., complete the *Learning Objectives*), respectively. In Figure 2.3, a student has created a note-form "*Learning Objective*" for Chapter 3 and described learning objectives as "*1. Master all the terms in this document; 2. Be able to draw out the hierarchy of control states; 3. Understand the conceptual structures of goals*". Thus, defining *Learning Objectives* facilitates students' practice of SRL skills (Beaudoin & Winne, 2009). Other features of *nStudy* that promote SRL are highlighting, bookmarking, tagging, and looking up definitions of terms.

Using the *nStudy* platform, Odilinye investigated students' tagging and highlighting behaviors and implemented a personalized recommendation system (Odilinye, 2019). The recommendation system was implemented as a plug-in. The initial articles were selected randomly. As students interact with *nStudy*, the system collected fine-grained interaction data related to reading and SRL, such as articles read, text marking, tags, and highlighting. The recommendation system requires at least one highlight to provide the student with a list of recommended articles. The recommendation system extracts keywords from highlighted texts to generate the article list using the latent Dirichlet allocation (LDA) method (Blei et al., 2003).

Odilinye conducted a user study with 49 undergraduate students to test the system. The task was: given two essay questions, students had to highlight texts from articles relevant to answers of questions. In the study, the author used two types of highlights: highlights *relevant to the two questions* and highlights for *recommending articles for further reading*. After marking a highlight, a pop-up prompted students to select types. Students were assigned randomly to either experimental or random recommendation group. The experimental recommendation recommends articles based on students' reading and SRL

interaction data. On the other hand, the random recommendation used a random number generator to assign articles. The results showed 93% of students in the experimental group preferred the highlighting based recommendation system.

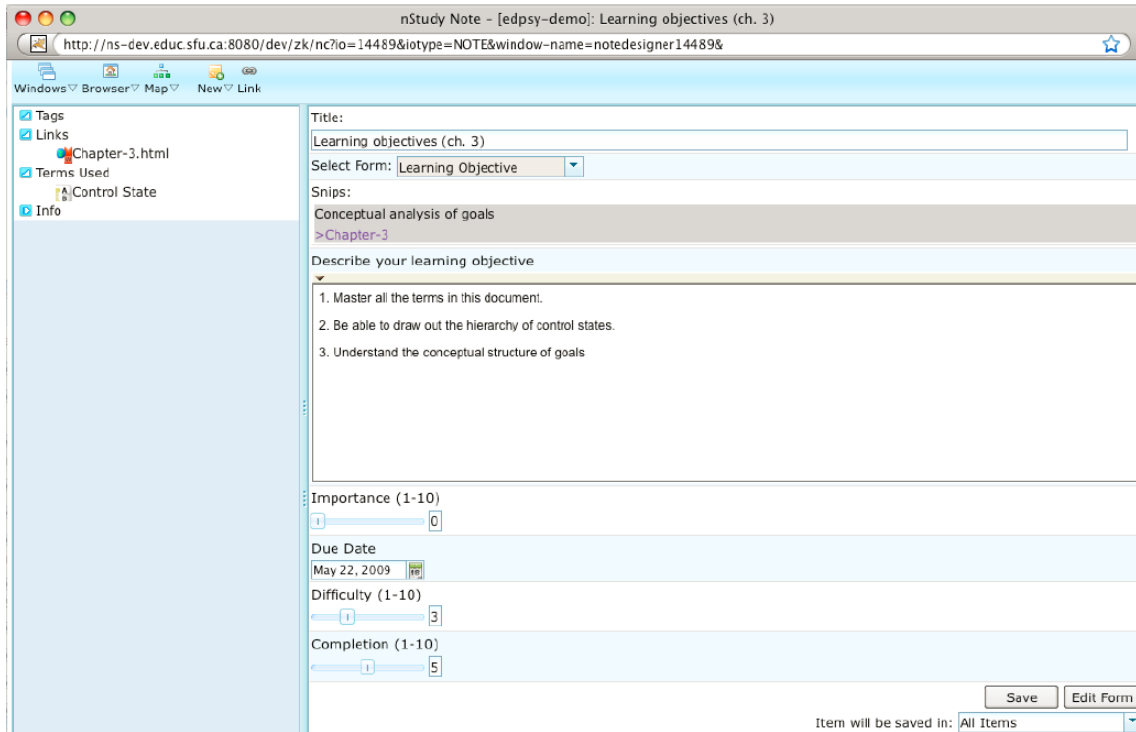


Figure 2.3: An interface of the *nstudy* platform for defining Learning Objectives (Beaudoin & Winne, 2009) .

iSTART. McNamara and colleagues designed a web-based system, *The Interactive Strategy Training for Active Reading and Thinking*, (*iSTART*), to train students reading strategies (Graesser & McNamara, 2010; McNamara et al., 2004, 2007), primarily targeting college students. Compared to *nStudy*, where system's features foster SRL, *iSTART* provides direct SRL training in the science domain using dialogues by animated agents (Graesser & McNamara, 2010). *iSTART* is built upon previous research of same authors showing that self-explanation enhances reading comprehension ability. With regards to SRL models, the system uses a combination of SRL models that are relevant to parts of the dialogue mechanism (Graesser & McNamara, 2010): Pintrich's model (Pintrich, 2000), Winne and Hadwin's model (Winne & Hadwin, 1998; Winne, 2011), Zimmerman's model (Schunk & Zimmerman, 2012), and Azevedo and colleagues hybrid model (Azevedo & Cromley, 2004).

iSTART provides training on five reading strategies to enhance self-explanation and SRL: (i) *comprehension monitoring*: being able to understand reading material, (ii) *paraphrasing*: describing the text in own words, (iii) *elaboration*: elaborating text using prior knowledge and paraphrasing, (iv) *prediction*: making a prediction of next content in text, and (v) *bridging*: linking different parts of the text by inference. The system provides reading comprehension training in three phases: *Introduction*, *Demonstration*, and *Practice*. The *Introduction* and *Demonstration* phases deliver video lectures covering self-explanation and five reading strategies. In the *Practice* phase, the agent reads a sentence and asks the student to type self-explanation. The agent assesses student's self-explanation, provides feedback, and asks students to resubmit if their explanation is unsatisfactory (Graesser & McNamara, 2010). The system measures SRL skills by the quality of students' self-explanation and the system's feedback.

To assess *iSTART*'s effectiveness, McNamara and her colleagues designed a study with 42 middle school students (McNamara et al., 2004). The task was to read and paraphrase a 13-sentence text about the thunderstorm. Students' paraphrased responses were by expert annotators and then compared to *iSTART*'s evaluation. Experts identified four categories: only paraphrase, an elaboration not directly related to text, an elaboration relevant to the target sentence, and an elaboration relevant to the global level of ideas presented in the text. Results showed *iSTART* can distinguish between paraphrase and elaboration responses at statistically significant level. Additionally, *iSTART* can distinguish between irrelevant, sentence level, and global level elaboration.

In another study, authors examined how low and high reading-ability students interact with *iSTART* and how those behavior are connected to outcome (Snow et al., 2014). The sample covered 40 high school students and *iSTART* with a game-based learning environment, *iSTART-ME*. Comparing with the *iSTART* platform, the *iSTART-ME* platform provides four types of extra practice selections relevant to game-based learning. Students' interaction data were logged from four types of practice selections: generative practice, identification mini-games, personalizable features, and achievements screens. Generative practice facilitates a student to practice self-explanation tasks for a given text. Identification min-games included six types of games, each presented an example text and corresponding self-explanation. Students were required to identify the strategy used in the example. Example of personalizable features included students' changing background, changing pedagogical agents, and editing avatars. Achievement screen included earning points and winning trophies in the game environment.

Researchers categorized high and low ability readers with 48 questions following the

Gates- MacGinitie Reading Test (Tests, 1978). Students participated in an 11-session experiments: one pretest session, eight training sessions, one post-test session, and one delayed retention test session held one week after the post-test. Each session was a 1-hour duration. The pretest, post-test, and delayed retention test each consisted of a self-explanation task of a given science text. A repeated ANOVA result showed students' self-explanation ability improved significantly from pretest ($M = 1.52$, $SD = .62$) to post-test ($M = 2.11$, $SD = .63$). Moreover, log data analysis showed students with low reading ability improved at a statistically significant level in self-explanation tasks over time and across sessions.

ReaderBench. Another multi-lingual platform is *ReaderBench* providing support to teachers and students with reading and writing facilities (Dascalu et al., 2013, 2015). *ReaderBench* facilitates i) text complexity analysis, ii) fostering SRL by automatic reading comprehension strategy identification, and iii) collaborative learning by chats and forums. Text complexity uses lexical indicators as word frequency, to syntactic and semantic levels (e.g., textual cohesion). Teachers can use text-complexity analysis to assign appropriate reading to the class as per students reading comprehension strategy. For reading strategy identification, the platform automatically identifies five types of strategies defined by McNamara et al. (McNamara et al., 2007): monitoring, causality, bridging, paraphrasing, and elaboration (Dascalu et al., 2014). *ReaderBench* helps students to improve SRL by automatic identification of reading comprehension strategies from their self-explanation (Dascalu et al., 2013).

In summary, *nStudy* and *ReaderBench* facilitate students' practice of SRL in reading whereas *iSTART* provides direct SRL training on reading. The AL system is similar to *nStudy* and *ReaderBench* in facilitating SRL by platform features.

2.4 Reading Literature in Educational Data Mining

In recent years, researchers in the Educational Data Mining (EDM) community have focused on students' reading behavior alongside SRL researchers. This line of research falls into two categories: incorporating reading into online textbook and student modelling. Below we discuss reading literature in EDM community.

2.4.1 Reading Speed and Student Modelling

In the context of student modelling, Eagle et al. incorporated reading rate in student modelling (Eagle et al., 2016). This study highlighted the positive effect of students' reading

rate to provide individual feedback. They extracted 12 features related to extracted reading time and text page revisited features into Bayesian Knowledge Tracing (BKT). Huang et al. modeled students' knowledge using textbook-based navigation (Huang et al., 2016). They made two assumptions: first, as students read a document, their knowledge associated with the document knowledge component (KC) will grow. Second, a student will spend less time on or entirely skip the document if it contains KCs that the student has already learned. In contrast, the student will read the document more carefully if it contains new KCs. However, the model's assumptions may not be applicable for all students. Thaker et al. addressed the limitations of Huang et al.'s work by incorporating individual student differences. (Thaker et al., 2018). According to the authors, students' reading speed varies because of individual motivations, reading proficiency, and other factors. They evaluated their model on 22,536 interactions from 22 students from a graduate-level Information Retrieval Course. Their proposed model learned a more accurate student knowledge state than Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994). Another work by the same authors incorporated reading rate in Performance Factor Analysis (PFA) model (Pavlik et al., 2009*b*). They proposed two models: one incorporating reading event and another reading rate evaluated their model with 777 students in a MOOC course and found the reading rate model performed better compared to PFA (Thaker et al., 2019).

Our RQ1 includes predictive student modelling taking students' reading-related SRL behavior. In particular, we predictive model extends the deep knowledge tracing (DKT) (Piech et al., 2015) by taking into account annotating, highlighting, and vocabulary behavior. Our predictive model differs from DKT in terms of scope—we limit students' SRL behavior within an assignment to predict a questions' score.

CHAPTER

3

ACTIVELY LEARN PLATFORM

3.1 The Actively Learn Platform

In this section we describe the Actively Learn (AL) ¹ platform and reading-related SRL in the AL.

3.1.1 The Actively Learn Platform

AL is a digital literacy platform aimed at students in primary education (K-12). AL is designed to improve students' reading proficiency. The platform allows teachers to assign readings to a class as assignments. Assignments in the AL platform can range from one page to multiple pages and have embedded questions. Questions in AL can be multiple choice (MCQ) and short answer (SA) questions, including free texts and fill in the blanks. Teachers may use predefined reading texts and questions of AL or introduce their own as assignments. MCQs are automatically graded whereas SAs are not. AL questions are graded on a scale of zero to four. Physical science reading texts in the AL platform are organized following the Next

¹<https://www.activelylearn.com/>

Generation Science Standards (NGSS) ² guidelines. Figure 3.1 shows a reading text on the AL interface.

3.1.2 SRL in AL

Our focus of this study is evaluating students' SRL usage in middle school science reading within the AL platform. For cross-domain comparison, we evaluated SRL usage in social studies assignments. Within our study, we adopt Winne and Hadwin's four-phase model of COPES model of SRL (Winne & Hadwin, 1998). First we describe measuring SRL as events, outlined by Winne and Hadwin (Winne & Perry, 2000) and then we describe how we adopt this model to analyze SRL.

Measuring SRL as Events

Winne and Hadwin's model operates in four phases described in Section 2.1.1. As stated by Winne and Perry (Winne & Perry, 2000), this model can be used to measure SRL as an *aptitude* or as an *event*. We discussed measuring SRL as an aptitude in Section 2.1.1. In this Section, we focus in detail on measuring SRL as an *event* in the AL system.

Winne describes three properties of SRL as an event: *occurrence*, *contingency*, and *patterned contingency*. *Occurrence* describes an event when SRL is taking place. It describes as an event transition from a state where the SRL was absent to a state SRL is present. An example of SRL occurrence is a student taking a note. *Contingency* refers to a conditional *if-then* relationship among events involving SRL. An example of SRL as contingency event is *if* a student attempts an SA, *then* they will highlight. A *patterned contingency* refers a combination of *if-then* relationships to solve a task. For example, *if* a student attempts an SA, *then* they will highlight, take notes, and look up vocabularies.

For measuring SRL in RQ1, RQ3 and RQ4, we consider SRL as *occurrence* events and in RQ2, we consider SRL as *contingency* and *patterned contingency* events.

Coding SRL in AL

To identify reading-support features of AL as SRL, we consult previous literature of SRL. For the first study, we considered four AL features as SRL (re-reading, annotating, highlighting, and vocabulary lookups) and for the remaining ones, we considered three. We identify three reading support features of AL as SRL: annotating (Makany et al., 2009), highlighting

²<https://www.nextgenscience.org/>

Inertia

What Is Inertia?

Inertia is the tendency of an object to resist a change in its motion. All objects have inertia, whether they are stationary or moving. Inertia explains **Newton's first law of motion**, which states that an object at rest will remain at rest and an object in motion will stay in motion unless it is acted on by an unbalanced force.

QUESTION 1 DOK 1 STANDARD RST.2 MS-PS2-2



How does inertia explain why it is difficult to stop a moving skateboard?

- Due to inertia, it may take as much effort to stop a moving skateboard as it did to start the moving skateboard.
- A moving skateboard, due to inertia, travels extremely quickly, making it very difficult to stop.
- Inertia affects the skateboard's direction, making it hard to stop it from moving.
- A skateboard is affected by various forces that act upon it all at once, allowing inertia to play a key role in its movement.

QUESTION 3 DOK 2 STANDARD RST.1 MS-PS2-2



Why don't moving objects on Earth continue in that direction forever, as Newton's first law states?

Figure 3.1: A reading text and embedded questions. Question 1 is an MCQ and question 3 is an SA.

(Winne et al., 2017), and vocabulary lookups (Biemiller & Slonim, 2001), as we believe these features serve as proxies for SRL behaviors. Science text involves domain-specific terms and vocabulary. Vocabulary lookups help students to understand concepts when they come across new vocabulary. Prior research showed that vocabulary acquisition is highly predictive for reading comprehension (Biemiller & Slonim, 2001). Annotating requires students to comprehend text and write down in their own words (Makany et al., 2009). Azevedo described taking notes, summarization, and reading notes in the context of SRL strategies for science learning with hypermedia (Azevedo, 2008). To select texts for highlighting, students monitor information and connect those to their prior knowledge (Winne et al., 2017).

3.1.3 Comparison with Other Reading Platforms

In this section, we compare SRL features and scope of AL with other reading platforms described in Section 2.3.

Comparing with the *nStudy* (Beaudoin & Winne, 2009) platform, AL catalogs curriculum-integrated reading whereas in the *nStudy*, a user can link in a text from the web for reading. *nStudy* is primarily designed for personalized reading. The platform supports collaborative features for group readings or projects. On the other hand, teachers use the AL in K-12 classrooms for in-class and homework assignments for science, English, and social studies. SRL in AL and *nStudy* are indirect in nature, i.e., platforms' features foster SRL. SRL supporting features in the *nStudy* are highlighting, bookmarking, annotating, goal-setting, and self-reflection features, including defining a *Learning Objective*, perceived difficulty, and deadline (Section 2.3). SRL features in the AL are reading-aid features: annotating, highlighting, and vocabulary-lookups.

Comparing AL with *iSTART*, *iSTART* is designed primarily targeting high school students and adult literacy (Dascalu et al., 2015). Comparing AL with *ReaderBench*, *ReaderBench* provides SRL support by automatic feedback to students self-explanation and reading strategies defined by McNamara et al. (McNamara et al., 2004). The *iSTART* differs from AL in nature of SRL support. *iSTART* provides explicit reading comprehension strategies training and feedback on students self-explanations to develop SRL.

The reading events in the AL logs total time spent on an assignment. When comparing reading events in *MetaTutor*, the system logs pages read by students and researchers coded relevant and irrelevant page-readings as SRL (Bouchet et al., 2012).

3.2 Dataset Construction

Studies of this thesis focus on middle school science reading assignments in the AL platform. The dataset covers student records who completed assignments in 2018. Table 3.1 shows our dataset filtering criteria for physical science dataset. We applied similar filtering criteria to construct the social studies dataset in RQ1.2 and RQ2 .

Table 3.1: AL Dataset Construction: Science

Criteria	Description
Initial	17,886 students in 1,033 classes
Filter 1	
Class size between 10 and 60	14,925 students
Removing duplicate entries	14,575 students
Filter 2	
Exclude ungraded students' SA responses (missing grade SAs was replaced by zero in RQ1.1)	12,548 students and 940 assignments

The initial dataset of RQ1.1 covered 69 middle school physical science readings, used in 1,033 classes by 17,886 user accounts. After examining the data we opted to exclude classes containing more than 60 or fewer than 10 participants from our analysis. This left 83.45% of the students; we also excluded any participant that was enrolled in multiple classes (an additional 1.95% students), as we believed that these accounts may have been used by teachers or for testing purposes. This resulted in a final sample size of 14,575 unique users in 701 classes with 1,003 assignments.

We applied similar filtering criteria to filter 1,151 social study assignments used in 1,045 classes by 19,993 user accounts. After examining the data we opted to exclude classes containing more than 75 or fewer than 10 participants from our analysis. This left us with 18,086 students (90.46% of the students). We excluded answers to ungraded SA questions. This resulted in a final sample size of 16,240 unique users in 781 classes with 857 assignments. Mean SAs per assignment is 18 and MCQ is 2.05. For science dataset, we also excluded students' SA responses but missing grades (were replaced by zero in RQ1.1). The resulting physical science dataset had 12,548 unique students and 940 assignments.

CHAPTER

4

SRL AND QUESTION FEATURES

In this Chapter 4, we will describe methodology and results of RQ1. For our analysis, we split RQ1 into following three sub-research questions

RQ1.1: Does the association of SRL vary depending on question formats?

RQ1.2 How do SA response text associate with question performance?

RQ1.3 How can we design predictive student modelling taking the temporal ordering of question and SRLs textual representation?

In RQ1.1, we discuss the association between SRL activities and question scores for two question formats, such as MCQ and SA. For our RQ1.1 analysis, we counted SRL actions and applied hierarchical linear models to measure association. In RQ1.2, we discuss students' responses to SA to understand how SA responses are connected with scores. In RQ1.3, we focused on *what* SRL actions students exhibited and *how* those SRL actions were connected to question score. Contrary to RQ1.1 where we simply counted SRL actions, here we examined the *textual content* of highlighting, annotating, and vocabulary lookup. Moreover, to measure how SRL activities may impact students' scores, we examined (i) similarity measures between SRL actions and question texts and (ii) temporal ordering of SRL actions and question attempts.

In this Chapter, first, we will discuss the background and related literature review of three sub-RQs in Section 4.1. Next, we will discuss methodology and results of RQ1.1 in Section 4.2, RQ1.2 in Section 4.3, and RQ1.3 in Section 4.4. Finally, in Section 4.5, we discuss how RQ1 results connect to the broader goal of this thesis and conclude this Chapter.

4.1 Background

4.1.1 Question Format

O'Neil and Brown (O'Neil Jr & Brown, 1998) investigated the effect of question format and SRL usage with 1,032 8-th grade students from 59 classes in Mathematics. Authors measured SRL behavior with two question formats: MCQ and open-ended for males and females with different ethnicity groups. The study comprised seven MCQs and two open-ended questions. Students were randomly assigned to MCQ or SA question assessment. Considering scores, the authors did not find any gender difference with respect to question formats. However, when considering ethnicity, they found Asian and White students scored higher in open-ended questions compared to Latino and African American students. Considering SRL behavior, authors found open-ended questions prompted more metacognitive processes, including cognitive strategies and self-checking. Females showed more SRL behavior than males. However, the authors did not find a significant correlation between SRL usage and ethnicity.

O'Reilly and colleagues (O'Reilly et al., 2004) examined students' reading comprehension with two types of questions: text-based questions and bridging questions within the iSTART reading platform. The iSTART platform assesses students' SRL skills from their paraphrased responses. Text-based questions require students to find answers from the reading text. This question type was introduced to assess students' basic reading skills. Answers to bridging questions can span multiple sentences in the text and require understanding connections between those sentences. Bridging questions were introduced to assess students' deeper level of reading comprehension. Thus, bridging questions required more SRL skills. The study with 1000 students in 35 high-school classrooms found students performed better in text-based questions compared to bridging ones. Our work is similar to previous work in that it focuses on question features. We also examine students' SA response text and its association with students' performance.

4.1.2 Predictive Student Modelling

The literature of predictive student modelling can be categorized into two broader categories: cognitive diagnostic models and deep learning-based models. We discuss relevant literature below.

Cognitive Diagnostic and Knowledge Tracing.

Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994) uses students' historical exercise data to estimate the probability that a student has mastered a specific skill. A student's knowledge state is represented as a set of binary variables. Each variable represents a skill. However, the original BKT algorithm does not take into account student-specific behavior. Researchers have proposed individualized BKT by taking account of the question difficulty model per student (Pardos & Heffernan, 2011; Yudelson et al., 2013). Student-specific parameters have also been used in the cognitive science domain. Item response theory (IRT) (Hambleton et al., 1991) uses students' historical performance to estimate their ability. The simplest form of IRT is 1-PL Rasch model (Rasch, 1960). The Rasch model is a logistic model in that it evaluates the latent ability level needed to get a 50% chance of getting a question right. The descendent of Rasch models are learning factor analysis (LFA) (Cen et al., 2006) and performance factor analysis (PFA) (Pavlik et al., 2009a). All these methods use student mastery or proficiency metrics to predict students' performance.

Deep Learning-based Knowledge Tracing. Inspired by deep neural models' success, Piech et. al proposed deep knowledge tracing (DKT) (Piech et al., 2015) in 2015. Since then, researchers have proposed variations of DKT methods. Among them, Nagatani and colleagues incorporated forgetting behavior, considering elapsed time between events (Nogatani et al., 2019). One limitation to the original DKT was it did not consider skills associated with each question. Zhang and colleagues (Zhang et al., 2017) improved the DKT to estimate students' latent knowledge state. Specifically, their model utilizes Dynamic Key-Value Memory Networks (DKVMN) to capture the skill-set associated with questions. Ai and colleagues also utilized DKVMN to capture concepts in a course and perform knowledge tracing (Ai et al., 2019).

Pandey et. al further proposed *attention-based knowledge tracing* (Pandey & Karypis, 2019; Pandey & Srivastava, 2020). In attention-based models, recurrent neural networks are no longer needed. Instead, positional embedding is used to track sequential behavior. The advantage of an attention-based model is its explanatory capability. Ghosh et. al (Ghosh et al., 2020) proposed time-decaying monotonic attention capturing the importance of previous question-solving behavior. Ghosh et. all argue that when students face a question,

past question-solving behavior from unrelated concepts or that isn't recent may not be relevant.

Su et. al (Su et al., 2018) proposed exercise-enhanced deep knowledge tracing. According to the authors, exercise texts may semantically represent underlying knowledge concepts. An extended version of the study was done by the same author proposing EKT (Liu et al., 2019). Our study is similar to Su et al., as we take into account question text to measure students' performance. As we do not have questions-to-skillset mapping, we took the embeddings question-text, similar to Pandey et. al (Pandey & Karypis, 2019) and Su et. al (Su et al., 2018). As stated by Su et. al, the learned question embedding captures the characteristics of each question. This approach does not require manual labeling of KC. Pandey et. al and Su et. al's predictive modelling took only previous question answering attempts to predict the score of a question. In contrast, our predictive modelling uses text embedding to capture the characteristics of question and reading-related activities, such as highlighting, annotating, and vocabulary lookup. We used an LSTM-based approach with an attention mechanism to capture students' sequential behavior within the AL system.

Sequential Student Behavior. Previous researchers applied sequence mining techniques in identifying learners' behavior, and predictive student modelling (Geden et al., 2020; Min et al., 2016). Gaden et. al similarly applied multi-task learning to capture sequential behavior within a game-based learning environment, Crystal Island (Geden et al., 2020). In the Crystal Island environment, students can engage with eight action types. The authors designed an LSTM to encode students' action types and corresponding action properties at each timestep to predict the students' post-test scores.

Our study is also similar to sequential student modelling (Geden et al., 2020; Min et al., 2016). We model students' action type at each timestamp to predict performance on a question. However, Gaden et. al considered the entire game-sequence behavior and predicted performance after finishing the game. Our predictive modelling differs from Gaden et. al as we adopt a window-based approach (Choffin et al., 2019) to take sequential actions before each question attempt.

4.2 SRL and Question Features

In this section we present methodology and results of *RQ1.1 Does the association of SRL vary depending on question formats?*

4.2.1 Methodology of RQ1.1

We first describe hierarchical linear model (HLM) and then describe our modelling.

Correlation Analysis

First we calculated correlation between assignment score and each SRL variable. We assessed three types of scores as assignment performance: total assignment score, MCQ score in assignment, and SA score in assignment. We applied non parametric Spearman test Spearman (1961) to compute correlation.

Hierarchical Linear Model

Hierarchical linear model (HLM) is a form of regression to analyze data with nested levels (Woltman et al., 2012). An example of nested level data is students nested within classrooms nested within school. Prior introducing HLM, researchers used fixed parameter regressions i.e., ordinary least square regression (OLS); ignoring variance among the nested data levels. HLM allows random co-efficient regression analysis for nested data (Snijders & Bosker, 2011). In HLM, level one (L1) variables are individual level variables. Group level variables are denoted by level two (L2) and higher levels. In the above example, students are level one (L1) variable, classes are level two (L2), and schools are level three (L3) variables.

The simple form of HLM is a random intercept model (Snijders & Bosker, 2011). Let i denoting L1 and j denoting L2 variables. The OLS model with outcome variable Y_{ij} and independent variable x_{ij} is

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + R_{ij} \quad (4.1)$$

Group-dependent regression model, (HLM), is

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + R_{ij} \quad (4.2)$$

In equation 4.1, OLS uses the same intercept, β_0 , for nested data. In equation 4.2, the intercept β_{0j} is group dependent. In other words, the intercept varies among groups in a random intercept model. Graphically, these random intercepts are parallel lines (see Figure 41. in Snijders & Bosker (2011)).

Modelling

Table 4.1 presents descriptive variables of our dataset.

Table 4.1: Mean and (Standard Deviation) of Descriptive Variables

Variable	Mean (SD) (n = 14,575)
Questions Per Assignment	7.092 (7.712)
MCQ Per Assignment	2.52 (2.53)
SA Per Assignment	4.19 (5.96)
Total Assignment Score	10 (10.37)
MCQ Score in Assignment	5.831 (8.474)
SA Score in Assignment	4.170 (6.583)

We used HLMs to model the relationship between observed behaviors and students assignment performance. We assessed three types of scores as assignment, as done for correlation analysis. We built three HLMs for the abovementioned three scores as response variables. We modelled assignment at level one (L1), nested within students (level two, L2), nested within classes (level three, L3). The fixed-effect variables were SRL features and number of questions in assignment; these variables were at *Level 1*. Assignment, student, and class were all modeled as random intercepts.

4.2.2 Results of RQ1.1

Results of Correlation Analysis

Table 4.2 presents correlation analysis results of each SRL variable. Considering total score nad SA score, all SRL variables are statistically significant and have positive correlation with score. However, for MCQ score, re-reading has negative correlation with score.

Results of HLM

Table 4.3 displays the results of three models. We report standardized effect size using the formula $\beta = (B * SD_x) / SD_y$ (see e.g., Rutherford et al. (2017)). All SRL-related variables had positive and statistically significant associations with total science score in assignment. Re-reading had the highest predictive power ($B = 1.667$, $\beta = 0.128$, $p < 0.001$), followed by note taking ($B = 0.582$, $\beta = 0.055$, $p < 0.001$), highlighting ($B = 0.492$, $\beta = 0.028$, $p < 0.001$),

Table 4.2: Results from Spearman correlation measuring association between SRL and science score

<i>LI (Assignment) Level</i>	<i>Correlation Value</i>	<i>p</i>
Total Score		
No. Times Reread	0.16	<0.001
No. Notes	0.16	<0.001
No. Highlights	0.06	<0.001
No. Vocabulary Lookups	0.13	<0.001
MCQ Score		
No. Times Reread	-0.04	<0.001
No. Notes	0.10	<0.001
No. Highlights	0.04	<0.001
No. Vocabulary Lookups	0.01	0.04
SA Score		
No. Times Reread	0.28	<0.001
No. Notes	0.08	<0.001
No. Highlights	0.06	<0.001
No. Vocabulary Lookups	0.14	<0.001

and vocabulary lookups ($B = 0.275$, $\beta = 0.021$, $p < 0.001$).

In order to address how predictive associations varied based on question format, we analyzed HLM results of two models: one with the MCQ score and another SA score as the dependent variable. All of the SRL strategies continued to be positive predictors of SA score with statistically significant associations; all but the vocabulary lookups continued to be statistically significant positive predictors of MCQ score. Standardized betas were also higher in SRL/SA associations.

4.2.3 Discussion of RQ1.1

In this section, we present our findings of RQ1.1.

Our findings show that reading-associated SRLs are linked with students assignment performance. Our findings are in-line with previous SRL researchers demonstrating reading-related SRLs leading to higher academic performance (Türkben, 2019). Considering question format and SRL, we found frequency of SRL was more predictive of SAs than MCQs. One possible explanation for this could be that when answering MCQ, students could leverage some partial knowledge to rule out one or more distractors. Thus, deeper processing, as is associated with SRL (Bliss, 1980), may be less necessary for answering MCQs.

Table 4.3: Results from HLM measuring association between SRL and science score

<i>L1 (Assignment) Level</i>	β	<i>B</i>	<i>SE</i>	<i>p</i>
Total Score				
Intercept		6.533	0.402	<0.001
No. Times Reread	0.128	1.667	0.069	<0.001
No. Notes	0.055	0.582	0.062	<0.001
No. Highlights	0.028	0.492	0.072	<0.001
No. Vocabulary Lookups	0.021	0.275	0.055	<0.001
MCQ Score				
Intercept		5.510	0.369	<0.001
No. Times Reread	0.032	0.345	0.041	<0.001
No. Notes	0.024	0.206	0.038	<0.001
No. Highlights	0.016	0.228	0.045	<0.001
No. Vocabulary Lookups	-0.003	-0.036	0.031	0.259
SA Score				
Intercept		1.699	0.232	<0.001
No. Times Reread	0.150	1.223	0.043	<0.001
No. Notes	0.040	0.271	0.038	<0.001
No. Highlights	0.019	0.210	0.043	<0.001
No. Vocabulary Lookups	0.036	0.289	0.035	<0.001

4.3 SA Response Analysis

In this section, we answer *RQ1.2 How do SA response text associate with question performance?* More specifically, we calculate the cosine similarity between question text and students response text. We examine (i) how cosine similarity vary between cross domain subjects and (ii) how cosine similarities are associated with question performance.

4.3.1 Methodology of RQ1.2

Dataset

We selected graded SA question responses from Section 3.2 dataset. The science dataset had total 32061 SA question submissions (unique SA questions = 1990) submitted by 6,245 students in 573 assignments. The social study dataset had 1,71,399 SA question submissions (10,948 unique questions) submitted by 12,865 students in 698 assignments.

Statistical Analysis

We calculated the cosine similarity between each SA text and corresponding response text. The value of cosine similarity ranges from -1 to 1.A -1 indicates entirely unrelated and 1

indicates highly related texts.

The cosine similarity statistics of science SA were (mean = 0.43, SD = 0.22) and social study (mean = 0.34, SD = 0.20). We conducted the non parametric Mann-Whitney U test (Mann & Whitney, 1947) to test whether the mean cosine similarities vary at a statistical significant level.

Score and Cosine similarity

To examine how the cosine similarities are related to question score, first we calculated correlation between the two variables. We applied non parametric Spearman correlation (Spearman, 1961) between cosine similarities and scores.

We also examined how cosine similarities are associated with SA scores taking class factors. For this purpose, we conducted HLM analysis on science and social study dataset. The independent variable was cosine similarity (fixed effect variable) and dependent variable was score. We modelled question tags as Level 1 (L1) variable, nested within assignments (L2) nested within students (L3) nested within class (L4).

4.3.2 Results of RQ1.2

4.3.3 Results of Statistical Analysis

The Mann-Whitney U test was statistically significant between mean cosine similarities between two domains ($p < 0.001$). We computed non parametric effect size (r), Cliff's-Delta (Cliff, 1993) between cosine similarities of two domains. The effect size was 0.241.

4.3.4 Results of Score and Cosine Similarity

The Spearman correlation (ρ value) between score and cosine similarity was statistically significant for both subject domains. The science was (ρ value = 0.18, $p < 0.001$) and social study (ρ value = 0.17, $p < 0.001$).

We report regression coefficient (B) and standardized effect size (β) of cosine similarities for HLM. The standardized effect size was greater in science than social study. For science HLM, values were ($B = 1.36$, $\beta = 0.25$, $p < 0.001$). For social study HLM, values were ($B = 0.78$, $\beta = 0.15$, $p < 0.001$).

Table 4.4: Results from HLM measuring association between cosine similarity (SA question text, response) and score

<i>L1 (Question ID) Level</i>	β	<i>B</i>	<i>SE</i>	<i>p</i>
Science				
Intercept		2.11	0.034	<0.001
Cosine Sim (SA Ques. text and resp)	0.25	1.36	0.01	<0.001
Social Study				
Intercept		2.64	0.021	<0.001
Cosine Sim (SA Ques. text and resp)	0.15	0.78	0.32	<0.001

4.4 Predictive Student Modelling & Temporal SRL Behavior

In this section, we answer *RQ1.3 How can we design predictive student modelling taking the temporal ordering of question and SRL's textual representation?*

The AL system logs students' actions within the system including reading behavior and question submission. For the purpose of our study, we aggregated student actions into a unified transaction log. Table 4.5 shows a hypothetical example of our log. The system logs student interaction data indexed by unique assignment ID, AId. The timestamp column logs actions in sorted order. Student actions are question answering and SRL activities, such as highlighting, annotating, and vocabulary lookup. In the above example, S101 highlighted once (at T1), took a note (at T2), looked up vocabulary once (at T4), and attempted three questions (at T3, T5, and T6, respectively). The 'Action Text' column contains the textual description of corresponding actions. For example, the action text at timestamp T1 includes the highlighted text of the student. Similarly, the action text at T5 contains the textual descriptions of the attempted questions.

Table 4.5: A hypothetical Student Action Log in AL

AId	Std.Id	Time	Action	Item Id	Action Text
123	S101	T1	Highlight	h101	<Highlighted Text>
123	S101	T2	Annotation	a101	<Note Text>
123	S101	T3	Question	q101	<Ques Text>
123	S101	T4	Vocab. Lookup	v101	<Vocab text>
123	S101	T5	Question	q102	<Ques. Text>
123	S101	T6	Question	q103	<Ques. Text>

We formulate our task as follows

PROBLEM DEFINITION *Given the log of students' actions and textual contents of each action, our goal is to predict students' question performance at time t considering previous actions within a window size, w .*

For our hypothetical example, assume our window size, $w = 5$. Thus, to predict S101's performance on question q103 at T6, we will consider actions from T1-T5.

Scope. Our study differs from prior work on DKT in terms of scope. In the DKT and its variations, students' question answering sequences are typically longer than our study (65.9 questions per student Su et al. (2018)). Questions are mapped to concepts. Question performance measures are generally correct or incorrect attempts. In contrast, the scope of our study is within a reading-comprehension assignment. Thus, the students' activity sequence length is much smaller compared to DKT.

Dataset

We selected a subset of Science dataset from Section 3.2. We selected students who have at least one SRL activity (i.e., highlighting, annotating, and vocabulary lookup) in an assignment. Next, we prepared our dataset as described in Table 4.5. This resulted in 13,066 samples. The total number of unique students in the dataset is 1,796 and the total number of unique assignments is 425. The descriptive statistics of number of student actions prior a question attempt is (Mean = 6.47, Median = 5.00, and standard deviation (SD) = 5.37).

4.4.1 Methodology of RQ1.3

In this Section, we describe our methodology. Figure 4.1 presents components of our model architecture. We describe our model's components below.

Input. We encode action texts to a fixed dimensional action feature vector. The action feature vector, \mathbf{X} is comprised of following parts:

- *Text Embedding* We used USE to represent each action text into a $d = 512$ dimensional vector, \mathbf{E} .

$$\mathbf{E} = USE(\text{ActionText}) \in \mathbb{R}^d \quad (4.3)$$

- *Action Type Encoding.* We used one-hot encoding to encode four different action types.

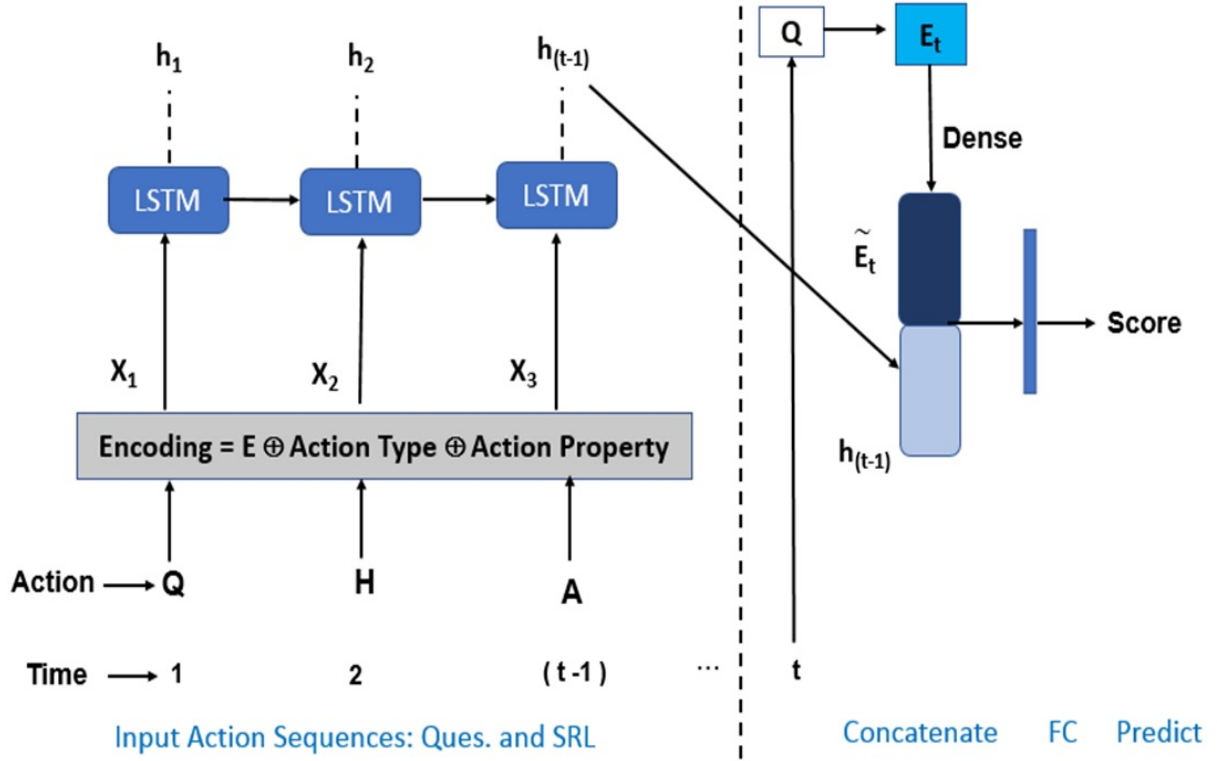


Figure 4.1: Proposed model. Q = Question Attempt, H = Highlighting, A = Annotation.

- *Action Property Encoding.* We encode two additional features: a boolean feature indicating whether the action is a question attempt or SRL feature. This feature will distinguish a question attempt from SRL actions. The second feature is a question score feature. This field contains a score of if the action a question attempts.

We concatenate the above-mentioned features to form an input action vector, \mathbf{X} .

Sequential Action Modelling. We used long short-term memory (LSTM) Hochreiter & Schmidhuber (1997) to encode students' sequential actions within the time window w . Let, $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_{t-1}$ represents sequential actions upto a question submission at time t .

$$\mathbf{h}_{(t-1)} = LSTM(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_{(t-1)}) \quad (4.4)$$

where \mathbf{h} is the dimension of the hidden layer of the LSTM.

LSTM models are designed to process sequential data using a series of input, output, and forgetting gates. Each gate in an LSTM model uses a sigmoid function to select feature values. A value of zero means that the gate will block everything, and one means the gate allowing all features to the next step. The equations of LSTM gates are

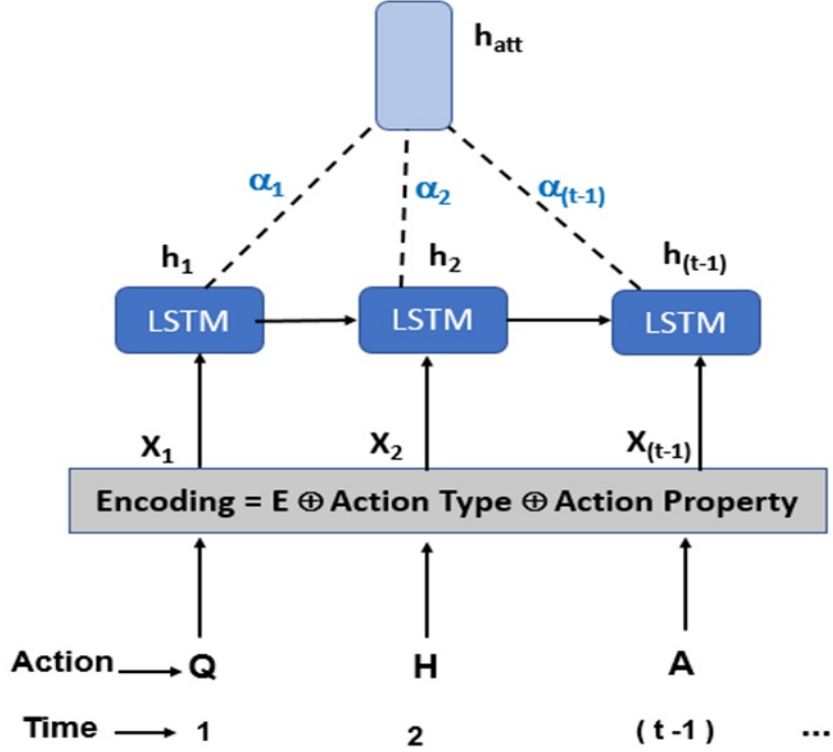


Figure 4.2: LSTM with Attention

$$\begin{aligned}
 i^t &= \sigma(W_h^i h^{t-1} + W_x^i x^t) \\
 f^t &= \sigma(W_h^f h^{t-1} + W_x^f x^t) \\
 o^t &= \sigma(W_h^o h^{t-1} + W_x^o x^t) \\
 \tilde{c}^t &= \tanh(W_h^c h^{t-1} + W_x^c x^t) \\
 c^t &= f^t \oplus c^{(t-1)} + i^t \oplus \tilde{c}^t \\
 h^t &= o^t \oplus \tanh(c^t)
 \end{aligned} \tag{4.5}$$

The cell state serves as memory block for LSTM. The cell state takes contents from most recent cell state, $c^{(t-1)}$ and hidden state, $h^{(t-1)}$, erases previous information through the forgetting gate, f^t , adds new information from the input gate, i^t , and generate a new candidate cell state, \tilde{c}^t . The output gate layer, o^t , generates new cell state, c^t and hidden state, h^t . We explored two types of sequential modelling of student actions.

- **Last Hidden State.** In vanilla LSTM, the last hidden state at $(t-1)$, $h_{(t-1)}$ captures

students' sequential action sequence. We concatenate $\mathbf{h}_{(t-1)}$ to question embedding at t (Figure 4.1).

- **With Attention.** Attention mechanisms assign weight to hidden states. In our study, we adopt a global attention mechanism (Luong et al., 2015). Global attention mechanism considers all previous actions to compute the final output of LSTM. The final output is a weighted function of all previous hidden states. We applied cosine similarity to compute the weighted sum between the exercise at time t , \mathbf{E}_t and previous actions. Cosine similarity-based attention model has been used in education contexts, such as in exercise-enhanced knowledge tracing (Su et al., 2018), question-difficulty prediction (Huang et al., 2017), and finding similar exercise-pairs in online education system (Liu et al., 2018). Mathematically, the hidden state computation of attention is as follows

$$\mathbf{h}_{att} = \sum_{i=1}^{(t-1)} \alpha_i \mathbf{h}_i, \alpha_i = \cos(\mathbf{E}_t, \mathbf{E}_i) \quad (4.6)$$

Instead of equation 4.4, we apply Equation 4.6 in the concatenation layer. Figure 4.2 shows \mathbf{h}_{att} computed from Equation 4.6.

Prediction. After obtaining students sequential action representations upto $(t-1)$, we predict student's question performance at time t . We transform the question text embedding vector \mathbf{E}_t by a dense layer and obtain hidden representation \mathbf{E}_t . Then we concatenate and the hidden state from LSTM, $\mathbf{h}_{(t-1)}$ (or \mathbf{h}_{att}) followed by a dense layer with a Rectified Linear Unit (ReLU). Finally, we apply a linear layer to minimize the mean squared error (MSE). The equations are

$$\begin{aligned} y &= ReLU(\mathbf{W}(\mathbf{E}_t \oplus \mathbf{h}_{(t-1)}) + \mathbf{b}) \\ Prediction &= Linear(y) \end{aligned} \quad (4.7)$$

where \mathbf{W}, \mathbf{b} are network parameters to learn and \oplus is concatenation operator.

4.4.2 Experiments of RQ1.3

Experiment Setting

Parameters.

We explored window size to capture sequential actions prior to attempting a question. Previous researchers have used window-based limit to capture users' sequential behavior data (Choffin et al., 2019; Henderson et al., 2020; Spathis et al., 2019). We examined descriptive statistics of the number of sequential actions. The descriptive statistics of number of student actions prior a question attempt is (Mean = 6.47, Median = 5.00, and standard deviation (SD) = 5.37). Based on our observation, we selected a window size of 7 and 10 for our study. We used front padding to pad sequences shorter than the window size.

We applied the standard k -fold cross-validation ($k = 5$) at student-level to evaluate our models. During the training stage in LSTM-based models, we used 5% of data as validation. For fair comparison among models, the dataset split remained the same across all models (i.e., LSTM-based models and baseline models).

Embedding. We used Universal Sentence Encoder (USE) ?. USE can take phrases, sentences, and short paragraphs as inputs and encodes inputs into a fixed-length vector of 512.

Implementation. We tuned hyperparameters on the validation set for each fold. The hyperparameters of vanilla LSTM model are LSTM hidden dimension = {128, 100}, dropout = {0.33, 0.66}, and learning rate = {0.0001, 5×10^{-5} }. Hyperparameters of LSTM with an attention model are: LSTM hidden dimension = {100, 80}, dropout rate = {0.33, 0.66}, and learning rate = {0.001, 5×10^{-5} }. The dense layer for \mathbf{E}_t and concatenation layer have dimensions 64 and 16, respectively.

We implemented our models in Keras with the Tensorflow backend. We used the RM-Sprop optimizer and a batch size of 146 and 50 epoch to train all models.

Baselines

We compared our model with the following static baseline methods. We consulted previous literature (Geden et al., 2020; Henderson et al., 2020) of sequential student modelling to select our baselines.

- Support Vector Regressor (SVR): Support vector regressor applies kernel to predict the output of a regression.

- Gradient Boosting Regressor (GBR): Gradient boosting regression applies tree ensembles to approximate the predicted variable.
- Linear Regression: We applied Elastic Net as linear regression. Elastic net applies L1 and L2 regularizations.

Features of Static Models. As static models can not handle sequential data; we prepared the following nine aggregated features to apply these models. We aggregated features within the window size, w of students’ sequential actions. These features are the number of SRL features, such as highlights, annotations, vocabulary lookups, number of question attempts, Avg. scores of previously attempted questions, Avg. Cosine Sim (E_t , highlights), Avg. Cosine Sim (E_t , annotation), Avg. Cosine Sim (E_t , vocab. lookups), Avg. Cosine Sim (E_t , previous question texts)

We implemented all our baseline static models in Scikit-learn Pedregosa et al. (2011a).

4.4.3 Results of RQ1.3

Table 4.6: Experimental Results. Mean (Standard Deviation) of 5-fold cross validation on test dataset. w = window size.

$w = 7$	Methods	MSE	MAE
	SLSTM	0.1358 (0.013)	0.2983 (0.036)
	LSTM+Att	0.1325 (0.008)	0.2981 (0.017)
	GBR	0.1482 (0.021)	0.0340 (0.047)
	SVR	0.1394 (0.014)	0.297 (0.029)
	Linear Reg	0.1376236 (4.69e-05)	0.30316(2.70745e-05)
$w = 10$			
	LSTM	0.1358 (0.013)	0.2986 (0.0356)
	LSTM+Att	0.1353 (0.011)	0.3063 (0.0236)
	GBR	0.1499 (0.0226)	0.3067(0.0488)
	SVR	0.1406 (0.0138)	0.2981 (0.02900)
	Linear Reg	0.13762 (4.69e-05)	0.30316(2.59 e-05)

In this section we describe our results.

Performance Statistics. Table 4.6 presents performance of all models. Boldface indicates the best performing model within each metric. From Table 4.6, we observe that

sequential models perform better than their non-sequential counterparts. Also, we observe that within sequential models, attention-based LSTM performs better. It has the lowest MSE and MAE in window size = 7. Additionally, the standard deviation is also the lowest.

Effect of Window Size. Table 4.6 also shows results for window size, $w = 7$ and $w = 10$. We observe that MSE increases for LSTM-attention based model, GBR, and SVR. For Elastic Net regression and the vanilla LSTM model, the MSE remains same upto four decimal places.

From Table 4.6 results, we can say that, the recency of an activity has greater impact in LSTM-based models performance. This could explain the models' relatively lower performance with $w = 10$ than $w = 7$. The window-based technique was introduced by Choffin et. al Choffin et al. (2019). Authors have assumptions that "students first have access to some theoretical knowledge about skills, but learning happens with retrieval practice". Authors also assume that mastery of a skill decreases as time goes by and students tend to forget. Observing our results we conclude that, smaller window size captures more recent activities resulting in more predictive power.

Attention Visualization

In this subsection, we present two case studies of attention visualization on the testing set. We describe two test cases. In the first case, the student scored 0.5 in question attempt at time t (Case (a), Figure 4.4). In the second case, a student got full score in question attempt at time t (Case (b), Figure 4.3). Both test examples are taken from window size = 7, experiments.

Case (a). Figure 4.4 shows the student's action sequence prior to a question attempt. The student had a 3-length action sequence. As we have used front padding, we have indexed the action timestamps as T 5, T 6, and T 7.

In this example, the student attempted one question (T 5), and performed two highlights (T 6, T7). The student scored 0.25 in the question attempt at T 5. We notice that the model put maximum weight at T 5 followed by event at T 6 and T 7.

Case (b). Figure 4.3 shows the students last 7 actions prior to a question attempt at T 8. The student performed two annotations {at T 1, T 2}, attempted one question (T 3), performed one annotation (T 4), and attempted three questions (T 5, T 6, T 7). To perform an annotation, a student first selects a text and then writes down their notes. The bar plot shows relative weight the model puts on student's action.

From the bar plot of Figure 4.3, we observe that the model puts relatively higher weights on actions at time T 2, T 3, and T 5. From the textual description of the question at T 8, we

Time	Action	Action Text
T 1	A	The tool that rusted is made of iron, and the other tool is made of aluminum. The ability to rust is a chemical property of iron but not aluminum. <i>Good to know</i>
T 2	A	Chemical properties are properties that can be measured or observed only when matter undergoes a change to become an entirely different kind of matter. <i>Definition of chemical properties</i>
T 3	Q	Paper can be torn into pieces. Is this an example of a physical or chemical property? Why? (Score: 1)
T 4	A	Reactivity is the ability of matter to combine chemically with other substances. <i>Definition of reactivity.</i>
T 5	Q	If potassium is reacting with water, what must be occurring for it to be considered a chemical change? Select the best 2 answers (Score: 0)
T 6	Q	Which of the following are chemical properties of silver, shown in the image below? Select all that apply. (Score: 0)
T 7	Q	Identify the metal if it is somewhat reactive, reacts with acids but not oxygen or water. (Score: 0)



T 8	Identify 3 ways a substance can undergo a chemical change and what evidence would you use to support that a chemical change has occurred?
-----	-------------------------------------------------------------------------------------------------------------------------------------------

Figure 4.3: Case (b) Attention Visualization at test time. True score: 1.00. Predicted score: 0.84. Scores are scaled in [0-1] range. Q = question attempt, A = Annotation. For annotations, *Blue* font = Student's note text and *black* font = selected text.

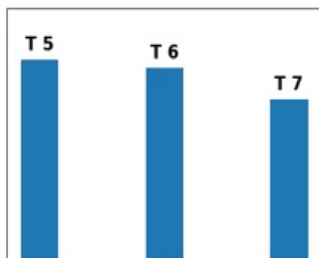
observe the question text has higher word overlapping with Action Text at T 2, T 3, and T5.

4.4.4 Discussion of RQ1.3

In this work, we investigated students' performance on questions considering their sequential behaviors. To achieve this goal, we developed predictive models based upon students' question attempts and SRL activities that contribute to learning. We summarize our findings and outline possible implications.

For Researchers. Although researchers have proposed more than dozens of variations on the basic DKT model (i.e., Ai et al. (2019); Cheung & Yang (2017); Ghosh et al. (2020); Liu

Time	Action	Action Text
		⋮
T 5	Q	What is the purpose of the periodic table? (Score = 0.25)
T 6	H	The modern table is based on Mendeleev's table, except the modern table arranges the elements by increasing atomic number instead of atomic mass
T 7	H	The modern table has more elements than Mendeleev's table because many elements have been discovered since Mendeleev's time



T 8	Compare and contrast Mendeleev's periodic table with the modern periodic table.
-----	---------------------------------------------------------------------------------

Figure 4.4: Case (a) Attention Visualization at test time. True score = 0.5. Predicted score = 0.59. Scores are scaled in [0-1] range. Q = question attempt, H = Highlighted text.

et al. (2019); Mongkhonvanit et al. (2019); Nagatani et al. (2019); Pandey & Karypis (2019); Zhang et al. (2017)), these enhancements consider only students' past problem-solving behavior to predict future question performance. We argue that it is important to consider question solving and other student interaction behavior during learning—an area still underexplored Choi et al. (2020). We have utilized question-answering and reading-related features in predictive model. We believe that researchers in educational data mining domain and learning science domain will find our work helpful.

Personalized Education. Our models can be helpful to design personalized education. Specifically, our proposed attention-based LSTM model has explanatory power. These explanations can be helpful for students to understand their sequential activities and how those activities can impact their question performance. A next step can be to design a recommendation system based on students' activity.

For Teachers. Our model can also provide support for educators. In a teaching context the interpretation of a model important for intervention. By providing an estimate of student performance that is tied to specific study habits and SRL features we can: (i) better

identify *what* contributed a students' performance in score and (ii) *what* study habits they can engage in to support better outcomes.

4.5 Conclusions

In this Chapter, we first examined the association between SRL behavior and scores of MCQ and SA questions in RQ1.1. Our findings of HLM showed that the frequency of SRL was more predictive of SAs than MCQs. We conclude that answering SAs may require more SRL and monitoring behavior, whereas students can rule out distractors in MCQs. In our RQ1.2, we measured the association between cosine similarity (SA text, SA response) and SA score. Our findings show that Spearman's correlation coefficient was significant for both science and social studies data, and the coefficient value was higher in science assignments. We conclude, science students' SA responses contained more similar words from the question text. One possible explanation could be SA question, and students' responses contained both domain-specific science vocabularies—resulting in higher overlapping with question texts. In RQ1.3, we designed predictive models considering the (i) *textual description* of question and SRL and (ii) *sequential ordering* of prior question attempts and SRL. We designed two predictive models: one with LSTM and another LSTM with an attention mechanism. We incorporated a window-based method to capture recent activities before each question attempt. Experiments showed that the smaller window size resulted in better prediction. From this observation, we conclude that the recency of action has better predictive power.

Overall, in this Chapter, we examined students' SRL behavior and question format considering SRL frequency (RQ1.1), students' responses to SA (RQ1.2), and textual description and sequential ordering of question and SRL (RQ1.3). The following Chapter presents techniques to identify question-solving and SRL patterns among productive and unproductive students.

CHAPTER

5

SRL PATTERNS AND PERFORMANCE

In this Chapter, we will examine how SRL reading behaviors relate to performance on two types of questions, MCQ and SA, across two subjects, science and social studies. We hypothesize that differential sequence mining (Kinnebrew et al., 2013) can identify patterns of student behaviors that

1. vary across groups of students based on performance score on questions within subject
2. vary across subject domain

To test our hypothesis, we split the RQ2 into two subquestions as follows.

RQ2.1: How can students be grouped according to their question performance?

RQ2.2 How observed reading and SRL patterns differ between groups of students within and across science and social study domains?

In this Chapter, first we will describe background and related literature of RQ2 in Section 5.1. Next, we will present our methodology and results of two sub-RQs in Section 5.2. As a part of our analysis, we will group students to identify high and low performance students within two subjects. To do so, we will cluster students using question performance features,

such as first attempt score, last attempt score, and number of attempts (RQ2.1). After clustering students by performance, we will identify frequent patterns of reading and SRL usage within each cluster (RQ2.2). We discuss our findings and conclude this Chapter in Sections 5.3 and 5.4, respectively.

5.1 Background

5.1.1 Sequential Pattern Mining

In order to better understand students' learning, researchers have applied sequential pattern mining techniques to different educational contexts. One such context is analyzing online learner behaviors in MOOCs. Researchers have analyzed MOOC students' clickstream data to understand how they transition between activities (Davis et al., 2016; Pardos et al., 2017; Wen & Rosé, 2014). Davis et al. identified common 8-gram event sequences to represent study patterns (Davis et al., 2016). They applied Markov models to trace transitions between activity patterns. Wen and Rose used the n -gram technique to describe learners' activities and grouped learners by their habitual behavior. Their grouping offered insights into how high and low performing student behaviors may differ (Wen & Rosé, 2014). Pardos et al. utilized MOOCs clickstream data to generate a personalized next-step recommendations system (Pardos et al., 2017). Their model extended a student behavior model which tracked students' time spent on pages and utilized this information for forecasting next-pages student may spend significant time.

Researchers have also used sequence mining techniques in intelligent tutoring systems to trace students' learning behaviors. Akram et al. employed sequence mining techniques to identify common problem-solving strategies in a game-based learning environment (Akram et al., 2018). They applied n -gram sequencing to represent students problem-solving patterns and clustered patterns to identify groups. Mirzaei et al. clustered students' interaction patterns for object-oriented programming learning and identified high and low performing students' patterns (Mirzaei et al., 2019). Gitinabard et al. applied sequence mining techniques to study students' transitions behavior among three platforms, namely GitHub, Piazza, and WebAssign (Gitinabard et al., 2019). They split students into two groups by final grade 'B+' as "Distinction" and "Non-Distinction" for two undergraduate level courses: Java Programming and Discrete Math course. Students belonging "Distinction" category made more transitions among platforms and participated in discussion forums compared to "Non-Distinction" groups.

One challenge in analyzing sequence mining is the number of patterns could be very large. Applying threshold can limit the sequences. However, finding a threshold value is challenging. For example, applying 80% threshold may result in filtering important sequences whereas a lower threshold may result in a large pool of patterns for analysis. To overcome this challenge, Kinnenbrew and Biswas developed differential sequence mining technique (Kinnebrew et al., 2013) to identify frequent patterns those differ within groups (e.g., control vs. experimental, high vs. low performing). Kinnenbrew and colleagues identified students' productive and unproductive behavior by applying the differential sequence mining technique in an open-ended learning environment, Betty's Brain (Kinnebrew et al., 2013; Kinnebrew & Biswas, 2012). In one study within MetaTutor, Bouchet and colleagues examined frequently used patterns in 148 undergraduate students (Bouchet et al., 2012). They clustered 51 students receiving experimental conditions into 3 groups using 13 performance features. They applied differential sequence mining technique (Kiladis et al., 2006) to identify SRL strategy usage by High (H) and Low (L) performing groups. Their result showed students of H cluster summarized, annotated more, and better identified relevant pages for subgoals. Students of L cluster could not identify relevant pages. When prompted for monitoring, they were mistaken in their evaluation (JOL, FOK, CE). Similar to Kinnenbrew et al. and Bouchet et. al, (Kinnebrew et al., 2013; Kinnebrew & Biswas, 2012), we apply differential sequence mining techniques to compare patterns among groups of students. Further, we apply differential sequence mining techniques to identify domain-specific reading and SRL patterns in science and social studies.

5.1.2 SRL in Cross Domain

SRL researchers emphasized SRL can vary across subject domains, (Greene et al., 2013; Poitras & Lajoie, 2013). Alexander and colleagues surveyed 77 literature on SRL to understand domain-general vs. domain-specific studies (Alexander et al., 2011). Their coding scheme showed the majority of the articles did not consider any specific domain (n = 24, 37%) and only three studies compared cross domains (3.90%). Science (physical, chemical, and biological science) was the most widely used to study SRL in a single domain (n = 20, 25.97%). The authors also coded SRL literature with domains into three groups: purposeful, convenience, and implied. Purposeful indicating authors' explanation of selecting subject domain to study SRL whereas convenience indicating no such explanation. Their study found 22 articles into both purposeful and convenient group and 5 articles into implicit group. Considering SRL literature on science domain, 9 studies were purposeful, 7 were con-

venient, and 3 were implied. In studies belonging to purposeful group, authors explained on complexity of the science topic and little prior research.

Poitras and Lajoie proposed a framework based on Winne and Hadwin's model to measure domain-specific SRL on history (Poitras & Lajoie, 2013). Using their model, the authors designed a CBLE with an SRL fostering agent for history (Poitras & Lajoie, 2014). The agent provides hint, prompt, and feedback to foster SRL skills in history. In another study, Rotgans and Schmidt found no significant difference in SRL behaviors for mathematics, science, and English course for 155 first-year students in at a Singaporean polytechnic (Rotgans & Schmidt, 2009). However, they relied on questionnaires to measure SRL, a measure which is not effective in measuring SRL (Rovers et al., 2019; Zhou & Winne, 2012). Greene and colleagues measured science and history domain-specific SRL with 91 undergraduate students in a digital library (DL) environment (Greene et al., 2015). DL is a curated set of multimedia representation (textual, video, and images). The history task focused on description of Blue Ridge Parkway of North Carolina. Science task was to understand change of matters from solid to liquid then gaseous state. They adopted think aloud protocol and Azevedo's scheme to code SRL (Azevedo & Cromley, 2004). Authors designed pre and post test with MCQ, fill-in the blanks, and essay questions for both domains. Common SRL behaviors across the two domains were corroborating sources (comparing information from multiple sources in DL), positive feelings of knowing, knowledge elaboration (draw conclusion by elaborating reading with prior knowledge), memorization, prior knowledge activation (relevant prior knowledge), self-knowledge activation (adopting a strategy because it is helpful to their personally), and time planning. Students assigned in science DL exhibited more monitoring activities, such as content evaluation and judgement of learning. Students assigned in history DL showed more knowledge elaboration and annotation behavior. Our study is similar to Greene et al. as we compare reading and SRL behaviors between science and social studies. However, we study middle school reading-related SRL and examine which SRL patterns differ in high and low performing students.

5.2 SRL Patterns and Performance

In this section, we present methodology and results of two subquestions of RQ2,

RQ2.1: How can students be grouped according to their assignment performance? and

RQ2.2 How observed reading and SRL patterns differ between groups of students within and across science and social study domains?

5.2.1 Methodology of RQ2

We first started extracting students performance features on embedded questions within the AL system. We clustered students by performance features for science and social studies. Next, we applied the n -gram technique to represent reading and SRL patterns within each cluster. We applied differential sequence mining techniques (Kinnebrew et al., 2013) to identify patterns between groups of students within subject domain and across subject domain and to determine if these patterns were different at statistically significant levels. Figure 5.1 shows the pipeline of our methodology.

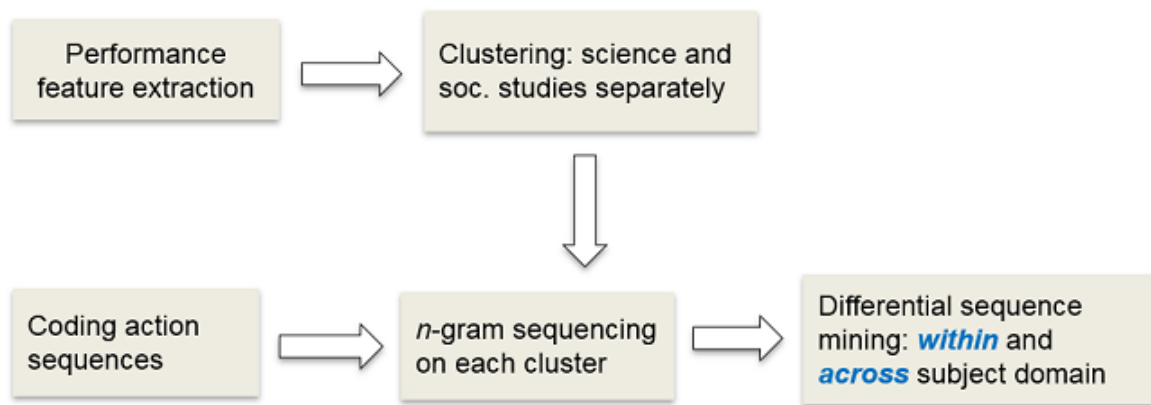


Figure 5.1: Methodology Overview of RQ2

Performance Feature Extraction

We calculated four types of scores for each MCQ and SA, resulting in eight performance features. Of these four features, the first two are raw features and the remaining ones are derived. The raw features are: the students score on their first attempt at the question, and their score on the last attempt. We also define a normalized attempt count calculated by counting the number of attempts for each question by a student. We then normalize the attempt count by total student attempts of that question within the assignment.

We multiplied the last attempt score by the normalized attempt count and refer it as the *Norm_Last* score. The *Norm_Last* score is a derived feature combining attempts and scores. The fourth feature counts the number of attempts for each question considering the median timestamp by which students of that class answered it . To compute the fourth

feature, we measured the proportion of attempts a student made after the median time for all students on that question. Here we use the time elapsed since the student began reading the assignment. We refer to this feature as *Long_Submit*. The purpose of introducing feature is inspired by the approach of Mirzaei et al. (Mirzaei et al., 2019). Mirzaei et al. clustered students of a programming language course according to their problem solving patterns. To code student behaviors, they defined short and long success for a successful question attempt considering the median time to attempt the question. Similarly, we want to get a sense of students' problem solving attempt behaviour relative to other students with this feature. We assume an assignment is available to all students at the same. A *Long_Submit* of a question indicates the student may be struggling with the question.

Table 5.1: Mean and (Standard Deviation) of Performance feature for Science and Social Studies

Feature	Science (n =12,548)	Social Studies (n = 16,240)
MCQ_First	2.80 (1.83)	2.17 (1.99)
MCQ_Last	2.84 (1.82)	2.19 (1.99)
MCQ_Norm_Last	2.71 (1.80)	2.12 (1.96)
MCQ_Long_Submit	0.67 (0.47)	0.71(0.45)
SA_First	2.46 (1.35)	2.56 (1.36)
SA_Last	2.58 (1.30)	2.62 (1.33)
SA_Norm_Last	2.05 (1.29)	2.28 (1.35)
SA_Long_Submit	0.66 (0.46)	0.63 (0.48)

Clustering Students by Performance

To prepare our dataset for clustering, we averaged all performance features for each student in science and social studies. This resulted in one row per student in our final dataset.

We applied K-means clustering for $K = 2$ to 10 (with 25 random start values for each K) to group students by their performance features. To determine the optimal number of clusters, we calculated the Silhouette width (Hennig et al., 2015) to assess the cluster separation. We found $K = 4$ clusters to be the optimal number for both domains.

Coding Student Action Sequences

In our study, a student action can be answer submissions of questions, reading, note taking, highlighting, and vocabulary lookups. We refer to answer submission activities as performance activities and the rest as reading and SRL activities. To code performance activities, we distinguished between the first attempts and consecutive attempts for MCQ and SAs. A first MCQ attempt is coded as (M), a resubmission of an MCQ as (m); a first SA attempt as (S), and a resubmission attempt as (s). We coded a reading event as (R), an annotating event as (A), a highlighting event as (H), and a vocabulary lookup event as (V). The AL system stores timestamps of events. In order to produce our action sequences, we aggregated students' assignment actions and timestamps related to reading, vocabulary lookups, highlighting, annotation and question answering into a unified transaction log (Sheshadri et al., 2018; Kovanović et al., 2015). The AL system does not record explicit student sessions. Therefore we adopted a data-driven approach described by Kovanovic et al. (Kovanović et al., 2015) and used by Adithya et al. (Sheshadri et al., 2018) to partition the student logs into sessions. More specifically we conducted an exploratory analysis of the time between observed student actions to identify outliers and to estimate the last action in any session. We began by plotting the intervals between two sequential actions within an assignment. Based upon this analysis we chose a conservative cutoff of 30 minutes. Any gap larger than 30 minutes represents a new reading session.

After defining sessions, we split all student activities within a single assignment by session. We compacted repeated events by + as done by Kinnenbrew et al. (Kinnebrew & Biswas, 2012; Kinnebrew et al., 2013). For example, RSSM was represented as RS+M.

Sequential Pattern Mining

We encoded sequences as n -grams using the Scikit-Learn library (Pedregosa et al., 2011*b*). The n -gram representation extracts sequences of n adjacent elements from a string. Our preliminary exploration of the data showed that frequent sequences ranged in length from two to four. Thus, we analyzed sequences of two to four events in length from the student sessions. As we are focusing on students' reading and SRL behavior, we focused on those n -grams which contained at least one letter from the set {R, A, V, H}.

Differential Sequence Mining

We applied differential sequence mining (Kinnebrew et al., 2013) to identify distinguishable patterns between two groups of students. Differential sequence mining requires two parameters: s-support and i-support. The s-support measures the frequency of a pattern within a group. In our context, the s-support counts the number of students exhibiting a pattern in a group. We employ $s\text{-support} = 0.5$ to analyze patterns showed by at least half of the students in a group. The i-support parameter measures the frequency of a pattern within one action sequence. For example, the i-support of 'R' in an action sequence 'RSRM' is 2. The i-support is important to measure the extent a student shows a particular pattern. To measure i-support of a pattern, we calculate the mean of a pattern's i-support value across all sequences in the group, similar as in (Kinnebrew et al., 2013).

To apply differential sequence mining technique, we first generated sequential patterns from a group of students using the n-gram sequencing technique. We then applied $s\text{-support} = 0.5$ to filter patterns exhibited by at least half of students within that group. We calculated mean i-support value of each filtered pattern within a group. Next, we applied the Kruskal-Wallis test, a nonparametric analogue to ANOVA, to identify if there is a significant difference in the mean i-support value within the groups (i.e. one group used a pattern more often than the other)

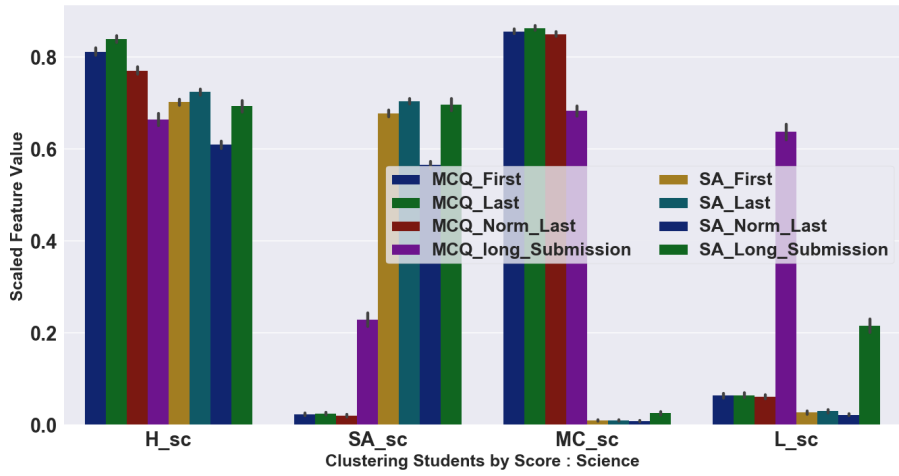
Differential sequence mining algorithm produces four categories of frequent patterns (Kinnebrew et al., 2013; Bouchet et al., 2012): two categories in which a pattern is present in both groups but one group exhibited the pattern significantly more frequently than the other and two categories in which the pattern is present and satisfies s-support in one group.

5.2.2 Results: RQ2.1

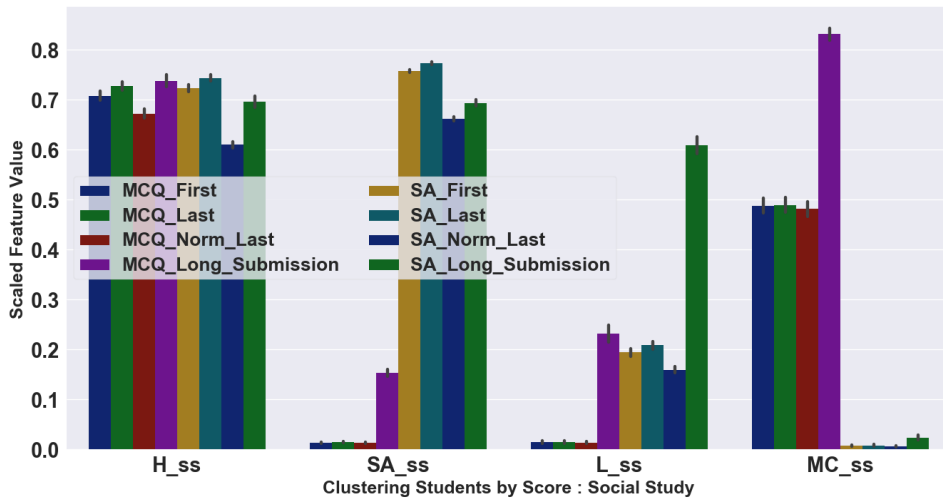
Clustering Results

The clustering results are shown in Figures 5.2a and 5.2b. When we compare the two figures, we notice some similarities between students' performance on science and social studies. One group of students performed well in both MCQ and SA questions. Similarly, one group of students performed well either in MCQ or SA question but not on both. This phenomenon can be attributed to teachers' behavior with number and types of question settings on assignments. We break these results down by cluster below.

- **Clusters in Science**



(a) Science



(b) Social Study

Figure 5.2: Clustering Students by Assignment Performance

Four science clusters are: cluster with high SA scores, cluster with high MCQ scores, cluster with low performance on both types of questions, and cluster with high scores on both SA and MCQ. We will refer cluster these clusters as SA_sc (SA performers in science), MC_sc (MCQ performers in science), L_sc (low performers in science), and H_sc (high performer in science) in the following sections. SA_sc students had generally high SA scores but low MCQ scores across the metrics. From the bar plots of 5.2a, the SA normalized last score, SA_Norm_Last, is lower than the raw last score. This indicates that students in this group students resubmitted the SA questions more. This cluster has 2,636 students. The MC_sc students, by contrast, had higher MCQ scores and low SA scores. This cluster has 4,471 students. While the L_sc has 2,341 students who are low performers on both the MCQ and SA questions. Additionally, students in this cluster had more long MCQ submissions. And finally, the H_sc has 3,100 students who were high performers in both question formats.

- **Clusters in Social Studies** Similar as in science student clustering, we have four different groups in social studies. We will refer these clusters as H_ss (high performer- soc. study), L_ss (low performer- soc. studies), MC_ss (MCQ performer-soc.studies), and SA_ss (SA performer soc. study). H_ss is the students with the largest cluster size 8,948 students. Comparing with high performers in students assignments H_sc, we observe H_sc students have higher MCQ First, MCQ Last score, and fewer MCQ Long submission (Figure 5.2a and 5.2b). H_ss students scored higher in SA questions than Sc_ss. L_ss cluster has 2760 students. Students in this cluster had low MCQ scores. They also had moderate SA scores and more SA_Long_ Submissions. This behavior indicates students might be struggling with SA questions as well. The MC_ss students, by contrast, performed better in MCQs than SAs. This cluster is the second largest one in size with 2,928 students. And finally, the SA_ss cluster has 1604 students with good SA scores compared to MCQ scores.

5.2.3 Results: RQ2.2

Performance Differences Connecting With Reading and SRL Patterns

We examined the students' differential sequences both within and across the domains. We considered a sequence to be differentiable within a domain if it can separate groups of students within the same subject. Cross-domain sequences, by contrast, are differentiable if they separate the two domains. The I-supp Diff column presents the mean i-support

difference of the pattern (i.e. frequency of the pattern within one action sequence) between the two groups. A positive value of mean i-support difference, I-supp Diff, indicates the pattern was more frequently used by the group on the left than on the right. For example, Table 5.2 shows RS pattern has i-support difference of 0.17 in H_sc vs L_sc group. This indicates RS pattern was used more frequently by H_sc than L_sc group, resulting in a positive difference. The sequences are sorted in descending order of mean i-support value, as in (Kinnebrew & Biswas, 2012). We primarily focus on identifying distinguishable patterns in high and low performing groups of students. The *Group* column of each table describes on which group the pattern belongs : both or only one.

Differentiable Patterns Within Domain: Science

Table 5.2: Differential patterns :Science. *** = $p < 0.001$, ** = $p < 0.05$

Pattern	H_sc vs L_sc			L_sc vs Rest			H_sc vs Rest		
	<i>I-supp Diff</i>	<i>p</i>	<i>Group</i>	<i>I-supp Diff</i>	<i>p</i>	<i>Group</i>	<i>I-supp Diff</i>	<i>p</i>	<i>Group</i>
RS	0.17	***	Both	-0.15	***	Both	0.07	***	Both
RS+	0.08	***	Both	- 0.07		Both	0.22	***	Both
SR	0.08	***	Both	-0.06	***	Both	0.04	***	Both
A+	0.03	***	Both	-0.02	***	Both	-	-	
AS	-	-	-	-	-	-	0.02	***	Both
V+M	-0.0001	***	Both	-	-	-	-	-	
RS+s	-	-	-	-	-	-	-0.002	***	Both
RH+M	-0.002	***	Both	0.001	***	Both	-	-	
+RH	-0.006	**	Both	-	-	Both	-	-	
RM	-0.16	***	Both	0.16	***	Both	-	-	
VsV	-0.0005	**	L_sc	-	-	-	-0.0005	***	Rest
AVAM	-0.0004	**	L_sc	-	-	-	-0.0005	**	Rest
ASA	0.005	***	H_sc	-0.002	***	Rest	-	-	
HS	0.006	***	H_sc	-0.003	**	Rest	-	-	

Pattern RS and RS+ Patterns RS and RS+ indicate that the students read text before attempting one (RS) or multiple short answers (RS+). These two patterns are common to both the H_sc and L_sc groups. The positive i-support differences observed (0.17 and 0.08 respectively) indicate high performing students read more before attempting a SA. It also indicates that students in L_sc group showed this pattern less frequently. Additionally, we can compare the i-support difference of RS and RS+ in cluster H_sc

vs L_sc and L_sc vs (H_sc, MC_sc and SA_sc). This finding reconfirms that the L_sc group spent less time reading before attempting to answer short answer questions. Similarly, we can confirm the reading behavior as H_sc students by observing the positive i-support difference values in H_sc vs Rest (L_sc, MC_sc and SA_sc)

Pattern SR Pattern SR indicates that after an SA submission, the student continued reading. This pattern also may indicate that students are revisiting the reading material after an SA attempt. When considering the H_sc and L_sc clusters, we also observe students in L_sc group showed this pattern less than H_sc group. This behavior of L_sc group students is also confirmed from L_sc vs Rest group (H_sc, MC_sc & SA_sc) result.

Pattern A+ An A+ pattern indicates that the student made multiple successive annotations. Similar to previous three patterns, H_sc group students show more annotations than L_sc group students. Less annotation behavior of L_sc group students is also observable from L_sc vs Rest (H_sc, MC_sc, & SA_sc) results.

Pattern RM The RM pattern indicates a reading behavior before an MCQ submission. Students of L_sc group exhibit this pattern more than the other three groups. From Figure 5.2a, we observe that the L_sc group students have higher long MCQ submissions overall. The MCQ_Last score, by contrast, is lower overall even after long MCQ submissions. Combining observation from Figure 5.2a and pattern RM, we conclude L_sc group students struggle in choosing the correct MCQ option.

Pattern RH+M and V+M Patterns RH+M and V+M are also related with MCQ score behavior of L_sc group students. RH+M indicates students are reading (R) followed by multiple highlighting (H+) before submitting an MCQ. Pattern V+M indicates multiple vocabulary lookups before attempting an MCQ. These two patterns and the previous one characterizes L_sc group students' MCQ answering pattern.

Pattern +RH Pattern +RH indicates a read and highlighting event after multiple actions of another event ($X = M, m, S, s, A, V, H$) other than reading (otherwise it would be R+). This pattern is common in both H_sc and L_sc group students but the L_sc group students exhibit more often. For the students in the H_sc group, this pattern may explain they were revisiting the reading material and highlighting important lines after multiple occurrences of X. For low performing students, it may explain they were struggling with the action, and thus continue reading and highlighting.

Patterns VsV and AVAM Only L_sc student groups exhibited these patterns frequently. The VsV pattern indicates vocabulary lookup behavior prior resubmitting a SA question. The AVAM pattern indicates annotation and vocabulary lookup behavior prior submitting an MCQ. As we can observe from 5.2a L_sc group students had lower MCQ last score and longer MCQ submission time, we conclude L_sc students were struggling in attempting MCQs.

Patterns ASA and HS Only H_sc student groups exhibited these patterns frequently. Both of these patterns are related to SA attempts. ASA patterns indicated annotation behavior prior and after an SA attempt and HS indicates highlighting prior an SA attempt.

Differentiable Patterns Within Domain: Social Studies

Table 5.3: Differential patterns :Social Study. *** = $p < 0.001$, ** = $p < 0.05$

Pattern	H_ss vs L_ss			L_ss vs Rest			H_ss vs Rest		
	I-supp Diff	p	Group	I-supp Diff	p	Group	I-supp Diff	p	Group
RS	0.06	***	Both	-0.04	***	Both	0.039	***	Both
RS+	0.04	***	Both	- 0.02	***	Both	0.03	***	Both
RS+M	0.02	***	Both	-	-	-	-	-	-
SR	0.14	***	Both	-	-	-	0.009	***	Both
VS	0.013	***	Both	-0.01	***	Both	-	-	-
s+R	-0.001	**	Both	0.001	**	Both	-	-	-
+R	-0.005	**	Both	0.01	***	Both	-	-	-
S+RS	-0.007	***	Both	-	-	-	-	-	-
VM+	0.002	***	H_ss	-0.0007	**	Rest	-	-	-
Rm	0.002	***	H_ss	-	-	-	-	-	-
RmS+	-	-	-	-	-	-	0.0004	***	H_ss
S+A+	-0.0003	**	L_ss	-	-	-	-	-	-
sVs+	-0.0003	**	L_ss	-	-	-	-	-	-

Pattern RS and RS+ As with the science domain we believe that patterns RS and RS+ occur more frequently in H_ss group than L_ss group. RS+M pattern is an extension of RS+. This indicates students read before submitting multiple SA questions and an MCQ. All these three patterns indicate that the student engaged in reading activities prior submitting questions.

Pattern SR The pattern SR indicates student performing a reading activity after an SA

submission. This finding may indicate reviewing the reading material behaviour of students. Students from L_ss group engaged in this behavior less frequently than the H_ss group.

Pattern VS Pattern VS, on the other hand, indicates that the student engaged in vocabulary lookup before attempting an SA. This pattern indicates students in high performing students looked up vocabulary more often than low performing students.

Pattern s+R Pattern s+R indicates multiple resubmission of SAs followed by a reading event. From the negative I-supp Diff value in H_ss vs L_ss group, we know that L_ss exhibited this pattern more often than H_ss. As Figure 5.2b shows, the L_ss group has a high value of long SA submission with respect to median submission timestamp for that question. Taken together, the results shown in Figure 5.2b and Table 5.3 allow us to conclude that the L_ss students were frequently struggling with the SA questions and, after multiple attempts, turned to reviewing the material.

Pattern +R Pattern +R is an important subset of the previous pattern, s+R. Our explanation for this pattern is that it is similar to +R pattern in science studies. L_ss group exhibited this pattern more than H_ss.

Pattern S+RS Finally pattern S+RS indicates a series of SA submissions, followed by reading and another SA submission. This pattern explains student's reading behavior in for first time SA attempt. L_ss group students attempted multiple SA questions before reading. This pattern explains an important distinguishable feature of H_ss and L_ss students in approaching SAs. As discussed above, RS and RS+ patterns exhibited more by H_ss students. Based upon the i-support differences shown in Table 5.3 for: RS, RS+, and S+RS, we can say that H_ss students were reading before attempting SAs and L_ss students were doing the opposite activity.

Patterns VM+, Rm, and RmS+ Only H_ss student groups exhibited these patterns frequently. The VM+ pattern indicates vocabulary lookup behavior prior attempting an MCQ. The Rm pattern indicates MCQ resubmission (m) behavior. H_ss students read more prior an MCQ resubmission. The RmS+ pattern is a superset of Rm pattern including multiple SA submissions followed by an MCQ resubmission (m).

Patterns S+A+ and sVs+ Only L_ss student groups exhibited these patterns frequently. Both of these patterns are related to SA attempt behavior of of L_ss students. The S+A+ pattern indicates L_ss students annotated *after* SA submissions. The sVs+ pattern

indicates L_{ss} students looked up vocabularies prior and after an SA resubmission (s). From figure 5.2b, we observe L_{ss} students had longer SA submission time and lower SA last score. We conclude that L_{sc} students were struggling in SA attempts.

Table 5.4: Differential patterns :Science vs Social Study. *** = $p < 0.001$, ** = $p < 0.05$

Pattern	H _{sc} vs H _{ss}			SA _{sc} vs SA _{ss}			MC _{sc} vs MC _{ss}		
	<i>I-supp Diff</i>	<i>p</i>	<i>Group</i>	<i>I-supp Diff</i>	<i>p</i>	<i>Group</i>	<i>I-supp Diff</i>	<i>p</i>	<i>Group</i>
SR	0.16	***	Both	0.16	***	Both	0.12	***	Both
+R	0.18	***	Both	0.18	***	Both	-	-	
S+R	0.14	***	Both	0.14	***	Both	0.11	***	Both
V+	0.04	***	Both	-	-		0.03	***	Both
R+	-0.14	***	Both	-	-		-0.042	***	Both
R+M	-	-		0.022	***	Both			
MR+	-	-		0.0192	***	Both	0.0179	***	Both
V+MV	0.003	***	H _{sc}	-	-		-	-	
SV+M	0.001	**	H _{sc}	-	-		-	-	

Differentiable Patterns: Cross Domain

In this section, we describe differentiable sequences across subjects. From descriptive statistics in Table 5.1, we observe the mean SA score is higher in social study assignments (SA First = 2.56, SA Last = 2.62) compared to science (SA First = 2.46, SA Last = 2.58) ones. Additionally, mean MCQ score is higher in science assignments. Thus we compared MCQ and SA attempting behavior between these subjects. In particular, we compared MC_{sc} vs MC_{ss} and SA_{sc} vs SA_{ss} group. Additionally, we compared the H_{sc} and H_{ss} groups to identify high performing student behavior in these two subjects. Table 5.4 presents our results.

H_{sc} vs H_{ss} We begin with our results for the H_{sc} vs H_{ss} comparison. Patterns SR and S+R indicate reading after one (S) or multiple SA submissions. We observe from Table 5.4 these two patterns are more often used by science overall high performers (H) and SA high performer students (SA) compared to social study assignments of the same group (H_{sc} and H_{ss}). Thus, reading the text before attempting SAs lead to higher mean higher SA scores for readings in social studies than readings in science. Similarly, pattern +R is also used more frequently by science students. This pattern indicates

science students read more often after performing consecutive action compared to social studies. The R+ pattern is more frequently used by H_ss group, indicating students' reading behaviors. Only H_sc students exhibited V+MV and SV+M patterns frequently. These patterns are related to vocabulary lookup behavior in MCQ (V+MV) and SA (SV+M) attempts. From these observations, we conclude science students looked up vocabularies more than social study students.

SA_sc vs SA_ss SA_sc and SA_ss groups share similar patterns, except for vocabulary lookups. The relatively lower mean SA score in science can be explained by SR and S+R as discussed above. We observe the difference in frequent patterns for H_sc group students while comparing with domain vs cross domain. RS+ pattern was differentially significant in H_sc group when compared to other groups of science students. In contrast, we do not observe RS+ pattern being frequent while comparing to social studies. Moreover, H_sc group students exhibited less reading activities while compared to H_ss. The i-support difference of R+ pattern conforms this behaviour.

MC_sc vs MC_ss Finally, we analyze MC_sc vs MC_ss group. From the mean i-support difference of R+M pattern, we can say students with science assignments exhibited more reading behavior before attempting MCQ.

5.3 Discussion of RQ2

In RQ2, we aim to understand productive and unproductive students' reading and SRL behaviours and how those behaviors vary *within* and *across* domain. One limitation in our study is we do not have access to confounding variables, such as students demographics, and teachers' instructions to work on the assignments (in-class or homework). To mitigate this limitation, we considered two features Norm_Last and Long_Submissions in our clustering. These two features take into account students' performance with respect to that assignment. Our analysis shows the higher-performing students in science annotate more often and read the text before attempting SAs, which is in-line with high performing students' learning pattern in human biology science learning with Metatutor (Bouchet et al., 2012). Low performing students read and highlighted more for MCQ questions. Despite this fact, they needed more time than their peers to answer MCQs and had lower scores. Our analysis shows higher-performing students in social study assignments read more often before attempting SA and MCQs. Also, they looked up more vocabulary. Vocabulary

lookup helps a student to understand a concept when they come across new words when reading. In contrast, low performing students read after attempting SAs. They also had higher resubmission rate of SA questions followed by read event. Our observed patterns explain the way high and low performing students navigated the SA questions. It appears reading and comprehending the concept prior answering a SA led to score differences. Our findings are in-line with findings of Butler and colleagues (see Table 3, (Butler & Cartier, 2005)). They examined domain specific SRLs with questionnaire with 102 8-th grade Humanities and 25 Information Technology students of grade 9/10. Their classroom study showed that Humanities students reread text more than Information Technology students (58% vs 32%).

5.4 Conclusion and Future Work

Our RQ1.1 examines the association of students' SRL with assignment performance. In RQ2, we examined students' reading and SRL patterns for productive and unproductive students across two subject domains. Our RQ2 findings show that low-performing science students struggled in MCQ attempts. Specifically, when we compared high vs low student groups, V+M pattern was statistically significant in low performing students group— indicating their vocabulary lookup behavior in MCQ attempts. From our RQ1.1 HLM results, we observed that vocabulary lookup behavior was negatively correlated with the MCQ score. Combining our RQ1.1 and RQ2 findings, we can have a better understanding of students' SRL behavior and their question performance.

Our findings may be useful to intervene in low-performing students' SRL behavior. Future work may investigate the generalizability of patterns in another ITS. Another direction of research could be applying temporal pattern mining Chen et al. (2016). Temporal pattern mining will discover time spent on each action sequence by high and low performing students. Findings from the analysis can be used for intervention, such as identifying the most intensive action of low-performing students to scaffold SRLs. However, our analysis of RQ2 only considers question formats, first submission, and resubmissions. We did not take into account the assignment description or the difficulty level of a question. The following Chapter covers question and assignment text-complexity to understand how students adapt their SRL behaviors and connect those behavior to their performance.

CHAPTER

6

SRL AND CONTENT DIFFICULTY

In this Chapter, we discuss question and text difficulty and students' SRL behavior. We answer two sub-RQs

RQ3.1 *How do students' reading and SRL strategies vary with question difficulty?*

RQ3.2 *How do students' reading, question performance, and SRL strategies vary with text complexity and question positioning in the article?*

We split RQ3.2 into two sub-RQs as follows

RQ3.2.1 How do students' score and SRL behavior vary by article text-complexity?

RQ3.2.2 How do students' reading and SRL behavior vary by question text-complexity and question positions within article text-complexity?

In this Chapter, first we will describe background and related literature of RQ3 in Section ?? . Then we will assess question difficulty at the *class-level* context in RQ3.1 in Section 6.2 . SRL researchers emphasized that analyzing SRL requires understanding students' learning contexts (Butler & Cartier, 2005). The context of learning is nested: geographical, socio-economical, within-school, and within-classroom. At the classroom level, students' engagement in learning is shaped by teachers' instructional approaches and by interactions with the teacher and peers. We will compare our proposed approach with the item response theory (IRT) Rasch (1960) approach.

Next, we will describe text-readability of articles and question (Flesch, 1948) and students SRL behavior in RQ3.2 in Section 6.3. Text-readability metrics (Flesch, 1948) measures easiness of reading a text. In addition to question readability, we will also analyze vertical positioning of questions in articles to measure how SRL behavior vary. Finally, we conclude in Section 6.4.

6.1 Background

6.1.1 Background: Question Difficulty

Understanding the difficulty level of test items has wide range of applications in educational data mining (EDM), such as curriculum arrangement (Hsieh & Wang, 2010) and designing sequence of test items in an intelligent tutoring system (ITS) for mastery learning of domain knowledge. Researchers have utilized student response log to identify relationships among test items. Examples include Q-matrix to map questions into skills (Barnes, 2005), learning factor analysis (Cen et al., 2006), Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994), and item response theory (IRT) (Hambleton et al., 1991).

Item Response Theory (IRT) (Hambleton et al., 1991) is regarded as “gold standard” of estimating question difficulties from student response data. The simplest model is named “Rasch Model” (Rasch, 1960), which associates a skill or ability to each student and a difficulty level to each question. Researchers have utilized student attempts coupled with IRT to estimate question difficulty (Ravi & Sosnovsky., 2013). Fouh et al. utilized the total number of attempts and guessing behavior to understand difficult topics in a Data Structure course (Fouh et al., 2016). Additionally, they compared their approach to IRT. Peckham and McCalla (Peckham & McCalla, 2012) examined the connection of reading behaviour and questions difficulty level of Bloom’s taxonomy (Anderson et al., 2000) with 28 grade 12 students in Adult Education English course. They grouped reading, scanning, and scrolling behaviors using k-means clustering into four clusters: Light Reading as 50% reading: 30% scanning: 20% scrolling (50:30:20), Light Medium Reading (60:30:10), Heavy Medium Reading (70:20:10), and Heavy Reading (80:10:10). They observed Light Reading was not found in any question with Bloom’s Level above 3. Students used more Heavy Reading as the Bloom levels increase in difficulty. They also observed students how adopted Heavy Reading for lower level of Bloom’s difficulty did not perform well.

Different intelligent tutoring systems (ITS) (Wenger, 2014) have utilized student logged data to estimate difficulty level of questions. Pardos and Heffernan (Pardos & Heffernan,

2011) extended the BKT model to handle item difficulty in a math tutoring system, ASSISTment. QuizGuide (Sosnovsky et al., 2008), an assessment system for Java programming, predicts subjective difficulty on questions from predefined weights and student performance. The predefined weights are assigned by domain experts. ELM-ART II, a web based Lisp programming tool (Weber & Brusilovsky, 2001), uses fixed difficulty and weight for each item. A student's knowledge level is updated based on correct or incorrect attempts on each item and difficulty level.

Our approach is similar to previous literature in assessing question difficulty from student performance data. However, abovementioned literature examined question difficulty with the goal of students' knowledge tracing, estimating difficult topic in a course and assessing student reading behavior. In contrast, we assess question difficulty as part of reading and SRL behavior analysis.

6.1.2 Background: Text Complexity and Question Positioning

Text Complexity and Formality. The goal of measuring text complexity is to assess how easy or difficult the text is to comprehend. Earlier attempts to measure text complexity relied primarily on syntactic and semantic variations in the text. Examples of such approaches include the Flesch Reading Ease Score (FRES) (Flesch, 1948), the Dale-Chall Readability formula (Chall & Dale, 1995), and the Flesch-Kincaid (FK) Grade Level (Kincaid et al., 1975). All three approaches use the average sentence length to assess syntactic complexity. To compute semantic complexity, FRES and FK Grade Level use syllables per word. The Dale-Chall formula assesses semantic difficulty as a measure of word familiarity to readers.

Recently, research progress in several interdisciplinary communities, including computational linguistics [25, 33] and discourse processing (Jurasky & Martin, 2000; Kintsch & Van Dijk, 1978), have enabled researchers to provide more detailed analysis of text complexity and reading comprehension. Deeper levels of reading comprehension include interpreting sentence meaning from words and connecting to background knowledge (Graesser et al., 1994; Graesser & McNamara, 2011; Snow, 2002*b*). Examples of text complexity using deeper comprehension levels is Coh-Metrix (Graesser, McNamara, Louwerse & Cai, 2004). The web version of Coh-Metrix provides 108 metrics to analyze text (see Dowell et al. (2016) for details). One such measure is text cohesion. Cohesion refers to text features (grammatical, lexical) those connect together to convey ideas presented in the text. Cohe-

sions communicate readers about ideas across sentences via explicit cues (such as because, and, or). Text-cohesion features has been shown an important indicator for reading comprehension task (Tapiero, 2007). Another example of automated text complexity analysis tool is TEXTEVALUATOR (Sheehan et al., 2014; Sheehan, 2015), developed by researchers at the educational testing services (ETS) ¹. The TEXTEVALUATOR tool was developed to measure accelerated text complexity metrics stated by the new Common Core Standard (CCS) (CCS, 2010). ETS researchers first collected Standard-aligned reading text materials. Next, expert educators determined the text complexity of reading materials and the process was automated. An analysis conducted by Sheehan and colleagues (Sheehan et al., 2014) showed that text complexity measured by TEXTEVALUATOR aligns with the new Common Core Standards.

To define a unified metric to assess text complexity, the Coh-Metrix team derived a unified term “Coh-Metrix Formality” (Graesser et al., 2014). Formality is a stylistic measure to assess how rigidly structured the text is. Formal texts are composed of careful word and sentence constructions (Jack C. Richards and Platt, 1992), such as law. Informal text has similar structure to spoken or conversational language (Graesser et al., 2014), such as daily conversation. In earlier research, researchers have measured formality of a text at word or phrase levels. Heylighen and Dewaele (Heylighen & Dewaele, 2002) defined an F-score measure for formality which relies on parts of speech of words. The F-score formality increases with higher frequency of nouns, adjectives, articles, and prepositions and a lower frequency of pronoun, adverb, and interjections.

In contrast, Graesser and colleagues derived Coh-Metrix formality from five dimensions of principal component analysis of Coh-Metrix output (Graesser et al., 2014). These five dimensions are (i) *Narrativity*: measures the extent to which text describes a story in terms of character, places, and things; (ii) *Syntactic Simplicity*: measures the extent to which the text contains simpler sentences; (iii) *Word Correctness*: measures the extent to which the text contains concrete words conveying explicit meaning; (iv) *Referential Cohesion*: measures the extent to which the text contains words and ideas overlap across sentences, and (iv) *Deep Cohesion*: measures the extent to which the text contains temporal, causal connectivities. Graesser and colleagues reported the Coh-Metrix formality has high correlation with text difficulty measures (Graesser et al., 2014; Dowell et al., 2016). Formal language has higher referential and deep cohesion and informal texts have more narrativity, syntactic simplicity, and word correctness.

¹<https://www.ets.org/>

Vertical Positioning and Reading. Reading behaviors and vertical text positioning has been widely studied. Researchers in this domain have focused on two concepts: fixation and saccade. *Fixation* refers to gazing at a single position for a brief period while *saccade* refers to changing fixation from one position to another (Reichle et al., 2003).

Li and colleagues found that users' fixation rate decays with vertical positioning in the text (Li et al., 2018). Forrin and colleagues investigated the "*TL;DR*" effect in reading behavior (Forrin et al., 2020). The authors found that longer text spans had more mind wandering behavior and also lower performance scores. In another study, Goedecke et al. (Goedecke et al., 2015) measured how reading engagement varies with text complexity measured by Coh-Metrix formality. They analyzed text complexity in three genres: informational, narrative, and persuasive. Informational texts had the highest formality score, followed by persuasive and narrative. Experiments with 254 Amazon mechanical turk (AMT) participants showed that reading disengagement was more related to vertical positioning on the text and genre rather than reading speed of participants.

Vega and colleagues investigated the relationship between reading time and text complexity with 64 participants in AMT (Vega et al., 2013). Authors defined a compound term - decoupling rate combining the FK grade level and reading time. The decoupling rate measured whether participants were allocating reading time depending on the text complexity. Their study showed that individual reading pace has connection with decoupling rate and text complexity.

AL articles used in our study were presented with embedded questions. We measured the text complexity at two levels: text complexity of articles and text complexity of questions. To assess text complexity of articles, we computed the FRES, FK grade level, and two measures of formality. For text complexity of questions, we compute FRES score only, as formality measures do not produce reliable measures in shorter texts (Heylighen & Dewaele, 2002). We assess how students' reading and SRL vary with article text complexity, question text complexity, and vertical positioning of questions.

6.2 SRL and Question Difficulty

In this section, we present methodology and results of *RQ3.1 How do students' reading and SRL strategies vary with question difficulty?*

6.2.1 Dataset

From the science dataset in Section 3.2, we selected 131 predefined AL questions used in at least two classes. The final resulting dataset has 11,832 students and 913 assignments used in 641 classes.

6.2.2 Methodology of RQ3.1

We utilize log trace data to assess how students across class-level may perceive question difficulty. We compare our proposed approach with IRT analysis. Next, we examine how reading and SRL are associated with question difficulty.

Question Difficulty and Student Performance: Student Interaction Data

We analyzed students' performance on each question to assess the difficulty of the question. To calculate a student's performance, we took the ratio of max score achieved to number of attempts on a question. Questions in AL are graded on a scale [0–4]. For our assessment, we scaled students' scores in [0–1] range. We defined the performance of a student i on a question q as

$$r_i = \frac{\text{scaled maximum score on } q}{\text{number of attempts on } q} = \frac{\text{maximum score on } q/4}{\text{number of attempts on } q} \quad (6.1)$$

Equation 6.1 computes (scaled) maximum score per attempt for each student. The value of r_i is in the range of zero to one, one representing good performance and zero representing a poor one.

We computed difficulty level (dl) of question q for all students in a class as

$$dl = 1 - \frac{\sum_{i=1}^n r_i}{n} \quad (6.2)$$

where n is the number of students in a class attempted q and r is the students' performance on q as defined by Equation 6.1. A value of $dl \sim 0$ value indicates an easy question and $dl \sim 1$ indicates a difficult one.

A predefined AL question can be used in multiple assignments assigned by class teachers. Teachers can use questions for in class reading, homework assignments, or extra reading. As we do not have information on how questions were used, we decided to compute question difficulty by *class level*. To analyze the difficulty of a question q across classes, we computed difficulty ratio of q across classes as follows:

$$\text{Difficulty ratio of question } q = \frac{\text{number of classes with } dl \geq 0.5 \text{ for } q}{\text{number of classes used } q \text{ in assignments}} \quad (6.3)$$

Equation 6.3 measures ratio of number of classes found a question hard ($dl > 0.5$) to number classes used the question.

We plotted histograms of difficult ratio for 131 questions. After examining the histogram, we observed more questions with difficulty ratio < 0.2 and fewer questions with difficulty ratio > 0.5 . We grouped questions into three categories by their difficulty ratio as shown in Table 6.1.

Table 6.1: Question category by difficulty ratio

Category	MCQ	SA	Total
Easy (<i>difficulty ratio</i> < 0.4)	6	75	81
Medium ($0.4 \leq \textit{difficulty ratio} < 0.6$)	5	26	31
Hard (<i>difficulty ratio</i> > 0.6)	11	8	19

Question Difficulty and Student Performance: IRT Analysis

The IRT method estimates the probability of a student will answer a question correctly with respect to item difficulty and student's ability. We applied the 1-parameter logistic IRT model (1PL) model, also known as the Rasch model (Rasch, 1960). The 1PL model describes test items considering only one parameter, item difficulty, b . Item difficulty is defined as how hard an item is. The 1PL model is a logistic curve, i.e., it evaluates how high a student's latent ability level needs to be in order to get a 50% chance of getting the item right.

To apply the 1PL model, we need to map students' responses on questions (test item) to zero or one computed in Equation 6.1. For this purpose, we assigned zero if $r < 0.5$ and 1 otherwise. We fit the 1PL model to 131 questions using the 'ltm' package in R (Rizopoulos, 2006).

What SRL Did Students Use Preceding and Following Each Question ?

We split students' actions into sessions following the same procedure in Section 5.2.1. Next, we counted reading and SRL activities (R, H, A, V) prior and after each question submission.

We counted these four features for Easy, Medium, and Hard questions separately.

We performed the above analysis at two levels: student-level and class-level. For class-level analysis, we took mean SRL value of a class on each question. Table 6.2 and 6.3 present mean and standard deviation of student and class level, respectively.

To test if there existed significant difference in means, we applied the non-parametric Kruskal-Wallis test Kruskal & Wallis (1952). In case with significant differences in mean, we performed a post-hoc Dunn test with Benjamini-Hochberg correction to identify pairwise significant groups using R package `dunn.test` (Dinno, 2017). We computed effect size (r) using a non-parametric test, Cliff's Delta (Cliff, 1993), for SRL metrics that showed significantly different between groups.

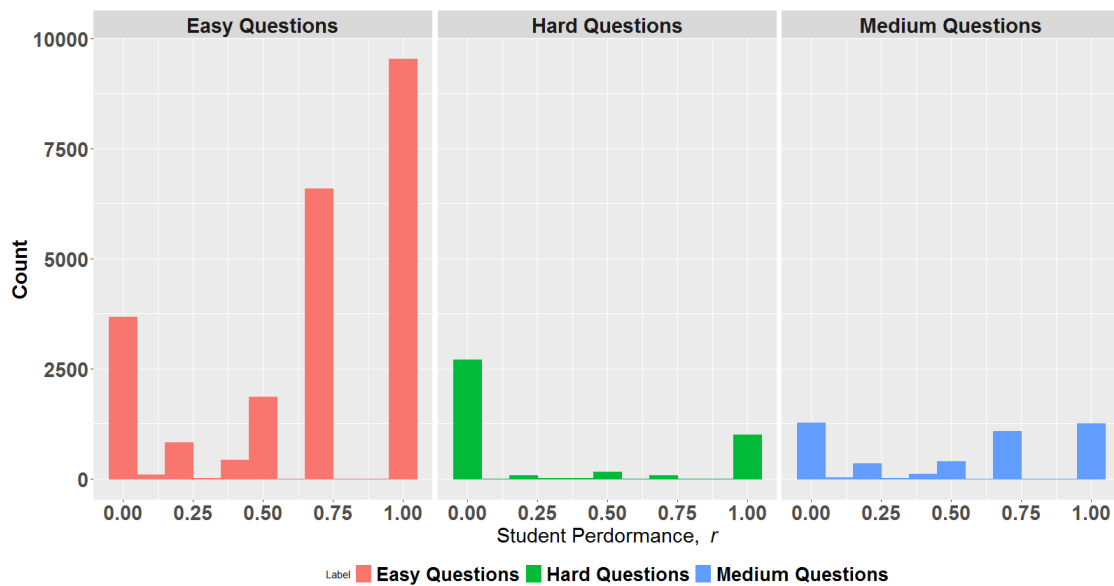


Figure 6.1: Student performance distribution by question difficulty

6.2.3 Result of RQ3.1

Question Difficulty and Student Performance

We plotted item characteristics curve (ICC) and item information curves (IIC) from the fitted IRT model. Figure 6.2 presents ICC and IIC curves. Each line of the ICC and IIC plot represents one question.

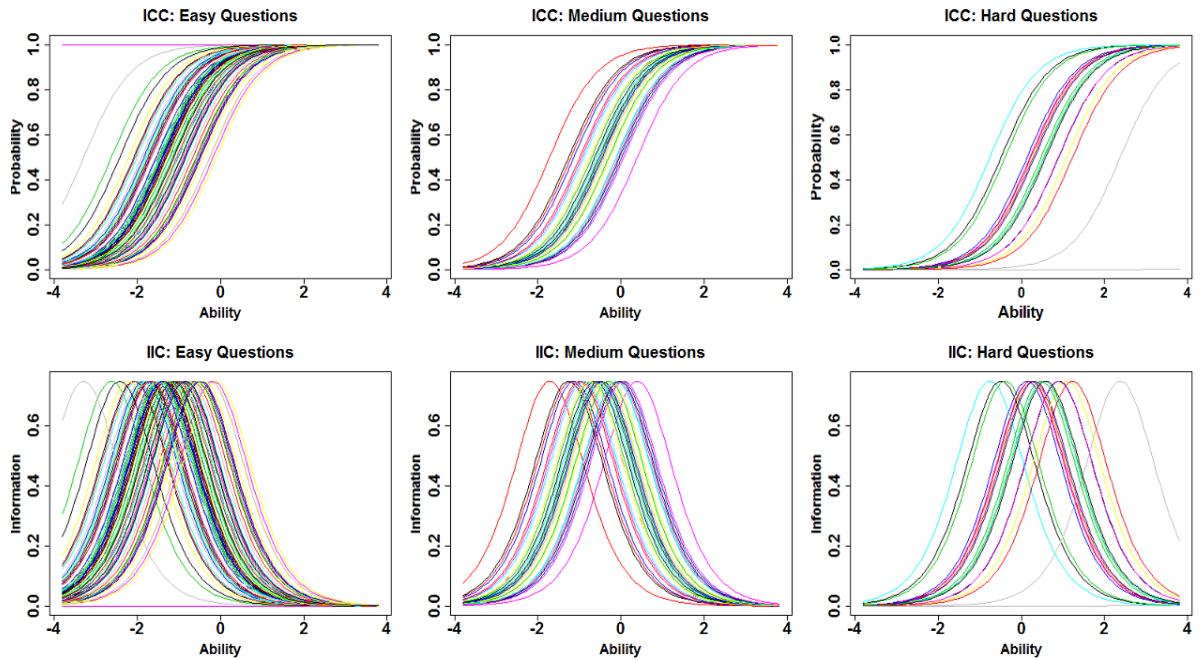


Figure 6.2: ICC and IIC plots from 1PL model

In ICC plot, the X axis represents students' latent ability and the Y axis represents the probability of students will answer the question correctly. The range of X axis is $[-4, 4]$ and $X = 0$ indicates average ability. We plotted ICC curves for Easy, Medium, and Hard questions separately to compare our approach with IRT two approaches. We observe that the ICC curves for Easy questions are mostly on the left side of zero, indicating Easy questions required lower ability for correct attempts. Comparing ICC curves of Easy and Hard questions, we notice Hard questions have curves more on the right side on X axis. The probability of answering a Hard question correctly decreases as curves go from left to right.

The IIC curves shows how much information about students' ability an item provides. A difficult item will provide little information about a student with low ability and vice versa for easy items. We plotted IIC curves for Easy, Medium, and Hard questions (Table 6.1) separately. From Figure 6.2, we observe Easy questions curves provide information about students with average and below average abilities (the peak of curves are mostly on the left side of $X = 0$. $X = 0$ refers to average ability). Similarly, IIC curves for Hard questions provide information about high ability (the peak of curves are mostly on the right side of $X = 0$) levels.

Question Difficulty and SRL: Student-level

Table 6.2: Mean with (Standard Deviation), and p value from KW = Kruskal-Wallis test for student behavior features on Easy, Medium, and Hard questions.

Feature	Easy	Medium	Hard	KW- <i>p</i>
Reading (R)	0.70 (0.71)	0.85 (0.74)	1.30 (0.68)	<0.001
Annotating (A)	0.034 (0.20)	0.021 (0.17)	0.012 (0.12)	<0.001
Highlighting (H)	0.007 (0.10)	0.004 (0.06)	0.002 (0.05)	0.003
Vocabulary lookup (V)	0.016 (0.13)	0.014 (0.12)	0.009 (0.1)	0.01956

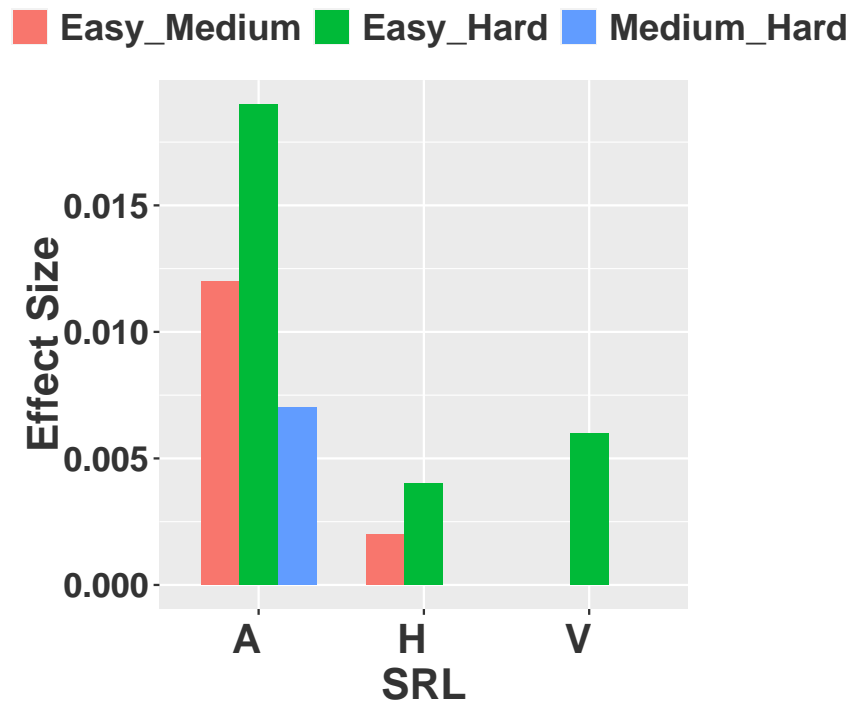


Figure 6.3: Effect Size Comparison of SRL among three pairs at student-level. A: Annotating, H: Highlighting, and V: Vocabulary lookup

The goal of the student-level analysis is to assess how students' reading and SRL behaviors vary by question type. As Table 6.2 shows, the mean of all features varied at statistically

significant levels across the three categories of questions. The number of reading activities was the highest for the Hard questions, followed by Medium and Easy. This indicates students had to read more before attempting a Hard question. It also indicates that they revisited the reading material after attempting a Hard question more frequently than they did for Easy and Medium questions. Annotating, highlighting, and vocabulary lookup counts were higher in Easy and Medium questions than Hard ones. We report the Dunn test and significant pairs for each feature below. For the reading feature (R), we found statistically significant differences among all three pairs Easy-Hard, Easy-Medium, and Medium-Hard. The p values of these pairs were ($p < 0.001, r = 0.43$), Easy-Medium ($p < 0.001, r = 0.11$), Medium-Hard ($p < 0.001, r = 0.33$).

When we consider the annotating feature (A), we also found statistically significant differences in means among all three pairs. Easy-Hard, Easy-Medium and Hard-Medium pairs had ($p < 0.001, r = 0.11$), ($p < 0.001, r = 0.012$), and ($p < 0.001, r = 0.007$), respectively.

And, when considering the highlighting feature (H), we found two pairs differed at statistically significant levels: Easy-Hard ($p = 0.004, r = 0.004$) and Easy-Medium ($p = 0.003, r = 0.003$).

Finally, for the vocabulary lookup (V) feature, we found one pair with statistically significant difference: Easy-Hard ($p = 0.008, r = 0.006$). Figure 6.3 shows the effect size of SRL across three pairs. From Figure 6.3, we observe that the effect size Easy-Hard pair is greater compared to the rest two pairs for all three SRL. We also observe only one SRL; vocabulary lookup differed at a statistically significant level for the Medium-Hard pair. .

Question Difficulty and SRL: Class-level

Table 6.3: Mean with (Standard Deviation), and p value from KW = Kruskal-Wallis test for class-level features on Easy, Medium, and Hard questions.

Feature	Easy	Medium	Hard	KW- p
Reading (R)	0.71 (0.50)	0.93 (0.50)	1.29 (0.47)	<0.001
Annotating (A)	0.03 (0.11)	0.02 (0.07)	0.01 (0.06)	0.0005
Highlighting (H)	0.007 (0.03)	0.003 (0.01)	0.002 (0.01)	0.03
Vocabulary lookup (V)	0.02 (0.08)	0.01 (0.07)	0.01 (0.06)	0.17

From Table 6.3, we observe the mean of all features except vocabulary lookup varied at statistically significant levels across the three categories of questions.

For the reading feature (R), we found statistically significant differences among all three pairs Easy-Hard, Easy-Medium, and Medium-Hard. The p values of these pairs were Easy-Hard ($p < 0.001$, $r = 0.60$), Easy-Medium ($p < 0.001$, $r = 0.24$), Medium-Hard ($p < 0.001$, $r = 0.41$).

When we consider the annotating feature (A), we found statistically significant differences in means among one pair: Easy-Hard ($p = 0.0004$, $r = 0.08$). Compared to student-level analysis, we did not observe a significant difference between Medium-Hard and Easy-Medium groups.

For highlighting (H) feature, we found significant difference in one pair: Easy-Hard ($p = 0.02$, $r = 0.04$).

To summarize, SRL difference Medium-Hard was not significant in the class-level analysis.

6.2.4 Discussion of RQ3.1

In this section, we present our findings of RQ3.1.

In RQ3.1, we used a data-driven approach on *class-level* student response data to group questions by difficulty levels. Our difficulty levels are consistent with findings from IRT analysis. ICC curves for Easy questions require lower student ability (Figure 6.2) and vice versa for Hard questions. Table 6.1. shows 11 MCQ questions belonged to the Hard category. We looked into the question texts and observed 10 out of 11 questions required students to select multiple options, i.e., ‘Select all that apply’. Our analysis of question difficulty and SRL indicates that students performed reading (R) more before and after answering Hard questions. Thus, our findings indicate that although students can rule out distractors in MCQs (Bliss, 1980), answering such questions is Hard when asked to select multiple correct answers. Table 6.1 shows 75 out of 81 Easy questions are SA. Our results also indicate students annotated (A), highlighted (H), and looked up vocabularies (V) more in answering Easy questions. We conclude the format of questions contributes to students’ SRL usage, even if the difficulty level is Easy. Ideally, we would have been able to control question format and student characteristics; secondary data mining allows for large-scale data, but the precision of results can be compromised by lack of these details. Nevertheless, we were able to demonstrate that SRL behaviors covary with question difficulty and/or format. It seems likely that as students encountered SA questions, they received metacognitive signals that encouraged their use of SRL behaviors and resulting in the relatively greater success of these questions (O’Neil Jr & Brown, 1998).

6.3 SRL and Text-Complexity

In this Section, we answer **RQ3.2** *How do students' reading, question performance, and SRL strategies vary with text complexity and question positioning in the article?* To conduct our study, first, we assess the text complexity of the reading article. Next, we calculate question positioning and text complexity of questions in articles. Then, we evaluate how students' reading, question score, and SRL behavior vary with article text complexity, question text complexity, and question positioning in the article.

Table 6.4: RQ3.2 Methodology Overview

	Single-Article Assignments	Multi-Article Assignments
RQ3.2.1		
Descriptive	42 articles in 881 assign.	59 Assign. Mean (SD) articles per assign. = 5.22 (2.97)
Article Text Complexity	<ul style="list-style-type: none"> • Compute 42 articles' text complexity • Map each assign. to an article 	<ul style="list-style-type: none"> • Combine all articles in a assign. • Compute assign.'s text complexity
Ques. Positioning	Vertical Ques. positioning in the article	Vertical Ques. positioning in the <i>combined articles</i> of an assign.
Ques. Text Complexity	Ques. in the article	Ques. in an assign.
RQ3.2.2 Connecting Text Complexity with Students' Ques. Score, Reading, and SRL		
Article Text Complexity	Compare single vs Multi-article assignment	
Ques. Positioning	Correlation analysis in both single and Multi-article assignment	
Ques. Text Complexity	Correlation analysis in both single and Multi-article assignment	

Teachers can use one reading article or combine multiple reading articles to create an AL assignment. In our science dataset in Section 3.2, 881 assignments used a single article and 59 assignments used more than one article. For multi-article assignments, we combined all articles in that assignment for our analysis. Table 6.4 presents an overview of our RQ3.2 methodology.

6.3.1 Methodology of RQ3.2.1

Text Readability and Articles

We computed the Flesch reading-ease score (FRES) for single and multi-article assignments using the science dataset of Section 3.2. The FRES is defined as follows

$$206.835 - 1.015 \frac{\text{total words}}{\text{total sentences}} - 84.6 \frac{\text{total syllables}}{\text{total words}} \quad (6.4)$$

The FRES ranges from [0-100]. A score of zero indicates extremely difficult to read, and a score of 100 indicates very easy to read. To assess text complexity, we measured three metrics: Flesch–Kincaid (FK) grade level and two formality measures. Formality scores represent the formal language level of a text. The more formal a text, the more difficult it is to comprehend.

1. **Flesch–Kincaid (FK) Grade Level** The FK grade level (Kincaid et al., 1975) reports a number corresponding to the U.S. grade level.
2. **F-score Formality** Heylighen and Dewaele (Heylighen & Dewaele, 2002) defined a formality measure, F-score, based on parts of speech frequency in the text. F-score is calculated as follows

$$\begin{aligned} \text{F-score} = & (\text{Noun freq.} + \text{Adj. freq.} + \text{Preposition freq.} + \text{Article freq.} \\ & - \text{Pronoun freq.} - \text{Verb freq.} - \text{Adverb freq.} - \text{Interjection freq.} + 100)/2 \quad (6.5) \end{aligned}$$

3. **Coh-Metrix Formality:** Coh-Metrix formality (Graesser et al., 2014) score is calculated from five dimensions' z-percentile scores. The formula is as follows

$$\begin{aligned} \text{Coh-Metrix Formality} = & [\text{Referential Cohesion} + \text{Deep Cohesion} - \text{Narrativity} \\ & - \text{Syntactic Simplicity} - \text{Word Correctness}]/5 \quad (6.6) \end{aligned}$$

Single-Article Assignments. Figure 6.4 shows the histogram of FRES of 42 articles used in single assignment.

Observing the Figure 6.4 histogram, we split 42 articles into three groups: Easy-Read, Medium-Read, and Hard-Read as follows

- **Hard-readability:** FRES < 60
- **Medium-readability:** 60 <= FRES <= 75
- **Easy-readability:** FRES > 75

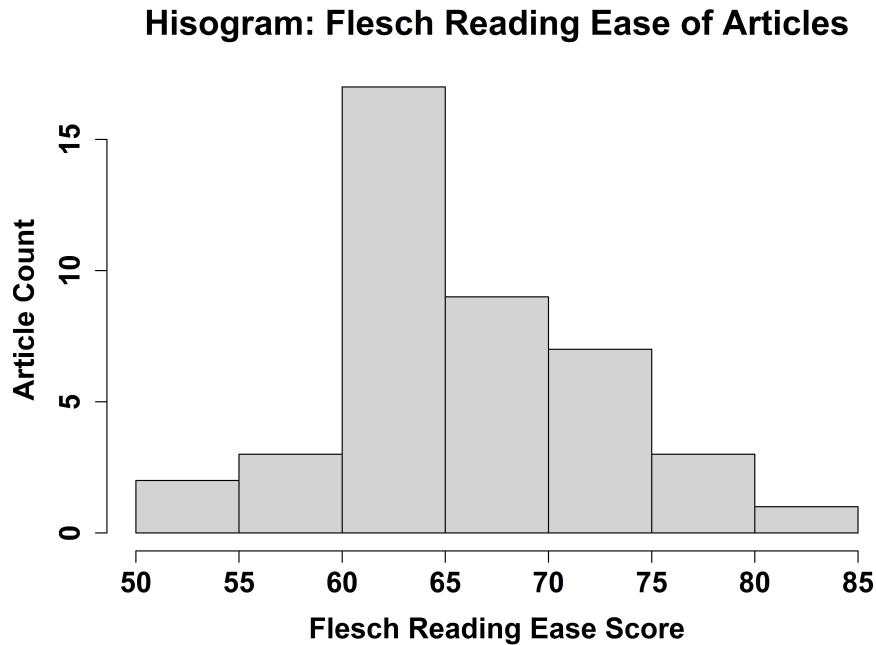


Figure 6.4: Article Category by Flesch Readability: Single-Article Per Assignment

Descriptive statistics of these three categories are in Table 6.5.

Multiple-Article Assignments. Table 6.6 presents descriptive statistics of multiple-articles per assignments.

We calculated FRES and FK grade level using the Python package, Textstat (Textstat, 2020). We used the R package qdap (Tyler Rinker, 2020) to calculate F-score formality. We used the Coh-Metrix Web tool version 3.0² to get Coh-Metrix measures of 42 articles and calculated Coh-Metrix formality using Equation 6.6. We conducted the Kruskal-Wallis test (Kruskal & Wallis, 1952) on FK grade level and F-score to identify how text complexity of articles varies among the four groups. We applied the Coh-Metrix formality to only 42 articles used in single-article assignments, as some multiple-article assignments exceed 15,000 character limit suggested by the Coh-Metrix website.

Score, SRL and Article Text Readability

In this section, we describe our methodology to calculate students' question performance and SRL among four groups: three groups with a single-article category and one group with

²<http://cohmetrix.com/>

Table 6.5: Descriptives of Article Category by FRES. Mean (SD) of features. n = No. of articles in each category.

Feature	Article Grouped by FRES: Single Article Per Assignment		
	Hard-Read (n = 5)	Medium-Read (n = 33)	Easy-Read (n = 4)
Sentences	44.20 (9.0)	43.58 (12.56)	41 (11.58)
Words	596.80 (119.63)	634.64 (172.5)	601(177.23)
Words Per Sentence	13.68 (2.04)	14.71 (1.45)	14.83 (2.21)
Readability Measure			
FRES	56.83(2.13)	65.88 (4.02)	78.97 (2.88)
An Article Used in Assignments			
Mean (SD)	15.2 (21.94)	22.90 (53.87)	12.25 (6.13)
Questions Per Assign.	1.80 (0.83)	3.45 (2.18)	2.00(1.16)

Table 6.6: Descriptives of Multiple Articles per Assignment. Mean (SD) and Median of features. n = No. of articles in each category.

Feature	Multi-Article Assignment
	(n = 59)
Sentences	308.95 (163.38), Median = 335
Words	4505.93 (2449.48), Median = 4739.00
Words Per Sentence	14.48 (0.76), Median = 14.40
Readability Measure	
FRES	65.89 (3.21), Median = 64.20
Article and Question Statistics	
Articles Per assign.	5.22 (2.97) Median = 5
Questions Per Assign.	8.88 (8.45) Median = 7
MCQ Per Assign.	3.88 (2.22) Median = 4
SA Per Assign.	5.00 (8.76) Median = 2

multiple articles. To compute reading and SRL behavior within each group, first, we split students' interactions into sessions following the same procedure in Section 5.2.1. After defining the session, we counted reading and SRL events for each question attempt. For this purpose, we calculated the number of SRL and readings in each session before the question attempt.

To understand the score difference among each group, we considered the score of each question attempt within that group. We applied the Kruskal-Wallis test (Kruskal & Wallis, 1952) to determine if there exist statistically significant differences in the score, reading, and SRL behavior among groups. To identify which pairs were statistically significant, we applied the Dunn test (Dinno, 2017). We also report effect size using a non-parametric test, Cliff's-delta, (Cliff, 1993) between pairs showing statistically significant differences.

6.3.2 Methodology of RQ3.2.2

Question Position and SRL

To calculate the vertical positioning of a question, we calculated the ratio of the question's beginning index to total character count in the article. This resulted in question positioning in a [0-1] range, where zero indicating at the beginning and 1 indicating at the end of the article. Table 6.7 presents descriptive statistics of vertical positioning of questions in all assignments grouped by FRES article category of RQ3.2.1.

We calculated Spearman's correlation (Spearman, 1961) between students' reading and SRL and question positions in three groups of articles of RQ3.2.1. To measure observed reading and SRL behavior, we computed observed behavior in each session before the question attempt—similar approach as in Section 6.3.1

Table 6.7: Vertical Positioning of Questions in Assignments Grouped by FRES.

Vertical Position	Single Article Per Assignment			Multiple Articles Per Assignment
	<i>Hard-Read</i>	<i>Medium-Read</i>	<i>Easy-Read</i>	
0 - < 0.25	44	429	44	179
0.25 - < 0.50	55	424	22	124
0.50 - < 0.75	34	331	59	101
0.75 - 1	47	2,335	72	118
Total Ques. in Assignments	180 Ques. in 76 Assign	3,519 Ques in 756 Assign.	197 Ques. 49 Assign.	524 Ques. in 59 Assigns.

Question Readability and SRL

We calculated question readability similarly to article readability. We computed FRES as question readability. Table 6.8 presents descriptive statistics of question readability of all assignments grouped by article category. Then, we calculated the Spearman's correlation between question readability and SRL behavior within each group of articles in RQ1. The approach is the same as described in Section 6.3.2.

Table 6.8: Question Readability in Assignments Grouped by FRES (Lower value ~ Difficult Readability)

Question FRES	Single Article Per Assignment			Multiple Articles Per Assignment
	Hard-Read	Medium-Read	Easy-Read	
0 - < 25	0	596	0	29
25 - < 50	21	567	6	75
50 - < 75	127	1,706	66	285
75 - 100	32	650	125	135
Total Ques. in Assignments	180 Ques. in 76 Assigns.	3,519 Ques in 756 Assigns.	197 Ques. 49 Assigns.	524 Ques. in 59 Assigns.

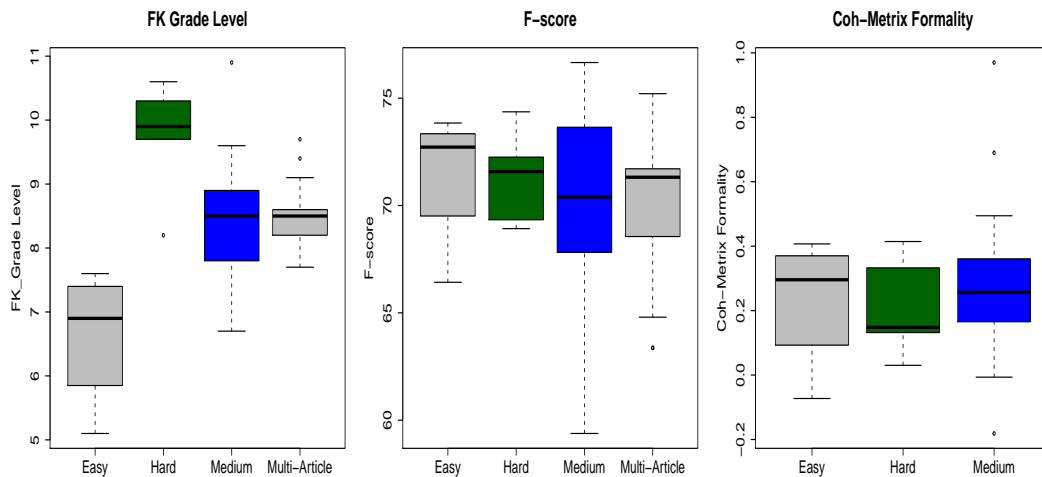


Figure 6.5: Formality and FK Grade Levels in RQ3.2.1

6.3.3 Results of RQ3.2.1

Text Complexity Analysis

Figure 6.5 presents boxplots of article complexity in terms of FK grade level and F-score formality among four groups of RQ3.2.1.

The F-score formality measure was not statistically significant ($p = 0.62$) among four groups of articles. Similarly, we do not find statistically significant differences in Coh-Matrix formality measure among three article groups ($p = 0.83$).

FK Grade level was statistically significant among four groups ($p = 0.00067$). Post-hoc Dunn test identified significant differences among all three pairs. The effect size and p-values of three pairs are: Easy-Read–Hard-Read ($p = 0.0001$, $r = 1$), Easy-Read–Medium-Read ($p = 0.003$, $r = 0.87$), Medium-Read–Hard-Read ($p = 0.01$, $r = 0.70$). When we consid-

ered single-article vs. multi-article category, we found two groups with FK Grade differing at a statistical significant level: Easy-Read–Multi-Article ($p = 0.003, r = 1.00$) and Hard-Read–Multi-Article ($p = 0.008, r = 0.70$).

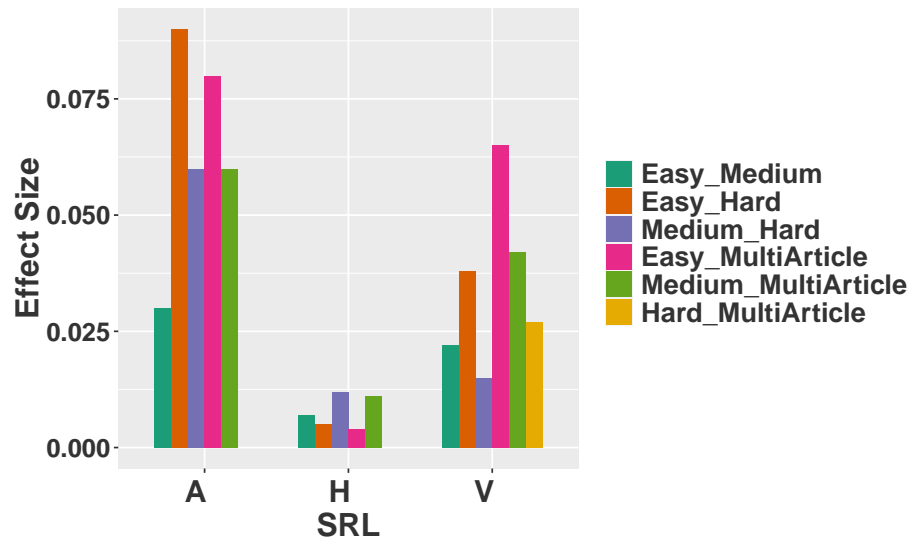


Figure 6.6: SRL and Article Text Complexity

Score, SRL and Text Readability

The Kruskal-Wallis result on the score was significant ($p < 0.001$) among four groups. Next, we applied the Dunn test to identify pairwise groups those differed at a statistically significant level by the Dunn test. We have six pairwise groups: three groups with single article assignments: Easy-Read – Medium-Read, Easy-Read–Hard-Read, and Medium-Read–Hard-Read. Rest three groups are Multi-article assignments paired with each of the Easy-read, Medium-Read, and Hard-Read categories. We report p -value and non parametric effect size (r), Cliff’s-Delta of those groups below. (Cliff, 1993) of pairs. Figure 6.7 and 6.6 shows plots of effect sizes among six groups.

First, we describe the question score variable. Considering pairs in single article assignments, the Easy-Read-Medium-Read had the lowest effect size ($p < 0.001, r = 0.17$). The

highest effect size was observed in Medium-Read-Hard-Read article pairs ($p < 0.001$, $r = 0.33$). The Easy-Read-Hard-Read pair had effect size in between of the above two pairs ($p < 0.001$, $r = 0.22$). When we consider pairs with multi-article assignments, we found the highest effect size in Medium-Read-Multi-Article ($p < 0.001$, $r = 0.22$), followed by Easy-Read-Multi-Article ($p < 0.001$, $r = 0.10$) and Hard-Read-Multi-Article ($p < 0.001$, $r = 0.09$).

Considering reading in single article category, we observed the highest effect size on the Easy-Read vs Medium-Read group ($p < 0.001$, $r = 0.31$), followed by the Medium-Read vs. Hard-Read group ($p < 0.001$, $r = 0.27$). The Easy-Read vs Hard-Read group had the lowest effect size ($p = 0.0006$, $r = 0.04$). When we consider pairs with multi-article assignments, we found the highest effect size in Medium-Read-Multi-Article ($p < 0.001$, $r = 0.56$), followed by Hard-Read-Multi-Article ($p < 0.001$, $r = 0.41$) and Easy-Read-Multi-Article ($p < 0.001$, $r = 0.39$).

In the single article assignment category, the annotation behavior (A) had the maximum effect size in the Easy-Read-Hard-Read group ($p < 0.001$, $r = 0.09$), followed by the Medium-Read vs Hard-Read pair ($p < 0.001$, $r = 0.06$), and Easy-Read vs Medium-Read pair ($p < 0.001$, $r = 0.03$). We found two pairs in Multi-Article assignments showing statistical significant difference in annotation: Easy-Read-Multi-Article ($p < 0.001$, $r = 0.08$) and Medium-Read-Multi-Article ($p < 0.001$, $r = 0.06$).

When we considered the highlighting behavior, the highest effect size was observed in Medium-Read vs Hard-Read group ($p < 0.001$, $r = 0.012$), followed by Easy-Read vs Medium-Read group ($p = 0.0001$, $r = 0.007$), and Easy-Read vs Hard-Read group ($p = 0.0190$, $r = 0.005$). We found two pairs in Multi-Article assignments showing statistical significant difference in highlighting: Easy-Read-Multi-Article ($p = 0.0212$, $r = 0.004$) and Medium-Read-Multi-Article ($p < 0.001$, $r = 0.011$).

Finally, the vocabulary lookup behavior had the highest effect size in Easy-Hard pair ($p < 0.001$, $r = 0.04$), followed by Easy-Medium ($p < 0.001$, $r = 0.02$), and Medium-Hard ($p < 0.001$, $r = 0.015$). When we consider pairs with multi-article assignments, we found the highest effect size in Easy-Read-Multi-Article ($p < 0.001$, $r = 0.065$) followed by Medium-Read-Multi-Article ($p < 0.001$, $r = 0.042$) and Hard-Read-Multi-Article ($p < 0.001$, $r = 0.027$).

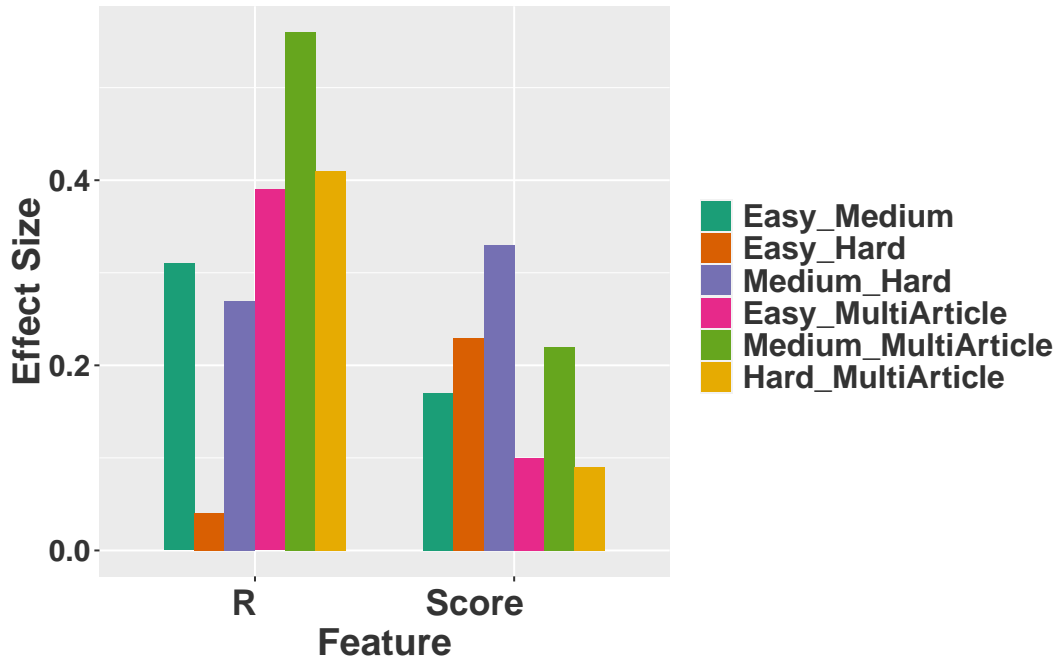


Figure 6.7: Score, Reading and Article Text Complexity

6.3.4 Results of RQ3.2.2

Question Position and SRL

Table 6.9 shows Spearman’s correlation results between observed students’ behavior (i.e, reading, SRL, and score) and question position in the text. A negative value indicates observed behavior decreased as the vertical question positioning increased (from up to down) and vice versa. We discuss results of single article per assignment followed by multiple articles per assignment.

Single Article Per Assignment. Reading behavior (R) was statistically significant in all groups. For Medium-Read articles, students’ reading behavior decreased as the vertical question position increased. For articles in the Easy-Read and Hard-Read group, reading behavior increased as the vertical position increased.

For annotation (A) behavior, we observed a statistically significant correlation with Medium and Easy-Read article groups. In both of these article categories, students showed more annotated behavior as the vertical question positioning increased.

Table 6.9: Spearman’s Correlation between Question Position and SRL by FRES. Bold = Statistically Significant. R =No. read, A=No. Annotation, H= No. Highlight, V= No. Vocab. lookup

Feature	Single Article Per Assignment			Multiple Articles Per Assignment
	<i>Hard-Read</i>	<i>Medium-Read</i>	<i>Easy-Read</i>	
R	0.22 ($p < 0.001$)	-0.31 ($p < 0.001$)	0.10 ($p < 0.001$)	0.30 ($p < 0.001$)
A	-0.02 ($p = 0.15$)	0.02 ($p < 0.001$)	0.07 ($p < 0.001$)	-0.002 ($p = 0.83$)
H	0.01 ($p = 0.48$)	-0.01 ($p = 0.006$)	-0.001 ($p = 0.92$)	0.02 ($p = 0.20$)
V	-0.07 ($p < 0.001$)	-0.06 ($p < 0.001$)	-0.004 ($p = 0.76$)	0.03 ($p = 0.0129$)
Score	0.29 ($p < 0.001$)	0.19 ($p < 0.001$)	-0.07 ($p < 0.001$)	-0.03 ($p = 0.02$)

For highlighting (H) behaviors, we found the Medium-Read group was statistically significant. Highlighting behavior decreased as the question position increased. The vocabulary lookup (V) behavior was statistically significant in Hard-Read and Medium-Read article groups. In both of these groups, students looked up less vocabulary as the question position increased. Finally, we found the question score statistically significant in all FRES groups of articles. Students’ performance of questions increased in the Hard and Medium-Read article categories as the vertical positioning increased. However, for Easy-Read articles, students’ performance decreased as the vertical question positioning increased.

Multiple Articles Per Assignment. For assignments with multiple articles, we found only reading (R), vocabulary lookup (V), and question score (Score) showed a statistically significant correlation with question positioning. Students’ scores on questions decreased for questions in the latter part of the assignment. However, students showed more reading and vocabulary behavior as vertical question positioning increased.

Question Readability and SRL

Table 6.10 shows the Spearman’s correlation results between observed students’ behavior (i.e., reading, SRL, and score) and question FRES readability. A positive value indicates the observed behavior increases for relatively easy questions (higher FRES value for question text) and vice versa.

Single Article Per Assignment. The reading behavior (R) was statistically significant with question easiness in Easy-Read and Medium-Read article categories. In both groups, students showed more reading behavior for easy questions.

The annotation behavior (A) was statistically significant in the Medium-Read and Easy-Read article group. In both of these categories, students showed more annotations before attempting a harder question, indicated by the negative correlation coefficient.

Table 6.10: Spearman's Correlation between Question Readability and SRL by FRES. Bold = Statistically Significant. R =No. read, A=No. Notes, H= No. Highlight, V= No. Vocab. lookup

Feature	Single Article Per Assignment			Multiple Articles Per Assignment
	<i>Hard-Read</i>	<i>Medium-Read</i>	<i>Easy-Read</i>	
R	0.03 ($p = 0.14$)	0.18 ($p < 0.001$)	0.12 ($p < 0.001$)	-0.05 ($p < 0.001$)
A	-0.01 ($p = 0.74$)	-0.05 ($p < 0.001$)	-0.16 ($p < 0.001$)	-0.02 ($p = 0.83$)
H	0.01 ($p = 0.45$)	-0.001 ($p = 0.099$)	-0.01 ($p = 0.46$)	-0.01 ($p = 0.52$)
V	0.044 ($p = 0.006$)	0.05 ($p < 0.001$)	-0.11 ($p < 0.001$)	-0.02 ($p = 0.13$)
Score	-0.09 ($p < 0.001$)	-0.15 ($p < 0.001$)	0.011 ($p = 0.46$)	-0.06 ($p < 0.001$)

The highlighting behavior (H) did not show statically significant relationships in any group.

Considering the vocabulary lookup (V) behavior, we found statistically significant relationships in all article category groups. In Hard-Read and Medium-Read groups, students looked up more vocabularies for easy questions. However, for Easy-Read articles, students looked up less vocabulary for easy readability questions.

Finally, the question performance (Score) was statistically significant in Hard-Read and Medium-Read article categories. In both of these groups, students' scores showed a negative correlation with question easiness.

Multiple Articles Per Assignment. For assignments with multiple articles, we found only reading (R) and question performance score showing a statistically significant correlation with question readability. The reading behavior showed a negative correlation, indicating students read more as questions became harder (i.e., question readability score decreases).

6.3.5 Discussion of RQ3.2

In this section, we summarize our findings and outline possible implications for teachers and intelligent tutoring system (ITS) (Wenger, 2014) design consideration. Our results in Section 6.3.3 shows that students' score on assignments varied significantly depending on text complexity among all three pairs: Easy-Read vs. Hard -Read, Easy-Read vs. Medium-Read, and Medium-Read vs. Hard-Read. The largest effect size was observed between the Medium-Read vs. Hard-Read pair. The reading behavior also had the largest effect size in Medium-Read vs. Hard-Read article categories. Similarly, all SRL behaviors were statistically significant and varied with text complexity.

As suggested by Graesser and colleagues (Graesser et al., 2014), students sometimes

are required to be challenged by complex text to improve their reading skills. Other times, students may be assigned to easier texts to comprehend and boost their confidence. Teachers can assign readings to students with varied text complexity depending on purpose: to promote SRL, to familiarize with complex texts, or to boost students' confidence by assigning easier texts.

Our results from Table 6.10 and 6.9 present students reading and SRL variance with question-positioning and question-readability depending on the text complexity of the article. As Table 6.9 shows, we observed a statistically significant relation between SRL behavior and question positioning for Medium-Read articles. Highlighting and vocabulary behavior decreases for questions in the lower part of the text. Considering reading behavior, we observe students' reading behavior decreases for Medium-Read texts but increases with Hard-Read texts. Our analysis shows that the text complexity and the positioning of the text are both essential factors in understanding reading behavior. Our findings are similar to Goedecke et al. (Goedecke et al., 2015) –where researchers found that the decoupling rate was more associated with vertical positioning on the screen and text formalities of informational, narrative, and persuasive texts rather than reading behavior (i.e., reading speed). From an ITS design perspective, reading materials can be split into chunks considering text complexity and vertical positioning in the screen for better comprehension.

Examining Table 6.10, we observe a statistically significant relation between question score and question readability in the Hard and Medium-Read category. In both of these categories, the association was negative, i.e., score decreased with easy readability questions (question readability increases easier). We examined assignments of these two categories and found that SA questions consisted of the majority. In the Medium-Read category with questions FRES [75-100]: 544 questions were SAs, and 106 were MCQs. In the Hard-Read category with question FRES [75- 100], 32 questions were SAs, and zero questions were MCQs. We conclude that although SA questions had easy readability measures, scores on these questions were lower than MCQs. Considering SRL, the vocabulary lookup behavior was positively correlated with question easiness in these two categories. Again, we conclude the SA question format resulted in high vocabulary lookups. From an ITS design viewpoint, we suggest considering both question readability and question format to support students. Examples of supports could be: providing more explanations to question texts and adding the meaning of frequently looked up vocabulary within the text.

6.4 Conclusion

In our RQ1.1 and RQ2, we examined how students' reading and SRL are connected with their assignment performance. In both studies, we considered question formats in AL, i.e., MCQ and SA. In RQ3.1, we take into account the question difficulty to analyze reading and SRL behavior. Our *class-level* question difficulty analysis is in-line with IRT results. In RQ3.2, we conducted an exploratory analysis to understand how students' observed behaviors, such as reading, SRL, and question performance vary with text features. To understand text features, we studied several features related to article and question texts, such as text complexity of articles measured by readability and formality, text complexity of questions measured by readability, and vertical positioning of questions in the article. Our analysis showed that students reading, SRL, and scores on questions vary at statistically significant levels with text readability of articles. To understand the impact of question features, we conducted a correlation analysis between question readability, positioning, and observed students' behavior. Our findings showed that the impact of questions features varied with text features of articles.

The findings of our study may help teachers and ITS designers to create reading materials that promote SRL and support students, such as provide more hints for 'Select all that apply' MCQs.

CHAPTER

7

FEEDBACK AND STUDENTS' SRL

In this Chapter we will perform an exploratory analysis to answer RQ4. We collected students' log trace data in AL who re-submitted questions after receiving teachers' feedback. We will describe how students' updated their response and adapt their reading and SRL behaviors upon receiving feedback on questions. We answer two research questions (RQ).

RQ4.1 How does students' score on questions vary upon receiving feedback?

RQ4.1.1 To what extent do students change their answers upon receiving feedback?

RQ4.2. How does students' reading and SRL usage vary upon receiving feedback?

We first describe background and related work in Section 7.1. Next, we describe our methodology and results of two RQs in Section 7.3 and 7.4, respectively. We discuss our findings in Section 7.5. Finally, we conclude in Section 7.6.

7.1 Background

In the context of feedback and SRL, researchers identified feedback as one of the influential environmental factors enhancing or impeding students SRL (Butler & Winne, 1995; Dweck, 2000; Garcia, 1995*a*). Butler and Winne proposed a model to integrate feedback in the SRL

process (Butler & Winne, 1995). They concluded feedback is important factor to engage students in self-regulated learning. Nicol and Macfarlane-Dick (Nicol & Macfarlane-Dick, 2006), proposed a model integrating formative feedback and self-regulated learning. they address the issue of how feedback principles help to promote self-regulation. Nicol and Macfarlane-Dick (Nicol & Macfarlane-Dick, 2006) categorized feedback into two categories: external feedback referring to teachers and peers) and internal feedback referring to students' awareness of the outcomes resulting from SRL activities. Self-regulated learners differ from their non-self-regulated peers by generating more internal feedback, responding positively to external feedback, and increasing efforts to achieve learning goals (Bose & Rengel, 2009).

Although feedback are powerful, not all feedback are useful. Positive, non-judgmental, and specific feedback foster students SRL (Hawk & Shah, 2008). Moreover, receiving feedback and acting upon it depend on students' self-belief and self-esteem (Garcia, 1995*a*), study habits (Dawson et al., 2018*a*), and demographics. How students act upon after receiving feedback is an area that is underexplored (Iraj et al., 2020; Tay, 2015). Tay (Tay, 2015) examined how students' SRL and formative assessment (FA) vary in different contexts with 13 students. Formative assessment is defined as providing feedback on performance to improve and accelerate learning (Sadler, 1998). The author examined two different tasks with the same cognitive level but in different contexts: one in pen and pencil (PP) format and another in an online forum of a national daily. Both tasks required reading and paraphrasing a text in written form. In the PP format, the text was printed whereas in online forum, students selected a featured newsletter. FA in PP format was teachers' assessment and in online forum was feedback from other readers. Students self-reported their SRL by questionnaire and a semi-structured interview. Findings showed students were more careful in monitoring their SRL to paraphrase in online forum. Students reported feedback in the online forum was more engaging. By observing other letters on the forum, students were able to self-assess their work, a key component of SRL. In contrast, feedback in the PP format came from the teacher. Students were not encouraged to exercise self-reflection on their work receiving teachers' feedback. Our study differs from Tay that we will evaluate SRL and teachers' feedback on academic readings in question answering context within the AL. Iraj et al. focused on system-generated feedback efficiency (Iraj et al., 2020). Our study differ from Iraj et al. that we will focus on feedback given by teachers. We will analyze students adaptation reading and SRL upon receiving feedback within the AL platform.

While earlier research on feedback focused on providing specification of a task, current research focus on change in student behavior (Dawson et al., 2018*b*). Nicol stressed

feedback as a two-way communication process (Nicol, 2010). That is, in order for feedback to be effective, students have to understand it, and be able to take action based upon it. However, feedback gaps (Evans, 2013) can happen due to lack of clarity in feedback (Burke, 2009), students' lack of understanding and how to address the feedback (Weaver, 2006) and the feedback paradox (Withey, 2013) — where students do not address feedback despite understanding its importance. To mitigate feedback gap, researchers have begun to emphasize *feedback actionability* i.e., the potential to change students' actions and behaviors (Carless & Boud, 2018). This line of research include understanding students' perception about feedback (Kung & Scholer, 2018; Rowe et al., 2008) and analyzing students behavior upon receiving feedback including learning strategies (Matcha et al., 2019) and responsiveness to teachers' feedback (Iraj et al., 2020).

Our work falls in this line, as we seek to understand students' feedback recipient behavior in the science reading context.

7.2 Dataset Preparation

From our science dataset in Section 3.2, we identified student submissions on which they received feedback. This reduced dataset included 1,819 unique students, 3,867 questions, and 5,373 submissions. As we were interested in students' behavior after receiving feedback, we applied filtering criteria : (i) a student submitted a question multiple times, (ii) received at least one instance of feedback, and (iii) re-submitted after receiving feedback. Based on our filtering criteria, we prepared two datasets:

Dataset D1: Student answer may contain an empty submission. The AL system records empty submission as “No response.” This dataset contained 670 unique students in 113 classes, 58 teachers, 156 assignments, 1,072 questions, and 2,502 submissions.

Dataset D2: Removing empty submissions from D1. This dataset contained 652 unique students in 107 classes, 54 teachers, 148 assignments, 1,048 questions, and 2,453 submissions.

All questions in our dataset are SA questions.

7.3 Methodology

7.3.1 RQ4.1 Methodology

We include students' submissions with multiple attempts after receiving feedback from teachers. We calculated the difference between the first and last submission scores on such questions. We observed three categories of submissions: score increased in the last submission, score decreased in the last submission, and score was unchanged in the last submission. To assess whether students were addressing teachers' feedback, we calculated similarity between subsequent submitted answers to a single question. We hypothesized that changes in submitted answers would result in a score difference in a question. Thus, we measured the cosine similarity between subsequent submissions of a question. More specifically, we calculated cosine similarities between i th and $(i-1)$ th submissions, for $i \Rightarrow 2$ attempts and took the average. In other words, we measured the cosine similarity of a submission and its immediately prior attempt and took the average of all similarity measures. The cosine similarity score measures the angle of two vectors defined as below. The smaller the angle, the more similar are two vectors, and the value is closer to 1. The value of cosine similarity ranges from -1 to 1.

To encode students' responses into vector representations, we used Universal Sentence Encoder (USE) (Cer et al., 2018). The USE can take phrases, sentences, and short paragraphs as inputs to calculate text similarity, classification, and other natural language processing (NLP) tasks. USE encodes sentences into a fixed length vector of 512. We used a deep averaging network (DAN) model of USE to encode questions and question-dependent texts into vectors (Cer et al., 2018). DAN averages unigrams and bi-grams of word embedding to construct sentence embedding.

To evaluate how answer modifications were connected to score differences, we calculated Spearman's correlation between mean cosine similarities and score difference.

7.3.2 RQ4.2 Methodology

We split students' actions into sessions following the same procedure in Section 5.2.1. Next, we counted SRL events *before* a student's resubmission of the question received feedback. An illustrative example of a student's action sequence is

Q2 [feedback] Q3 R R A H A Q2

In the above example, a student submitted Q2 twice. Teacher provided feedback on the first attempt of Q2. Prior resubmitting Q2, the student performed two reading events (R), two annotations (A), and one highlight (H). We counted the number of reading and SRL events prior to re-submission after a question received feedback. We applied a four-level hierarchical linear model (HLM) to predict the last score of a question. We applied HLM with questions at level one (L1), nested within assignment ID (level two, L2), nested within student ID (level three, L3), nested within teacher ID (level four, L4). Fixed effect variables were students' first score on questions and features of SRL usage during attempting questions. All grouping variables were modelled as random intercepts.

7.4 Results

7.4.1 RQ4.1 Result

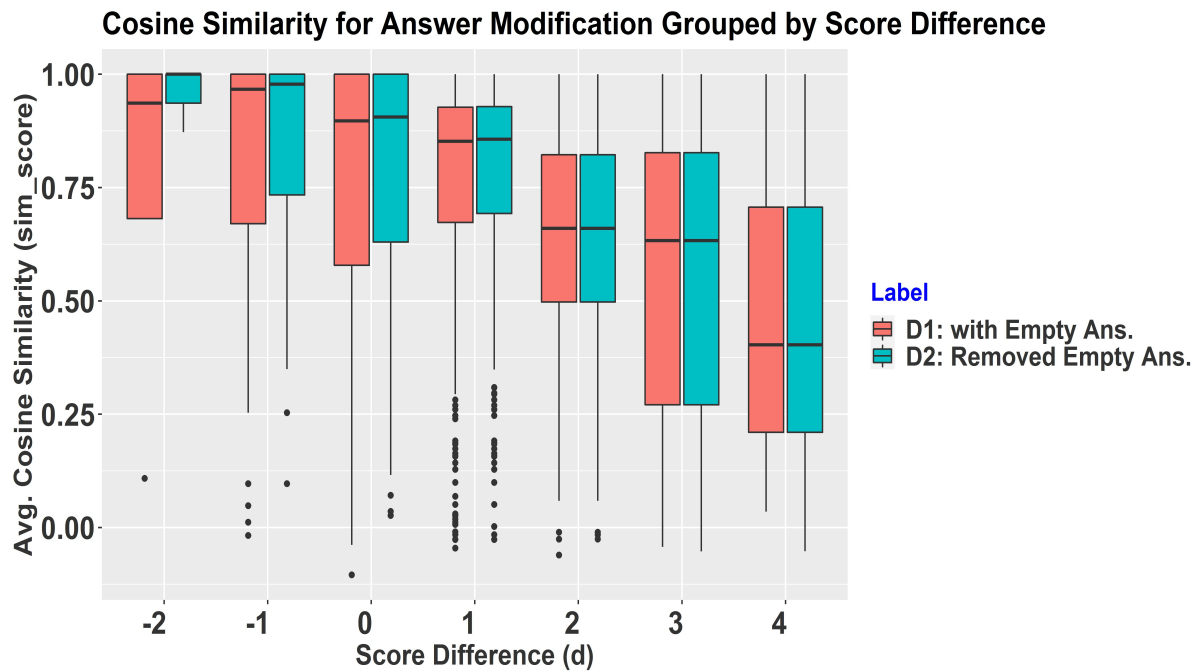


Figure 7.1: Avg. cosine similarities of students' responses to a question

Figure 7.1 presents average cosine similarities between subsequent submissions (sim_score) and score difference (d). A higher sim_score indicates submitted answers are more

similar to each other. Conversely, a lower `sim_score` value indicates subsequent answers are different from each other.

Frequencies of six different score difference categories and question counts (n) in D1 dataset are: -2 (n = 4), -1 (n = 53), 0 (n = 187), 1 (n = 474), 2 (n = 252), 3 (n = 87), and 4 (n = 15). Total questions = 1,072. After removing empty submissions and subsequently removing single attempt questions in D2 dataset: -2 (n = 3), -1 (n = 50), 0 (n = 178), 1 (n = 463), 2 (n = 252), 3 (n = 87), 4 (n = 15). Total questions = 1,048.

The Spearman correlation test between score difference (d) and mean cosine similarities (`sim_score`) were D1 (coefficient = -0.315, $p < 0.001$) and D2 (coefficient = -0.364, $p < 0.001$). The negative coefficient indicates when the mean cosine similarity score decreases, the score difference increases. In other words, the more changes are present in students' subsequent answers, the greater the score difference. We present examples of student answers (S), corresponding teacher's feedback (T), mean cosine similarity score, and score difference in the following.

Score Increased. Descriptive statistics of D1 dataset in this category are: 828 unique questions, 1,963 submissions by 543 students. First attempt score ranged from 0 to 3 with a mean 1.41. Last attempt score varied from 1 to 4 with a mean 2.98. D2 dataset contained 817 unique questions, 1,940 attempts by 535 students. First attempt score ranged from 0 to 3 with a mean 1.41. Last attempt score varied from 1 to 4 with a mean 2.99.

From Figure 7.1, we observe positive score change groups have increased by 1, 2, 3, and 4 points. In these four groups, `sim_score` has a lower median value compared to the rest. This observation indicates that students with greater score increases had submissions that differed more than their original answers, as represented by a lower `sim_score`.

We examined student submissions with identical responses (`sim_score` = 1) but an increase in last score (n = 40) submissions. We identified following cases in D1 dataset:

12 questions from score difference +1 include "Please start sentence with a capital letter," "Use punctuation!"; The USE removes punctuation and stop words, and converts lower cases—the pipeline in NLP preprocessing tasks. The preprocessing step resulted in `sim_score` = 1.

For the remaining cases, teachers increased scores despite students submitting the same answers.

Score Decreased. Descriptive statistics of D1 dataset in this category are: 57 unique questions, 124 submissions by 49 students. First attempt scores ranged from 1 to 4 with a mean 1.58. Last attempt score varied from 0 to 3 with a mean 0.51. D2 dataset contained 53

unique questions, 116 attempts by 45 students. First attempt score ranged from 1 to 4 with a mean 1.57. Last attempt score varied from 0 to 3 with a mean 0.51.

Score Unchanged. Descriptive statistics of D1 dataset in this category are: 187 unique questions, 415 submissions by 165 students. First and last attempt scores have the same statistics in this category. First and last attempt scores ranged from 0 to 4 with a mean 1.69. D2 dataset contained 178 unique questions, 397 attempts by 156 students. First and last attempt scores ranged from 0 to 4 with a mean 1.71. The median *sim_score* in D1 and D2 dataset are 0.90 and 0.91 respectively. We present some examples from D1 below.

57 submissions have *sim_score* = 1. Among these, 16 submissions received feedback related to capitalization and punctuation such as “Use capitals and periods for a complete sentence.” For the remaining cases, students submitted identical answers and their final score was also unchanged.

7.4.2 RQ4.2 Results

Table 7.1: Results from HLM measuring association between SRL and science score

	β	B	SE	p
Intercept		2.12	0.10	<0.001
No. Times Read	0.05	0.08	0.04	0.04
No. Notes	0.03	0.23	0.20	0.24
No. Highlights	-0.04	-0.35	0.21	0.098
No. Vocabulary Lookups	0.01	0.06	0.13	0.63
First Score	0.29	0.33	0.03	<0.001

Table 7.1 presents our HLM results on D1 dataset. We calculated standardized effect sizes using the formula, $\beta = (B \cdot SD_x) / (SD_y)$ (Rutherford et al., 2017). First attempt score had the highest predictive power ($B = 0.32$, $\beta = 0.28$, $p < 0.001$). Among reading and SRL variables, reading was a statistically significant positive predictor. Highlighting behavior was negatively associated with the last score. HLM with removing empty answers on D2 dataset showed similar results: reading ($B = 0.06$, $\beta = 0.04$, $p = 0.08$), annotations ($B = 0.23$, $\beta = 0.03$, $p = 0.24$), highlights ($B = -0.34$, $\beta = -0.04$, $p = 0.099$), vocabulary lookups ($B = 0.5$, $\beta = 0.01$, $p = 0.67$), first score ($B = 0.32$, $\beta = 0.28$, $p < 0.001$). However, reading was not statistically significant in this case.

7.5 Discussion

Our RQ1 results show that students' who modified their answers had greater score differences at levels of statistical significance. Our findings are in line with Zhu and colleagues (Zhu et al., 2017) who inspected the impact of automatic feedback on argumentative writing in high school students. They found statistically significant differences in the first score between students who addressed feedback and who did not. They also found a statistically significant difference in the last score between those who revised their answer and those who did not. Unlike Zhu's analysis, our study focuses on teacher-provided feedback instead of automated feedback. We observe that teachers may decrease points after students revise their responses. When examining students' responses, we found students sometimes submitted the identical answer or an empty answer ("No response") despite the teacher asking for explanation or suggesting additional correction. This phenomenon can be explained by the "feedback gap" as described by Evans (Evans, 2013), where students do not address teachers' feedback. Another reason could be that students found feedback difficult to decipher (Burke, 2009). Moreover, receiving feedback and acting upon it depends on students' self-beliefs (Garcia, 1995*b*) and study habits (Dawson et al., 2018*c*). As we do not have access to students demographic information or perception data, we do not know the exact reasons they were unresponsive.

Our HLM analysis from RQ2 shows that reading and the first score were statistically significant predictors of the last score. However, SRL variables such as, annotation, highlighting, and vocabulary lookup were not statistically significant.

7.6 Conclusion

In this Chapter, we empirically evaluated students' response changes to short answer questions upon receiving feedback. Our study contributes to two lines of research: reading and SRL. As a part of our analysis, we calculated cosine similarities between consecutive answer submissions to assess students' answer changes upon receiving feedback.

Our findings show that students who revised their answers showed statistically significant differences in their scores. We also observed some students submitted an empty answer after receiving feedback on a question ($n = 49$, obtained from D1-D2 in Dataset Set preparation). Additionally, we observed some students re-submitting the same answer. We conclude that students were not acting on feedback in the above mentioned cases.

One solution to increase actionability could be pointing to additional learning materials,

as done by Herodotou et al. in an automated generated feedback setting (Herodotou et al., 2017). Broos and colleagues (Broos et al., 2017) designed a button “*Okay, what now?*” in a dashboard to provide actionable feedback. Students can click the button to view extra reading content. Similar to their Broos and colleagues’ design, a nudge can be implemented in AL “*Are you sure you want to submit that empty answer?*”

Our RQ1, RQ2, and RQ3 analysis uses student interaction data. In contrast, our RQ4 analysis utilizes both teachers’ and students’ interaction with the AL. Thus, combining our analyses of four RQs, we can have a better understanding of how teachers’ and students interacted with the AL system and how those interaction shaped students reading and SRL behavior.

CHAPTER

8

CONCLUSIONS AND FUTURE WORK

In RQ1 in Chapter 4, first, we measured the association of SRL frequency and science question scores. Our findings showed that the association was greater in SAs than in MCQs. Next, we computed the cosine similarity between the SA question text and students' responses. We assessed how the cosine similarity varies (i) across two subject domains and (ii) across the class factor. In the later Chapter 4, we proposed predictive student modelling taking inputs: prior question attempts and SRL. We applied a window-based technique to handle the temporal ordering of inputs. From our empirical experimental observation, we conclude that recency of action has better predictive power.

In RQ2 in Chapter 5, we aimed to understand how reading and SRL behavior vary by (i) subject domain and (ii) with students' academic performance. We considered two subject domains: science and social studies. For our analysis, we first grouped students of two subject domains separately by their question performance. Next, we applied the n -gram sequence mining on each group to extract frequent reading and SRL patterns. We used differential sequence mining to test if observed patterns vary at a statistically significant level. Based on our analysis, we concluded: science high-performing students exhibited more reading and annotations than low-performing students. Considering cross-domain SRL, students with social study assignments showed more reading behavior than science.

In RQ3 in Chapter 6, we empirically assessed question difficulty considering the class-level factor. We included questions used in at least two classes. We grouped questions into three categories: Easy, Medium, and Hard. We examined students' reading and SRL behavior among three groups. Our findings showed that Easy questions prompted more SRL. Examining the data, we concluded, the SA question format (75 out of 81 in the Easy category) prompted more SRL. We also conclude that MCQs with '*Select all that apply*' are Hard for students. In the later part of Chapter, we examined how text-complexity and vertical question positioning correlate with performance, reading, and SRL. We studied several features related to article and question texts, such as text complexity of articles measured by readability and formality, text complexity of questions measured by readability, and vertical positioning of questions in the article. Based on our analysis, we conclude that the text-readability of an article and vertical question positioning are important to evaluate the reading and SRL behavior of a question.

In RQ4 in Chapter 7, we examined students' SRL and question response behavior after receiving feedback. Based on our analysis, we conclude students' modification of response upon receiving feedback had a statistically significant correlation with their last scores. Additionally, we observed some students submitted a blank response or did not update their responses. Based on our observation, we conclude this phenomenon as *feedback gap*. As we do not have students' perception data, we can not evaluate why students did not address teachers' feedback.

Overall, the high-level goal of this thesis is to understand how students' reading and SRL behavior are connected to their academic performance within the AL system. We outline contributions of this thesis below.

8.1 Contributions

To the EDM Research Community. We pioneered the research direction in combining temporal ordering of question attempts and learning activities in predictive student modelling—an underexplored area as identified by Choi et. al (Choi et al., 2020). Most prior research, such as Learning Factors Analysis (LFA) Chi et al. (2011), BKT (Corbett & Anderson, 1994), and DKT (Piech et al., 2015) use on students' previous question-interaction data to predict performance on question. However, these methods do not utilize the temporal ordering of students' interaction behavior within systems, such as reading, video watching, and others. Contrary to previous approaches that rely on question-solving actions, we consider the

temporal ordering of question-solving behavior and other student interaction behaviors within the system. In particular, we proposed two predictive models to capture the reading interaction data with question performance. One is an LSTM based model, and another LSTM with an attention-based mechanism, providing explanations of the model's input in making predictions.

For Student Support. My thesis presents a set of experiments to understand students' reading and SRL behaviors within the AL, including

1. An experiment to identify productive and unproductive students' reading and SRL patterns vary across subject domains, such as science and social studies (Farhana et al., 2020*b*).
2. An empirical analysis of students' SRL activities when they face questions with varying difficulty levels and comparison with the IRT (Farhana et al., 2020*a*).
3. An exploratory analysis to understand how students adapt their SRL after receiving teachers' feedback on questions (Farhana et al., 2021).
4. An empirical account to understand students' SRL behavior when they read articles with varying text complexities.

Findings from this study can be helpful to design recommendation systems targeting specific groups of users. Additionally, our findings may aid intelligent tutoring system (ITS) designers to create reading materials that promote students' SRL in science reading.

8.2 Design Implications

In this Section, we discuss possible design implications for the EDM, learning science community, and the AL platform.

8.2.1 Design Implication: EDM Community

Question Format

Our HLM analysis of RQ1 shows a negative association of vocabulary lookup behavior for MCQ score. Our RQ2 analysis shows that RH+M and V+M patterns were frequent among low-performing students in science. Additionally, low-performing science students had

lower MCQ last scores and longer submission time. Our RQ3.1 findings show that MCQ with '*Select all that apply*' was found hard across classes.

The above finding can be used to support low-performing students in MCQ questions. Specifically, if the system detects a student (i) is taking longer time on an MCQ compared to other students' on the same question, (ii) reading, doing SRL activities before attempting an MCQ, and (iii) can't answer correctly, then the system can redirect the student to extra practice questions. The system can also point to relevant reading articles by analyzing the question text. Additionally, for MCQs with *Select all that apply*, the system can explain each MCQ option.

Vocabulary Lookup for Science Domain

In our RQ2, we analyzed within and cross-domain SRL patterns for science and social studies students. Our within domain analysis for science showed V+M patterns were more frequently used by low-performing science students than the high-performing group. Considering cross-domain SRL behavior, we found only high-performing science students exhibited V+MV and SV+M patterns frequently. These patterns are related to vocabulary lookup behavior in MCQ (V+MV) and SA (SV+M) attempts. From these observations, we conclude science students looked up vocabulary more than social study students. From an ITS design perspective, we recommend providing more vocabulary lookup support in science reading.

Question Text Complexity

Our RQ3 focuses on question text complexity and its association with students' performance. We observed students showed more annotation behavior if the question is difficult to read. However, students' score was negatively associated with question readability. Further examining question texts, we observed easy readability questions are mostly SAs. For example, *Which property of liquids explains why a glass of water can be filled over the brim without spilling? Explain how this works.* is an SA question with FRES score of 85.69. Although this question is easy to read, students may find it difficult for its open-ended format.

The above observations can be useful in designing an ITS. For hard-readability questions, the system can provide hints or explanations of the question texts. Additionally, for easy-readability SA questions, the system can also provide similar support.

Feedback Actionability

We observed students submitted empty responses and identical responses after receiving feedback. To enhance students' feedback actionability, ITS can implement prompts in the above-mentioned cases.

8.2.2 Design Implication: Learning Science

Question Format

Our HLM analysis of RQ1 shows that SRL behaviors were more predictive of SAs than MCQ scores. This finding is in line with previous literature (O'Neil Jr & Brown, 1998). Our RQ3.1 findings showed that students found MCQs with '*Select all that apply*' hard across classes. One limitation of our study is we do not have access to information about students, teachers, and other confounding variables. A follow up study could be designing a user study to understand how students perceive MCQs with '*Select all that apply*' and adapt their SRL behaviors.

Question Positioning

Previous literature examined human reading attention decreases with vertical positioning of the text (Forrin et al., 2020; Li et al., 2018). In our RQ3.2 analysis, we found that students' performance on questions depends on questions' vertical positioning and article text complexity. Thus, we recommend considering both the text-complexity and the length of the text to conduct such experiments.

8.2.3 Design Implication: The AL Platform

Fine-grained Logging of Reading

In our analysis, we found that the AL platform does not capture fine-grained reading behavior, such as which part of the assignment students are reading. Also, we observed some students attempted MCQs and SAs before reading the assignment even though the questions were in the middle of the assignment. We recommend the AL system designers to log students' reading behavior precisely. Specifically, whether students are reading textual description, or mathematical equations, or viewing images in the text.

Grade and Feedback Viewing Time

The AL system logs the grading time of MCQs and SAs. However, it does not log when students are viewing their grades. Similarly, the system does not log when students are viewing their teachers' feedback. Incorporating these timestamps can enable researchers to conduct a more fine-grained analysis of students' interaction with the system.

8.3 Future Work

This thesis lays several future direction of studies described below.

Data Guided Intervention & Winne and Hadwin's Model The basic assumption of Winne and Hadwin's model is to understand patterns' of students' behavior from log data. More specifically, this model assumes *if Then* pattern—which implies if student performs an action X Then the outcome will be Y. Our RQ1 and RQ2 were based upon detailed temporal analysis. In other RQs, we analyzed sequential student behavior by splitting into sessions and considering SRL prior to the question attempt. **Automating Feedback** Middle school teachers may not have enough time to provide detailed feedback to all students; providing feedback to open-ended questions asking explanations is challenging as students' answers can vary (Botelho & Heffernan, 2019). As stressed by Botelho and Heffernan (Botelho & Heffernan, 2019)

“For a 7th grade math teacher with approximately 100 students, a teacher would have 30 seconds each day to provide feedback on submitted work. This alarming statistics suggest that teachers are in need of better support in order to provide students with beneficial feedback.”

To assist teachers in providing better quality feedback and reduce their workload, researchers have begun to focus on automated feedback generation and suggestion (Botelho & Heffernan, 2019; Cavalcanti et al., 2020). Cavalcanti et al. analyzed the quality of feedback generated in the Moodle platform and automatic feedback generation process following the Hattie and Temperly's (Hattie & Timperley, 2007) coding scheme (Cavalcanti et al., 2020). Similarly, Botelho and Heffernan (Botelho & Heffernan, 2019) proposed a feedback suggestion system, DRIVER-SEAT (Heffernan, 2020) in a middle school mathematics tutoring system, ASSISTments. Their project aims at developing several tools, such as QUICK-Comments, which will generate three customized feedback per student and send them to teachers. Teachers can select one of the suggested feedback. It may be the case that such similar methods of automated feedback could be helpful for AL science questions.

However, one challenge in automating feedback is that science question answers are

often in free text form and need to connect information from different paragraphs. The state-of-the-art NLP research to automatically infer information from paragraphs in reading comprehension is still in the early stage (Dua et al., 2019; Joshi et al., 2020). In one study, the authors found the F1-score automated system was 32%, whereas humans achieved 96% (Dua et al., 2019).

One possible automation design could be to collect other teachers' feedback on the same question and provide suggestions to the teacher. In the AL platform, teachers can use pre-built questions or create their own. For pre-built questions, an automated system could suggest feedback collected from other teachers or expert human feedback generators on the same question.

Generalizability and Transferrability Across Platform and Subjects In a recent study, Baker outlined six challenges in the future EDM research (Baker, 2019). Among these, one challenge is related to transferability and another one generalizability. The first problem is formally defined as "*Transferability: The (learning system) Wall*". The problem aims to identify whether a student's observed behavior in one platform generalizes to other platforms for the same student. The second problem is formally defined as "*Generalizability: The New York City and Marfa Problem*". The second challenge states that most educational research data are collected from smaller cities or upper-middle-class suburbs from the U.S. In contrast, little research is done in areas with restrictive policies to gather data (i.e., New York) or a small town far from the airport (i.e., Marfa, Texas). This challenge aims for EDM research and learning analytics valid for students at New York and Marfa, as well.

The study of this thesis is based on AL Science student SRL behavior. A follow-up study could be to test the generalizability of these results on another middle school STEM learning platform and subject domain, say, ASSISTments. ASSISTments is a middle-school mathematics tutoring system with MCQ and SA questions. To test the generalizability, Baker suggests to collect a new set of students interacting with another system and test generalizability.

REFERENCES

- ACT (2006), American college testing, *in* 'Reading between the lines: What the ACT reveals about college readiness in reading.'
- Ai, F., Chen, Y., Guo, Y., Zhao, Y., Wang, Z., Fu, G. & Wang, G. (2019), 'Concept-aware deep knowledge tracing and exercise recommendation in an online learning system.', *International Educational Data Mining Society*.
- Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E. & Lester, J. (2018), 'Improving stealth assessment in game-based learning with lstm-based analytics.', *International Educational Data Mining Society*.
- Alexander, P. A., Dinsmore, D. L., Parkinson, M. M. & Winters, F. I. (2011), 'Self-regulated learning in academic domains', *Handbook of self-regulation of learning and performance* pp. 393–407.
- Anderson, L. W., Krathwohl, D. R. & Bloom, B. S. (2000), 'A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives'
- Araka, E., Maina, E., Gitonga, R. & Oboko, R. (2020), 'Research trends in measurement and intervention tools for self-regulated learning for e-learning environments—systematic review (2008–2018)', *Research and Practice in Technology Enhanced Learning* **15**(1), 1–21.
- Arnold, K. E. & Pistilli, M. D. (2012), Course signals at purdue: Using learning analytics to increase student success, LAK '12, Association for Computing Machinery, New York, NY, USA, p. 267–270.
URL: <https://doi.org/10.1145/2330601.2330666>
- Azevedo, R. (2008), 'The role of self-regulated learning about science with hypermedia', *Recent innovations in educational technology that facilitate student learning* pp. 127–156.
- Azevedo, R. & Cromley, J. G. (2004), 'Does training on self-regulated learning facilitate students' learning with hypermedia?', *Journal of educational psychology* **96**(3), 523.
- Azevedo, R., Johnson, A., Chauncey, A. & Burkett, C. (2010), Self-regulated learning with metatutor: Advancing the science of learning with metacognitive tools, *in* 'New science of learning', Springer, pp. 225–247.
- Azevedo, R., Moos, D. C., Johnson, A. M. & Chauncey, A. D. (2010), 'Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges', *Educational psychologist* **45**(4), 210–223.
- Azevedo, R., Witherspoon, A. M., Graesser, A. C., McNamara, D. S., Chauncey, A. & Siler, E. (2009), Metatutor: Analyzing self-regulated learning in a tutoring system for biology, AIED.

- Baker, L. (1985), 'How do we know when we don't understand? standards for evaluating text comprehension', *Metacognition, cognition, and human performance* **1**, 155–205.
- Baker, R. S. (2019), 'Challenges for the future of educational data mining: The baker learning analytics prizes', *JEDM| Journal of Educational Data Mining* **11**(1), 1–17.
- Bannert, M. & Mengelkamp, C. (2008), 'Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. does the verbalisation method affect learning?', *Metacognition and Learning* **3**(1), 39–58.
- Barnes, T. (2005), The q-matrix method: Mining student response data for knowledge, *in* 'American Association for Artificial Intelligence 2005 Educational Data Mining Workshop'.
- Beaudoin, L. & Winne, P. (2009), nstudy: An internet tool to support learning, collaboration and researching learning strategies, *in* 'Canadian e-Learning Conference'.
- Biemiller, A. & Slonim, N. (2001), 'Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition.', *Journal of educational psychology* **93**(3), 498.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *J. Mach. Learn. Res.* **3**(null), 993–1022.
- Bliss, L. B. (1980), 'A test of lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students', *Journal of Educational Measurement* pp. 147–153.
- Bose, J. & Rengel, Z. (2009), 'A model formative assessment strategy to promote student-centered self-regulated learning in higher education.', *Online Submission* **6**(12), 29–35.
- Botelho, A. F. & Heffernan, N. T. (2019), 'Crowdsourcing chapter 11–feedback to support teachers and students', *Design Recommendations for Intelligent Tutoring Systems: Volume 7-Self-Improving Systems* p. 101.
- Bouchet, E., Azevedo, R., Kinnebrew, J. S. & Biswas, G. (2012), 'Identifying students' characteristic learning behaviors in an intelligent tutoring system fostering self-regulated learning.', *International Educational Data Mining Society*.
- Bouffard-Bouchard, T., Parent, S. & Larivee, S. (1991), 'Influence of self-efficacy on self-regulation and performance among junior and senior high-school age students', *International journal of behavioral development* **14**(2), 153–164.
- Broos, T., Peeters, L., Verbert, K., Van Soom, C., Langie, G. & De Laet, T. (2017), Dashboard for actionable feedback on learning skills: scalability and usefulness, *in* 'International Conference on Learning and Collaboration Technologies', Springer, pp. 229–241.
- Buehl, D. (2017), *Developing readers in the academic disciplines*, Stenhouse Publishers.

- Burke, D. (2009), 'Strategies for using feedback students bring to higher education', *Assessment & Evaluation in Higher Education* **34**(1), 41–50.
- Butler, D. L. & Cartier, S. C. (2005), 'Multiple complementary methods for understanding self-regulated learning as situated in context', *Annual meetings of the American Educational Research Association, Montreal, QC*.
- Butler, D. L. & Winne, P. H. (1995), 'Feedback and self-regulated learning: A theoretical synthesis', *Review of educational research* **65**(3), 245–281.
- Carless, D. & Boud, D. (2018), 'The development of student feedback literacy: enabling uptake of feedback', *Assessment & Evaluation in Higher Education* **43**(8), 1315–1325.
- Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F & Gašević, D. (2020), How good is my feedback? a content analysis of written feedback, *in* 'Proceedings of the Tenth International Conference on Learning Analytics & Knowledge', pp. 428–437.
- CCS (2010), Common core state standards initiative, *in* 'Common Core State Standards for English language arts literacy in history/social studies, science and technical subjects', Washington, DC: Council of Chief State School Officers National Governors Association.
- Cen, H., Koedinger, K. & Junker, B. (2006), Learning factors analysis—a general method for cognitive model evaluation and improvement., *in* 'International Conference on Intelligent Tutoring Systems'.
- Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. et al. (2018), Universal sentence encoder for english, *in* 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', pp. 169–174.
- Chall, J. S. & Dale, E. (1995), *Readability revisited: The new Dale-Chall readability formula*, Brookline Books.
- Chen, Y.-C., Peng, W.-C. & Lee, S.-Y. (2016), Mining temporal patterns in interval-based data, *in* '2016 IEEE 32nd International Conference on Data Engineering (ICDE)', IEEE, pp. 1506–1507.
- Cheung, L. P. & Yang, H. (2017), Heterogeneous features integration in deep knowledge tracing, *in* 'International Conference on Neural Information Processing', Springer, pp. 653–662.
- Chi, M., Koedinger, K. R., Gordon, G. J., Jordan, P. W. & VanLehn, K. (2011), Instructional factors analysis: A cognitive model for multiple instructional interventions, *in* M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero & J. C. Stamper, eds, 'Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, July 6-8, 2011', www.educationaldatamining.org, pp. 61–70.

- URL:** http://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm2011_paper1_full_Chi.pdf
- Cho, M.-H. & Shen, D. (2013), 'Self-regulation in online learning', *Distance Education* **34**, 290–301.
- Choffin, B., Popineau, F., Bourda, Y. & Vie, J.-J. (2019), Das3h: Modeling student learning and forgetting for optimally scheduling distributed practice of skills, *in* 'International Conference on Educational Data Mining (EDM 2019)'.
- Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B. & Heo, J. (2020), Ednet: A large-scale hierarchical dataset in education, *in* 'International Conference on Artificial Intelligence in Education', Springer, pp. 69–73.
- Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C. & Lindstädt, S. (2018), Finding traces of self-regulated learning in activity streams, *in* 'Proceedings of the 8th International Conference on Learning Analytics and Knowledge', LAK '18, Association for Computing Machinery, New York, NY, USA, p. 191–200.
URL: <https://doi.org/10.1145/3170358.3170381>
- Cliff, N. (1993), 'Dominance statistics: Ordinal analyses to answer ordinal questions.', *Psychological Bulletin* **114**(3), 494–509.
- Corbett, A. T. & Anderson, J. R. (1994), 'Knowledge tracing: Modeling the acquisition of procedural knowledge', *User modeling and user-adapted interaction* **4**(4), 253–278.
- Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S. & Nardy, A. (2014), Mining texts, learner productions and strategies with readerbench, *in* 'Educational Data Mining', Springer, pp. 345–377.
- Dascalu, M., Dessus, P., Trausan-Matu, Ş., Bianco, M. & Nardy, A. (2013), Readerbench, an environment for analyzing text complexity and reading strategies, *in* 'International Conference on Artificial Intelligence in Education', Springer, pp. 379–388.
- Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P., Bianco, M. & McNamara, D. S. (2015), Readerbench: An integrated tool supporting both individual and collaborative learning, *in* 'Proceedings of the Fifth International Conference on Learning Analytics And Knowledge', LAK '15, Association for Computing Machinery, New York, NY, USA, p. 436–437.
URL: <https://doi.org/10.1145/2723576.2723647>
- Davis, D., Chen, G., Hauff, C. & Houben, G.-J. (2016), 'Gauging mooc learners' adherence to the designed learning path.', *International Educational Data Mining Society*.
- Dawson, P., Henderson, M., Ryan, T., Mahoney, P., Boud, D., Phillips, M. & Molloy, E. (2018a), 'Technology and feedback design', *Learning, design, and technology*.

- Dawson, P., Henderson, M., Ryan, T., Mahoney, P., Boud, D., Phillips, M. & Molloy, E. (2018b), 'Technology and feedback design', *Learning, design, and technology* .
- Dawson, P., Henderson, M., Ryan, T., Mahoney, P., Boud, D., Phillips, M. & Molloy, E. (2018c), Technology and feedback design, *in* 'Learning, design, and technology.'
- Dinno, A. (2017), 'Pacage dunn.test', <https://cran.r-project.org/web/packages/dunn.test/dunn.test.pdf>.
- Dowell, N. M., Graesser, A. C. & Cai, Z. (2016), 'Language and discourse analysis with coh-metrix: Applications from educational material to learning environments at scale', *Journal of Learning Analytics* **3**(3), 72–95.
- Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S. & Gardner, M. (2019), 'Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs', *arXiv preprint arXiv:1903.00161* .
- Dweck, C. S. (2000), *Self-theories: Their role in motivation, personality, and development*, Psychology press.
- Eagle, M., Corbett, A., Stamper, J., McLaren, B. M., Wagner, A., MacLaren, B. & Mitchell, A. (2016), Estimating individual differences for student modeling in intelligent tutors from reading and pretest data, *in* 'International Conference on Intelligent Tutoring Systems', Springer, pp. 133–143.
- Evans, C. (2013), 'Making sense of assessment feedback in higher education', *Review of educational research* **83**(1), 70–120.
- Farhana, E., Potter, A., Rutherford, T. & Lynch, C. F. (2021), 'Feedback and self-regulated learning in science reading', *International Educational Data Mining Society (In Press)* .
- Farhana, E., Rutherford, T. & Lynch, C. F. (2020a), 'Investigating relations between self-regulated reading behaviors and science question difficulty.', *International Educational Data Mining Society* .
- Farhana, E., Rutherford, T. & Lynch, C. F. (2020b), Understanding reading behaviors of middle school students, *in* 'Proceedings of the Seventh ACM Conference on Learning @ Scale', L@S '20, Association for Computing Machinery, New York, NY, USA, p. 385–388.
URL: <https://doi.org/10.1145/3386527.3405948>
- Flesch, R. (1948), 'A new readability yardstick.', *Journal of applied psychology* **32**(3), 221.
- Forrest-Pressley, D.-L. & Waller, T. G. (2013), *Cognition, metacognition, and reading*, Vol. 18, Springer Science & Business Media.
- Forrin, N. D., Mills, C., D'Mello, S. K., Risko, E. F., Smilek, D. & Seli, P. (2020), 'Tl; dr: longer sections of text increase rates of unintentional mind-wandering', *The Journal of Experimental Education* pp. 1–13.

- Fouh, E., Farghally, M., Hamouda, S., Koh, K. H. & Shaffer, C. A. (2016), 'Investigating difficult topics in a data structures course using item response theory and logged data analysis.', *International Educational Data Mining Society* .
- Garcia, T. (1995*a*), 'The role of motivational strategies in self-regulated learning.', *New Directions for Teaching and Learning* **63**, 29–42.
- Garcia, T. (1995*b*), 'The role of motivational strategies in self-regulated learning', *New Directions for Teaching and Learning* **1995**, 29–42.
- Garner, R. & Kraus, C. (1981), 'Good and poor comprehender differences in knowing and regulating reading behaviors.', *Educational Research Quarterly* .
- Garner, R. & Reis, R. (1981), 'Monitoring and resolving comprehension obstacles: An investigation of spontaneous text lookbacks among upper-grade good and poor comprehenders', *Reading Research Quarterly* pp. 569–582.
- Gaupp, R., Fabry, G. & Körner, M. (2018), 'Self-regulated learning and critical reflection in an e-learning on patient safety for third-year medical students', *International Journal of Medical Education* **9**, 189 – 194.
- Geden, M., Emerson, A., Rowe, J., Azevedo, R. & Lester, J. (2020), Predictive student modeling in educational games with multi-task learning, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 654–661.
- Ghosh, A., Heffernan, N. & Lan, A. S. (2020), Context-aware attentive knowledge tracing, *in* 'Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', pp. 2330–2339.
- Gitinabard, N., Heckman, S., Barnes, T. & Lynch, C. F. (2019), 'What will you do next? a sequence analysis on the student transitions between online platforms in blended courses', *International Educational Data Mining Society* .
- Goedecke, P. J., Dong, D., Shi, G., Feng, S., Risko, E., Olney, A. M., D'Mello, S. K. & Graesser, A. C. (2015), 'Breaking off engagement: Readers' disengagement as a function of reader and text characteristics.', *International Educational Data Mining Society* .
- Graesser, A. C., Jackson, G. T., Mathews, E., Mitchell, H. H., Olney, A., Ventura, M., Chipman, P., Franceschetti, D. R., Hu, X., Louwerse, M. M. & Person, N. K. (2003), Why/autotutor: A test of learning gains from a physics tutor with natural language dialog.
- Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. & Louwerse, M. M. (2004), 'Autotutor: A tutor with dialogue in natural language', *Behavior Research Methods, Instruments, & Computers* **36**(2), 180–192.
- Graesser, A. C. & McNamara, D. S. (2011), 'Computational analyses of multilevel discourse comprehension', *Topics in cognitive science* **3**(2), 371–398.

- Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H. & Pennebaker, J. (2014), 'Coh-metrix measures text characteristics at multiple levels of language and discourse', *The Elementary School Journal* **115**(2), 210–229.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004), 'Coh-metrix: Analysis of text on cohesion and language', *Behavior research methods, instruments, & computers* **36**(2), 193–202.
- Graesser, A. C., Singer, M. & Trabasso, T. (1994), 'Constructing inferences during narrative text comprehension.', *Psychological review* **101**(3), 371.
- Graesser, A. & McNamara, D. (2010), 'Self-regulated learning in learning environments with pedagogical agents that interact in natural language', *Educational Psychologist* **45**(4), 234–244.
- Greene, J. A. & Azevedo, R. (2007), 'A theoretical review of winne and hadwin's model of self-regulated learning: New perspectives and directions', *Review of educational research* **77**(3), 334–372.
- Greene, J. A., Bolick, C. M., Jackson, W. P., Caprino, A. M., Oswald, C. & McVea, M. (2015), 'Domain-specificity of self-regulated learning processing in science and history', *Contemporary Educational Psychology* **42**, 111–128.
- Greene, J. A., Dellinger, K. R., Tüysüzoğlu, B. B. & Costa, L.-J. (2013), A two-tiered approach to analyzing self-regulated learning data to inform the design of hypermedia learning environments, *in* 'International handbook of metacognition and learning technologies', Springer, pp. 117–128.
- Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991), *Fundamentals of item response theory*, Sage.
- Hashemyolia, S., Asmuni, A., Ayub, A. F. M., Daud, S. M. & Shah, J. A. (2014), 'Motivation to use self-regulated learning strategies in learning management system amongst science and social science undergraduates', *Asian Social Science* **11**, 49.
- Hattie, J. & Timperley, H. (2007), 'The power of feedback', *Review of educational research* **77**(1), 81–112.
- Hawk, T. F. & Shah, A. J. (2008), 'A revised feedback model for task and self-regulated learning', *The Coastal Business Journal* pp. 66–81.
- Heffernan, N. (2020), 'Project driver, *eat*', .
- Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E. et al. (2020), 'Enhancing student competency models for game-based learning with a hybrid stealth assessment framework.', *International Educational Data Mining Society* .

Hennig, C., Meila, M., Murtagh, F. & Rocci, R. (2015), *Handbook of cluster analysis*, CRC Press.

Herodotou, C., Heiser, S. & Rienties, B. (2017), 'Implementing randomised control trials in open and distance learning: a feasibility study', *Open Learning: The Journal of Open, Distance and e-Learning* **32**, 147 – 162.

Heylighen, F. & Dewaele, J.-M. (2002), 'Variation in the contextuality of language: An empirical measure', *Foundations of science* **7**(3), 293–340.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Hsieh, T. & Wang, T. (2010), 'A mining-based approach on discovering courses pattern for constructing suitable learning path.', *Expert systems with applications* **37**(6), 4156–4167.

Huang, Y., Yudelson, M., Han, S., He, D. & Brusilovsky, P. (2016), A framework for dynamic knowledge modeling in textbook-based learning, in 'Proceedings of the 2016 conference on user modeling adaptation and personalization', ACM, pp. 141–150.

Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y. & Hu, G. (2017), Question difficulty prediction for reading problems in standard tests, in 'Thirty-First AAAI Conference on Artificial Intelligence'.

Iraj, H., Fudge, A., Faulkner, M., Pardo, A. & Kovanović, V. (2020), Understanding students' engagement with personalised feedback messages, in 'Proceedings of the Tenth International Conference on Learning Analytics & Knowledge', pp. 438–447.

Jack C. Richards and J. P. & Platt, H. (1992), Longman dictionary of language teaching applied linguistics.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L. & Levy, O. (2020), 'Spanbert: Improving pre-training by representing and predicting spans', *Transactions of the Association for Computational Linguistics* **8**, 64–77.

Jurasky, D. & Martin, J. H. (2000), 'Speech and language processing: An introduction to natural language processing', *Computational Linguistics and Speech Recognition*. Prentice Hall, New Jersey .

Keenan, S. A. (1984), 'Effects of chunking and line length on reading efficiency', *Visible Language* **18**(1), 61.

Kiladis, G. N., Thorncroft, C. D. & Hall, N. M. (2006), 'Three-dimensional structure and dynamics of African easterly waves. Part I: Observations', *J. Atmos. Sci.* **63**(9), 2212–2230.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. & Chissom, B. S. (1975), Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Technical report, Naval Technical Training Command Millington TN Research Branch.

Kinnebrew, J. S. & Biswas, G. (2012), 'Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution.', *International Educational Data Mining Society*.

Kinnebrew, J. S., Biswas, G. & Sulcer, W. B. (2010), Modeling and measuring self-regulated learning in teachable agent environments, in '2010 AAAI Fall Symposium Series'.

Kinnebrew, J. S., Loretz, K. M. & Biswas, G. (2013), 'A contextualized, differential sequence mining method to derive students' learning behavior patterns', *JEDM| Journal of Educational Data Mining* **5**(1), 190–219.

Kintsch, W. & Van Dijk, T. A. (1978), 'Toward a model of text comprehension and production.', *Psychological review* **85**(5), 363.

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S. & Hatala, M. (2015), Penetrating the black box of time-on-task estimation, in 'Proceedings of the fifth international conference on learning analytics and knowledge', pp. 184–193.

Kruskal, W. H. & Wallis, W. A. (1952), *Use of ranks in one-criterion variance analysis*, Vol. 47, Taylor & Francis Group.

Kung, F. Y. & Scholer, A. A. (2018), 'Message framing influences perceptions of feedback (in directness)', *Social Cognition* **36**(6), 626–670.

Lee, D., Watson, S. L. & Watson, W. R. (2019), 'Systematic literature review on self-regulated learning in massive open online courses', *Australasian Journal of Educational Technology* **35**(1).

Leelawong, K. & Biswas, G. (2008), 'Designing learning by teaching agents: The betty's brain system', *International Journal of Artificial Intelligence in Education* **18**(3), 181–208.

Lester, J. C., Mott, B. W., Robison, J. L., Rowe, J. P. & Shores, L. R. (2013), Supporting self-regulated science learning in narrative-centered learning environments, in 'International handbook of metacognition and learning technologies', Springer, pp. 471–483.

Li, X., Liu, Y., Mao, J., He, Z., Zhang, M. & Ma, S. (2018), Understanding reading attention distribution during relevance judgement, in 'Proceedings of the 27th ACM International Conference on Information and Knowledge Management', pp. 733–742.

- Littlejohn, A., Hood, N., Milligan, C. & Mustain, P. (2016), 'Learning in moocs: Motivations and self-regulated learning in moocs', *The Internet and Higher Education* **29**, 40 – 48.
URL: <http://www.sciencedirect.com/science/article/pii/S1096751615300099>
- Liu, Q., Huang, Z., Huang, Z., Liu, C., Chen, E., Su, Y. & Hu, G. (2018), Finding similar exercises in online education systems, *in* 'Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', pp. 1821–1830.
- Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y. & Hu, G. (2019), 'Ekt: Exercise-aware knowledge tracing for student performance prediction', *IEEE Transactions on Knowledge and Data Engineering* **33**(1), 100–115.
- Louis Gomez, P. H. & Gomez, K. (2007), 'Integrating text in content-area classes: Better supports for teachers and students.', *Voices in Urban Education* **14**, 22–29.
- Luong, M.-T., Pham, H. & Manning, C. D. (2015), Effective approaches to attention-based neural machine translation, *in* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', pp. 1412–1421.
- Makany, T., Kemp, J. & Dror., I. E. (2009), 'Optimising the use of note-taking as an external cognitive aid for increasing learning.', *British Journal of Educational Technology* **40**(4), 619–635.
- Mann, H. B. & Whitney, D. R. (1947), 'On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other', *The Annals of Mathematical Statistics* **18**(1), 50 – 60.
URL: <https://doi.org/10.1214/aoms/1177730491>
- Mason, L. H., Meadan-Kaplansky, H., Hedin, L. & Taft, R. (2013), 'Self-regulating informational text reading comprehension: Perceptions of low-achieving students', *Exceptionality* **21**(2), 69–86.
- Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J. & Pardo, A. (2019), Analytics of learning strategies: Associations with academic performance and feedback, *in* 'Proceedings of the 9th International Conference on Learning Analytics & Knowledge', pp. 461–470.
- Matsuda, N., Griger, C. L., Barbalios, N., Stylianides, G. J., Cohen, W. W. & Koedinger, K. R. (2014), Investigating the effect of meta-cognitive scaffolding for learning by teaching, *in* 'International Conference on Intelligent Tutoring Systems', Springer, pp. 104–113.
- McNamara, D. S., Levinstein, I. B. & Boonthum, C. (2004), 'istart: Interactive strategy training for active reading and thinking', *Behavior Research Methods, Instruments, & Computers* **36**(2), 222–233.

- McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C. & Levinstein, I. B. (2007), 'istart: A web-based tutor that teaches self-explanation and metacognitive reading strategies', *Reading comprehension strategies: Theories, interventions, and technologies* pp. 397–421.
- Min, W., Mott, B. W., Rowe, J. P., Liu, B. & Lester, J. C. (2016), Player goal recognition in open-world digital games with long short-term memory networks., *in* 'IJCAI', pp. 2590–2596.
- Mirzaei, M., Sahebi, S. & Brusilovsky, P. (2019), Annotated examples and parameterized exercises: Analyzing students' behavior patterns, *in* 'International Conference on Artificial Intelligence in Education', Springer, pp. 308–319.
- Mongkhonvanit, K., Kanopka, K. & Lang, D. (2019), Deep knowledge tracing and engagement with moocs, *in* 'Proceedings of the 9th International Conference on Learning Analytics & Knowledge', pp. 340–342.
- Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S. & Paquette, L. (2018), Modeling learners' cognitive and affective states to scaffold srl in open-ended learning environments, *in* 'Proceedings of the 26th conference on user modeling, adaptation and personalization', pp. 131–138.
- Nagatani, K., Zhang, Q., Sato, M., Chen, Y.-Y., Chen, F. & Ohkuma, T. (2019), Augmenting knowledge tracing by considering forgetting behavior, *in* 'The world wide web conference', pp. 3101–3107.
- Nicol, D. (2010), 'From monologue to dialogue: improving written feedback processes in mass higher education', *Assessment & Evaluation in Higher Education* **35**(5), 501–517.
- Nicol, D. J. & Macfarlane-Dick, D. (2006), 'Formative assessment and self-regulated learning: A model and seven principles of good feedback practice', *Studies in higher education* **31**(2), 199–218.
- NRC (2012), A framework for k-12 science education: Practices, crosscutting concepts, and core ideas, *in* 'National Research Council', Washington, DC: National Academies Press.
- Odilinye, L. (2019), Personalized recommender system for technology enhanced learning using learners' metacognitive activities, PhD thesis, School of Computing Science, SIMON FRASER UNIVERSITY.
- O'Neil Jr, H. F. & Brown, R. S. (1998), 'Differential effects of question formats in math assessment on metacognition and affect', *Applied measurement in Education* **11**(4), 331–351.

- O'Reilly, T., Sinclair, G. & McNamara, D. S. (2004), istart: A web-based reading strategy intervention that improves students's science comprehension., *in* 'IADIS International Conference Cognition and Exploratory Learning in Digital Age (CELDA 2004)', pp. 173–180.
- Panadero, E. (2017), 'A review of self-regulated learning: Six models and four directions for research', *Frontiers in psychology* **8**, 422.
- Panadero, E., Klug, J. & Järvelä, S. (2016), 'Third wave of measurement in the self-regulated learning field: when measurement and intervention come hand in hand', *Scandinavian Journal of Educational Research* **60**(6), 723–735.
- Pandey, S. & Karypis, G. (2019), 'A self-attentive model for knowledge tracing.', *International Educational Data Mining Society*.
- Pandey, S. & Srivastava, J. (2020), Rkt: Relation-aware self-attention for knowledge tracing, *in* 'Proceedings of the 29th ACM International Conference on Information & Knowledge Management', pp. 1205–1214.
- Pardos, Z. A. & Heffernan, N. T. (2011), Kt-idem: Introducing item difficulty to the knowledge tracing model, *in* 'International conference on user modeling, adaptation, and personalization', Springer, pp. 243–254.
- Pardos, Z. A., Tang, S., Davis, D. & Le, C. V. (2017), Enabling real-time adaptivity in moocs with a personalized next-step recommendation framework, *in* 'Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale', pp. 23–32.
- Paris, S. G. & Paris, A. H. (2001), 'Classroom applications of research on self-regulated learning', *Educational psychologist* **36**(2), 89–101.
- Pavlik, P. I., Cen, H. & Koedinger, K. R. (2009a), Performance factors analysis –a new alternative to knowledge tracing, IOS Press, NLD, p. 531–538.
- Pavlik, P. I., Cen, H. & Koedinger, K. R. (2009b), Performance factors analysis–a new alternative to knowledge tracing, *in* 'Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling', IOS Press, pp. 531–538.
- Peckham, T. & McCalla, G. (2012), 'Mining student behavior patterns in reading comprehension tasks.', *International Educational Data Mining Society*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011a), 'Scikit-learn: Machine learning in python', *the Journal of machine Learning research* **12**, 2825–2830.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011*b*), 'Scikit-learn: Machine learning in python', *Journal of machine learning research* **12**(Oct), 2825–2830.
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. & Sohl-Dickstein, J. (2015), Deep knowledge tracing, *in* 'Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1', pp. 505–513.
- Pieschl, S., Stahl, E. & Bromme, R. (2008), 'Epistemological beliefs and self-regulated learning with hypertext', *Metacognition and Learning* **3**(1), 17–37.
- Pintrich, P. R. (1999), 'The role of motivation in promoting and sustaining self-regulated learning', *International Journal of Educational Research* **31**(6), 459 – 470.
URL: <http://www.sciencedirect.com/science/article/pii/S0883035599000154>
- Pintrich, P. R. (2000), The role of goal orientation in self-regulated learning, *in* 'Handbook of self-regulation', Elsevier, pp. 451–502.
- Poitras, E. G. & Lajoie, S. P. (2013), 'A domain-specific account of self-regulated learning: The cognitive and metacognitive activities involved in learning through historical inquiry', *Metacognition and Learning* **8**(3), 213–234.
- Poitras, E. G. & Lajoie, S. P. (2014), 'Developing an agent-based adaptive system for scaffolding self-regulated inquiry learning in history education', *Educational Technology Research and Development* **62**(3), 335–366.
- Rasch, G. (1960), 'Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.'
- Ravi, G. A. & Sosnovsky, S. (2013), Exercise difficulty calibration based on student log mining., *in* 'Proceedings of DAILE'.
- Reichle, E. D., Rayner, K. & Pollatsek, A. (2003), 'The ez reader model of eye-movement control in reading: Comparisons to other models', *Behavioral and brain sciences* **26**(4), 445.
- Rizopoulos, D. (2006), 'ltm: An r package for latent variable modeling and item response theory analyses', *Journal of statistical software* **17**(5), 1 – 25.
- Rotgans, J. & Schmidt, H. (2009), 'Examination of the context-specific nature of self-regulated learning', *Educational Studies* **35**(3), 239–253.
- Rovers, S. F., Clarebout, G., Savelberg, H. H., de Bruin, A. B. & van Merriënboer, J. J. (2019), 'Granularity matters: comparing different ways of measuring self-regulated learning', *Metacognition and Learning* **14**(1), 1–19.

- Rowe, A. D., Wood, L. N. et al. (2008), 'Student perceptions and preferences for feedback'.
- Rutherford, T., Long, J. J. & Farkas, G. (2017), 'Teacher value for professional development, self-efficacy, and student outcomes within a digital mathematics intervention', *Contemporary educational psychology* **51**, 22–36.
- Sabourin, J. L. (2013), *Stealth Assessment of Self-Regulated Learning in Game-Based Learning Environments*, PhD thesis, Department of Computer Sciences, North Carolina State University.
- Sabourin, J. L., Mott, B. W. & Lester, J. C. (2012), 'Early prediction of student self-regulation strategies by combining multiple models.', *International Educational Data Mining Society*.
- Sabourin, J., Shores, L. R., Mott, B. W. & Lester, J. C. (2012), Predicting student self-regulation strategies in game-based learning environments, in 'International Conference on Intelligent Tutoring Systems', Springer, pp. 141–150.
- Sadler, D. R. (1998), 'Formative assessment: Revisiting the territory', *Assessment in education: principles, policy & practice* **5**(1), 77–84.
- Salinger, T. (2003), 'Helping older, struggling readers.', *Preventing School Failure: Alternative Education for Children and Youth* **47**(2), 79–85.
- Schunk, D. H. & Zimmerman, B. J. (2012), *Motivation and self-regulated learning: Theory, research, and applications*, Routledge.
- Sheehan, K. M. (2015), 'Aligning textevaluator® scores with the accelerated text complexity guidelines specified in the common core state standards', *ETS Research Report Series* **2015**(2), 1–20.
- Sheehan, K. M., Kostin, I., Napolitano, D. & Flor, M. (2014), 'The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment', *The Elementary School Journal* **115**(2), 184–209.
- Sheshadri, A., Gitinabard, N., Lynch, C. F., Barnes, T. & Heckman, S. (2018), 'Predicting student performance based on online study habits: A study of blended courses.', *International Educational Data Mining Society*.
- Snijders, T. A. & Bosker, R. J. (2011), *Multilevel analysis: An introduction to basic and advanced multilevel modeling*, Sage.
- Snow, C. (2002a), *Reading for understanding: Toward an rd program in reading comprehension*, Santa Monica, CA: RAND Corporation,.

- Snow, C. (2002*b*), *Reading for understanding: Toward an R&D program in reading comprehension*, Rand Corporation.
- Snow, E. L., Jackson, G. T. & McNamara, D. S. (2014), 'Emergent behaviors in computer-based learning environments: Computational signals of catching up', *Comput. Hum. Behav.* **41**, 62–70.
- Sosnovsky, S., Brusilovsky, P., Lee, D. H., Zadorozhny, V. & Zhou, X. (2008), Re-assessing the value of adaptive navigation support in e-learning context, *in* 'International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems', Springer, pp. 193–203.
- Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C. & Rentfrow, J. (2019), Sequence multi-task learning to forecast mental wellbeing from sparse self-reported data, *in* 'Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', pp. 2886–2894.
- Spearman, C. (1961), 'The proof and measurement of association between two things.'
- Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S. & Hu, G. (2018), Exercise-enhanced sequential modeling for student performance prediction, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 32, pp. 2435–2443.
- Tapiero, I. (2007), *Situation models and levels of coherence: Toward a definition of comprehension*, Taylor & Francis.
- Tay, H. Y. (2015), 'Setting formative assessments in real-world contexts to facilitate self-regulated learning', *Educational Research for Policy and Practice* **14**(2), 169–187.
- Tests, G.-M. R. (1978), 'Chicago: Riverside', *MacGinitie, WH, MacGinitie, RK, Maria, K., & Dreyer, LG (2000). Gates .*
- Textstat (2020), 'Python PackageTextstat', <https://github.com/shivam5992/textstat>. Online; accessed 5 March 2021.
- Thaker, K., Carvalho, P. & Koedinger, K. (2019), Comprehension factor analysis: Modeling student's reading behaviour: Accounting for reading practice in predicting students' learning in moocs, *in* 'Proceedings of the 9th International Conference on Learning Analytics & Knowledge', pp. 111–115.
- Thaker, K., Huang, Y., Brusilovsky, P. & Daqing, H. (2018), Dynamic knowledge modeling with heterogeneous activities for adaptive textbooks, *in* 'The 11th International Conference on Educational Data Mining', pp. 592–595.

- Türkben, T. (2019), 'The effect of self-regulation based strategic reading education on comprehension, motivation, and self-regulation skills.', *International Journal of Progressive Education* **15**(4), 27–46.
- Tyler Rinker (2020), 'R Package qdap', <https://cran.r-project.org/web/packages/qdap/index.html>. Version 2.4.3.
- VanLehn, K., Jones, R. M. & Chi, M. T. (1992), 'A model of the self-explanation effect', *The journal of the learning sciences* **2**(1), 1–59.
- Veenman, M. V. (2007), 'The assessment and instruction of self-regulation in computer-based environments: a discussion', *Metacognition and Learning* **2**(2-3), 177–183.
- Vega, B., Feng, S., Lehman, B., Graesser, A. & D'Mello, S. (2013), Reading into the text: Investigating the influence of text complexity on cognitive engagement, *in* 'Educational Data Mining 2013'.
- Weaver, M. R. (2006), 'Do students value feedback? student perceptions of tutors' written responses', *Assessment & Evaluation in Higher Education* **31**(3), 379–394.
- Weber, G. & Brusilovsky, P. (2001), 'Elm-art: An adaptive versatile system for web-based instruction', *International Journal of Artificial Intelligence in Education* **12**, 351–384.
- Weinstein, C. E., Husman, J. & Dierking, D. R. (2000), Self-regulation interventions with a focus on learning strategies, *in* 'Handbook of self-regulation', Elsevier, pp. 727–747.
- Wen, M. & Rosé, C. P. (2014), Identifying latent study habits by mining learner behavior patterns in massive open online courses, *in* 'Proceedings of the 23rd ACM international conference on conference on information and knowledge management', pp. 1983–1986.
- Wenger, E. (2014), *Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge*, Morgan Kaufmann.
- Williamson, G. L. (2006), 'Aligning the journey with a destination: A model for k–16 reading standards.', *MetaMetrics, Durham, NC*.
- Winne, P. H. (2011), A cognitive and metacognitive analysis of self-regulated learning: Faculty of education, simon fraser university, burnaby, canada, *in* 'Handbook of self-regulation of learning and performance', Routledge, pp. 29–46.
- Winne, P. H. & Hadwin, A. F. (1998), 'Studying as self-regulated learning..', *The educational psychology series. Metacognition in educational theory and practice*.

Winne, P. H., Nesbit, J. C., Ram, I., Marzouk, Z., Vytasek, J., Samadi, D. & Stewart, J. (2017), 'Tracing metacognition by highlighting and tagging to predict recall and transfer.', *AERA Online Paper Repository*.

Winne, P. H. & Perry, N. E. (2000), Measuring self-regulated learning, in 'Handbook of self-regulation', Elsevier, pp. 531–566.

Winstone, N. E., Nash, R. A., Rowntree, J. & Parker, M. (2017), "it'd be useful, but i wouldn't use it': barriers to university students' feedback seeking and recipience', *Studies in Higher Education* **42**(11), 2026–2041.

Withey, C. (2013), 'Feedback engagement: forcing feed-forward amongst law students', *The Law Teacher* **47**(3), 319–344.

Woltman, H., Feldstain, A., MacKay, J. C. & Rocchi, M. (2012), 'An introduction to hierarchical linear modeling', *Tutorials in quantitative methods for psychology* **8**(1), 52–69.

Yore, L. D. (2012), Science literacy for all: More than a slogan, logo, or rally flag!, in 'Issues and challenges in science education research', Springer, pp. 5–23.

Yore, L. D. & Tippett, C. D. (2014), *Reading and Science Learning*, Springer Netherlands, Dordrecht, pp. 1–9.

URL: https://doi.org/10.1007/978-94-007-6165-0_130-2

Yudelson, M. V., Koedinger, K. R. & Gordon, G. J. (2013), *Individualized bayesian knowledge tracing models*, in 'International conference on artificial intelligence in education', Springer, pp. 171–180.

Zhang, J., Shi, X., King, I. & Yeung, D.-Y. (2017), *Dynamic key-value memory networks for knowledge tracing*, in 'Proceedings of the 26th international conference on World Wide Web', pp. 765–774.

Zheng, Y., Mao, J., Liu, Y., Ye, Z., Zhang, M. & Ma, S. (2019), *Human behavior inspired machine reading comprehension*, in 'Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 425–434.

Zhou, M. & Winne, P. H. (2012), 'Modeling academic achievement by self-reported versus traced goal orientation', *Learning and Instruction* **22**(6), 413–419.

Zhu, M., Lee, H.-S., Wang, T., Liu, O., Belur, V. K. & Pallant, A. (2017), 'Investigating the impact of automated feedback on students' scientific argumentation', *International Journal of Science Education* **39**, 1648 – 1668.

Zimmerman, B. J. (1989), 'A social cognitive view of self-regulated academic learning.', *Journal of educational psychology* **81** (3), 329.

Zimmerman, B. J. (2000a), *Attaining self-regulation: A social cognitive perspective*, in 'Handbook of self-regulation', Elsevier, pp. 13–39.

Zimmerman, B. J. (2000b), 'Self-efficacy: An essential motive to learn', *Contemporary educational psychology* **25**(1), 82–91.

Zimmerman, B. J. (2008a), 'Goal setting: A key proactive source of', *Motivation and self-regulated learning: Theory, research, and applications* **267**.

Zimmerman, B. J. (2008b), 'Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects', *American educational research journal* **45**(1), 166–183.

Zimmerman, B. J. & Bandura, A. (1994), 'Impact of self-regulatory influences on writing course attainment', *American educational research journal* **31** (4), 845–862.

Zimmerman, B. J., Bandura, A. & Martinez-Pons, M. (1992), 'Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting', *American educational research journal* **29**(3), 663–676.