ALL-SUBSET REGRESSION AS A MEANS FOR SELECTION OF SELF-REGULATED LEARNING PROCESSES MEASURED USING THINK ALOUD PROTOCOL DATA

Christopher A. Oswald

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Educational Psychology, Measurement, and Evaluation in the Schools of Education.

Chapel Hill
2018

Approved by:

Jeffrey A. Greene

Jill Hamm

Matt Bernacki

©2019
Christopher A Oswald
ALL RIGHTS RESERVED

ABSTRACT

Christopher A Oswald: All Subset Regression as a Means for Selection of Self-Regulated Learning Processes Measured Using Think Aloud Protocol Data
(Under the Direction of Jeffrey A. Greene)

During the 1990s computers were placed into most educational classrooms; however, they sat underused or not used at all. One reason for this is students lacked the skills to use computers effectively. One set of skills that can help students make use of computers is self-regulated learning. By using think aloud protocol analysis while students complete a task on the computer, a trace of their cognition, metacognition, and behavior can be created. Analyzing these traces, however, has proven difficult due to the high number of variables compared to the typical number of participants. A solution to dealing with this problem is to analyze all possible combinations of variables. In this thesis, I compared the results of two pre-existing variable reduction methods and Best All Subset Regression. It was found that Best All Subset Regression outperformed the existing methods, by fitting better models without diagnostic problems or extensive time demands. Best All Subset Regression also retained more information than the prior methods, so I suggest using it moving forward instead of the aggregation-based methods used previously.

ACKNOWLEDGEMENTS

I have been blessed to have great mentors to guide my development into research design and quantitative methods. I would like to thank to Mary Nelson for providing me with a solid foundation of statistics and guiding me in my early days, showing me how to get involved in research and the faults in the common ways people analyze data. I would like to thank Shane Murphy for allowing me to gain experience with several different methodologies and giving me experiences that few undergraduate students are granted. I would like to thank William Ware for helping me to see the beautiful simplicity of the general linear model and its extensions. I would also like to thank Jill Hamm that helped me to better understand how large texts and proposals differ from articles. I will would like to thank Matt Bernacki for coming into the second half of this thesis on no notice. While many get one expert, I am honored to have two of the names in my field of study on this committee. Finally, without Jeff Greene's immeasurable levels of guidance and support during my stay at UNC this thesis would not be possible. I came to UNC having research experience, but I am leaving a researcher in my own right.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The problem of variable selection is one of the most persistent and difficult problems in statistics (George, 2000; Ratner, 2010). In simple terms, variable selection is the process of separating meaningful information from extraneous information. In an educational framework, information is gathered about a great many things. For instance, information was gathered for 1,793 variables by the National Center for Education Statistics (2015) in the National Assessment of Educational Progress survey to study the area of mathematic achievement by fourth graders in the United States. These variables included student factors such as gender or race, teacher factors such as years of teaching and attitudes, and other factors such as peer relationships or use of time outside of school. Variable selection when used with these data involves finding a small subset of predictor variables, also referred to as a *regression model* that correlates with mathematical achievement in test scores (Ratner, 2010). As more and more data are gathered, in part due to the rise of big data, separating the meaningful predictors from the predictors that are not useful is increasingly important. Within the context of education, variable selection is the process of separating the meaningful variables related to learning from other variables that were also collected. Once the separation is done, the meaningful variables can be examined in more detail, whereas the ones that were not found meaningful could be put to the side. One area where this is an important issue is the area of skills that children need to succeed in education and the modern world (Hattie, 2009).

With the subsequent rise of the “information superhighway” in the 1990’s (Becker, 2000), school reformers claimed that computers and computer-based learning environments would revolutionize education (Sheingold, Hadley, & Thesiar Lieliillan, 1990). It was claimed that computers were naturally engaging and able to meet individual learners' needs through multiple representations of information, such as passages of text, diagrams, videos, and interactive elements, within a learning environment (Cuban, 2001). Computers were already such a large part of the lives of students it was claimed that the modern student is a *digital native* (Prensky, 2001). These students, Prensky claimed, thrived on multitasking and could process several different sources of information at the same time with ease. This would cause the students to seek out non-linear computer-based learning environments to get instant access to the information they wanted in the format they wanted the information in (Prensky, 2001). For instance, a student interested in the brain would prefer to go to a website, find a diagram of the brain, and then follow links to pages with more information on individual parts.

In the last decade researchers' into computer-based instruction have found that digital natives are more mythical than real, however (Bennett, Maton, & Kervin, 2008; Selwyn, 2009). Prensky (2001) predicted that the digital native would be able to use technology with ease, engage multiple sources and critically review their differences, and compare and contrast arguments (Margaryan, Littlejohn, & Vojt, 2011). Subsequently, researchers found that, whereas digital natives could use familiar tools like word processing or email with ease, they struggled when using more advanced features of technology. For example, students tended to prefer more traditional, non-computer-based pedagogies and had trouble interacting with elements such as Blackboard or university-based computer interfaces (Margaryan, Littlejohn, & Vojt, 2011).

Researchers studying computer use in the classroom have found that not all computer use is effective at prompting learning. Scheiter et al. (2009) found that students who watched a realistic visualization of cellular mitosis had far worse scores than students given a schematic of how mitosis worked. Darabi, Nelson, and Palanki (2007) found similar results where students who worked with a simulated water treatment experiencing malfunctions did not gain as much knowledge of the nature and cause of the malfunctions and how to best fix them when compared to controls groups who did not use the simulation. These negative findings demonstrate that in more complex computer-based learning environments, computers use by itself will not increase learning and may not lead to the planned learning outcomes.

Self-regulated Learning

Researchers, wanting to see the benefits of computers-based learning environments, began to rethink research about computers. Instead of assuming students would natively be able to work with computers, researchers instead asked what skills are needed to be most effective at learning with computers. One set of skills that researchers study as students learn with computer-based learning environments is self-regulated learning (Pintrich, 2000).

Self-regulated learning can best be described as “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation and behavior, guided and constrained by their goals and the contextual features in the environment” (Pintrich, 2000, p. 453). Several components stand out in this definition. First, learners that employ self-regulated learning skills are actively involved in their own knowledge construction. Second, they engage in goal-directed behavior that defines the task they are trying to complete. Next, they engage in metacognitive monitoring and metacognitive control while learning. Flavell defined metacognition as “knowledge and cognition about cognitive phenomena” (Flavell, 1979, p. 906). Metacognition is used to monitor and regulate motivational, cognitive, and behavioral strategies while problem solving, monitoring their effect. Next, learning is guided by and constrained by the environment. For example, knowledge that students use while learning is limited to the learning materials they have available for use during that learning task. Finally, students need to be motivated to engage in self-regulated learning (de Boer, Donker-Bergstra, & Kostons, 2012). Self-regulated learning use has been shown to influence student learning with a moderate effect size (Hattie, 2009).

Research has shown that self-regulated learning is particularly effective for learning in digital environments, where non-linear design and a depth of options can negatively affect student learning (Azevedo, Moos, Johnson & Chauncey, 2010; Greene & Azevedo, 2007). Researchers have found that during learning tasks using digital environments, students' self-regulated learning use is associated with better learning outcomes (Littlejohn, Hood, Milligan & Mustain, 2016; Zimmerman, 2008). Furthermore, scaffolding (Devolder, van Braak & Tondeur (2012) or even prompting learners (Bannert & Reimann 2012; Bannert et al., 2015; Müller & Seufert, 2018) to use self-regulated learning during online learning tasks has also been shown to increase student performance on learning tasks.

Problems with Measuring Self-Regulated Learning

Researchers demonstrated that one strength of the field of self-regulated learning is that it incorporates many areas (e.g., motivation, goal setting, self-evaluation) that positively relate to learning and achievement (Pintrich, 2000). However, as self-regulated learning theory draws from many different fields of study, it is often hard to break down and separate which self-regulated learning processes are important to learning during specific tasks (Boekaerts, 1996). This leads to fragmentation (Zeidner, Boekaerts, & Pintrich, 2005), making comparing studies confusing as authors examine self-regulated learning through different theoretical lenses, each with their own labels, terminology, and construct definitions (Dent & Hoyle, 2015). Not only is this a problem, but the self-regulated learning processes used by learners may vary across academic domains (Greene, Bolick, & Robertson, 2010; Moos & Miller, 2015; Wolters & Pintrich, 1998), as well as vary from task to task (Lichtinger & Kaplan, 2015; McCardle & Hadwin, 2015; Vandevelde et al., 2015).

One method that researchers use to study self-regulated learning processes is *think-aloud protocols* (Bannert, Reimann, & Sonnenberg, 2014; Greene, Robertson & Costa, 2011; Greene et al., 2015; Moos & Miller, 2015; Schellings, van Hout-Wolters, Veenman, & Meijer, 2013; Vandevelde et al., 2015). Think aloud protocols is methods of collecting data where students engage in a task while verbalizing their thoughts (Greene, Robertson, et al., 2011). This method allows researchers to document self-regulation processes as they occur (Greene, Dellinger, Tüysüzoğlu, & Costa, 2013). Empirical evidence shows that the results from think-aloud protocol analyses are more accurate measures of self-regulated learning than self-report survey measures, as survey methods often require learners to try to remember or judge how often they engage in certain actions after the fact or estimate how often they typically use certain strategies or engage in various actions (Veenman, 2007; Winne & Jamieson-Noel, 2003; Winne, Jamieson-Noel, & Muis, 2000). Without an existing high level of metacognition skills and ability, learners would not make accurate assessments of how often they did engage in actions.

The problem with think-aloud methods is two fold. First, it can take a great deal of time to conduct the experiment, transcribe what was spoken aloud, and perform the protocal analysis itself. The protocal analysis most used in self-regulated learning research involves a list of codes where each can be matched to verbalizations of students' self-regulated learning use (Greene, Robertson, et al., 2011). These lists of codes can range from just coding at a macro-level for planning, motivation, or using strategies to a fine micro-level, where these macro-levels are subdivided into 50 or more codes (Greene et al., 2015) with no upper limit. This is complicated further because self-regulated learning incorporates elements from so many different fields (Boekaerts, 1996), some of these studies involve measurements of more processes than there are participants, which leads to problems when it comes to analyzing the data (Greene et al., 2013; 2014). Traditional statistics assumes that there will be more subjects than predictors, but in some of these studies there are more predictor variables than there are subjects, making common methods of analysis difficult to employ (Hastie, Tibshirani, & Friedman, 2009).

Currently, two methods exist to try to deal with transforming the data from the protocol analysis into a version that researchers can analyze using traditional methods (Greene et al., 2013, 2015). The first is a method of full aggregation where researchers examine self-regulated learning on a theory-based macro-level. The second method is data-driven aggregation that involves finding the processes that most predict learning gains and those most predictive of learning losses and then combining them into new predictor variables that are a linear combination of the skills that compose them (e.g., Greene et al., 2015).

Purpose of this Thesis

In this thesis, I propose a new method for determining which self-regulated learning processes are most predictive of learning for a specific learning task. Whereas traditional regression methods fit data to a defined model, I will be employing all-subset regression to let the data define a model (Ratner, 2010). In the past, researchers considered this method too computationally heavy for use with datasets over 20 (Miller, 2002) to 40 variables (Hastie et al., 2009), but recent advances in statistical software have allowed this method to be employed without intensive computational or time resources (Yang, 2013). Researchers have attempted to determine, indirectly, a best subset of self-regulated learning processes that were predictive of learning by using data-driven aggregation that used predictors correlated with task performance measures (Greene et al., 2014, 2015), but my proposed method should find the model with the best set of predictors, i.e., the predictors that maximize adjusted R^2.

Potential Contributions

The two contributions of this study are in the areas of subset selection and model assessment. Currently, researchers in this area focus on understanding the direct impact of a pre-defined model and use data driven models as a secondary method to estimate other models that may produce a better fitting model (Greene et al., 2015; Greene, Yu, & Copeland 2014). The goal of this study is to change this question to asking what is the best model that the data can produce. Researchers can use all subset regression to directly find the variable set with the greatest effect size.

Using best all subset regression is important for several reasons. First, researchers using this method can produce models that are useful in situations where there are more processes than participants (known as $p > n$). This is becoming more and more problematic in recent studies (Greene et al., 2014, 2015), so researchers need a method to pick which processes are useful as the process list grows, because it is not possible to calculate a full regression model in situations where $p > n$. Second, in statistical learning there are many methods to deal with the problem of subset selection, such as penalized regression, random forests, and multivariate adaptive regression splines. However, they come with a tradeoff in that they are not as easy to understand in terms of how they get their solutions, or how to interpret their solutions (Hastie, Tibshirani & Friedman, 2009). The best all subset regression method can provide a starting point to determine if the increase in complexity of interpretation is worth the improvements in model fit.

This method also allows for a way to compare the results of several studies. Researchers can use this method on several of the studies that have used the Azevedo and Greene and colleagues' method of think aloud protocol analysis (Azevedo et al., 2004; Azevedo & Cromley, 2004b; Greene & Azevedo, 2009; Greene et al., 2010), and then compare the results across studies. The learning task and knowledge measures are aligned across these studies, with each study using the same learning task about the human heart, and the same methods to assess learning. Despite some variation in the skills used, the skills in each study generalize back to the same areas of planning, monitoring, and strategy use, with most skills from the early studies being present in the later studies. So far, no attempts have been made to compare the results of these studies in a systematic manner, but this method should provide a uniform method to determine in each study which skills predict learning in that study, and the degree to which they do so, with all else held constant by examining the best models that can be created from each study and the beta weights of the skills selected. This benefit is, of course, limited to studies that use the same general methodology, learning task, knowledge measures and codebook, therefore researchers should exercise caution extending it beyond these studies.

Research Question

This thesis will examine the following research question:

1. Which variable selection method (full aggregation, data-driven aggregation, or best all subset regression) best fits the relationship between self-regulated learning and knowledge gain in the examined dataset, defined as the one that maximizes adjusted R^2.

CHAPTER TWO: LITERATURE REVIEW

Studying self-regulated learning use with computer-based learning environments requires a task and three models: a theoretical model, a measurement model, and a statistical model (Schraw, 2010). In this chapter, I briefly describe each model, beginning with a review of what computer-based learning environments are and how self-regulated learning skills could improve a learner's ability to learn within a computer-based learning environment. Next, I present the theoretical models of self-regulated learning, with a focus on the Winne and Hadwin model of self-regulated learning as that is the theoretical model used for this study. Then, I discuss measuring self-regulated learning processes, first modeling them as an attribute, as older theories have done (Pintrich, Smith, Garcia, & Mckeachie, 1993; Weinstein, Schulte, & Palmer, 1987), then as event, as the Winne and Hadwin Model of self-regulated learning does (Winne, 2010). Following that discussion of the measurement model for this thesis, I describe the measurement protocol for this study, which employs data generated from think-aloud protocols (Chi, 1997; Ericsson & Simon, 1980; Greene & Azevedo, 2010). Finally, I present the statistical model that I use in this thesis, all-subset regression, and demonstrate how it is an incrementally better method than the current models researchers use in analyses using the same theoretical and measurement models.

Computer Based Learning Environments and their Challenges

Computer based learning environments, such as hypermedia environments, offer an advantage over traditional classroom learners in that they can offer different sequences and modes for accessing information using multiple representations, such as texts, videos, and interactive applications. Learners control the paths they take through computer-based learning environment by a series of text or image links that connect different sections of the learning environment. However, while the wide variety of options presented to learners may seem like a good thing, often the reality is that the lack of a linear structure in online learning environments hinders learning (Scheiter & Gerjets, 2007). As information becomes more complex or there is a lot of elements to interact with, students with poor metacognitive skills are not able to best use the learning material, as their working memory becomes overloaded (Scheiter & Gerjets, 2007). When learning environments use different representations that lack integration, such as textual features and video features, the competition for learners's working memory increases as they encounter more and more elements (Swezller & Sweller, 1994). This situation creates what Salomon (1998) called the "butterfly defect," where learners click from link to link within a computer based learning environment, finding interesting but often irrelevant information (Kirschner & van Merriënboer, 2013).

Adding to the challenges associated with computer-based learning environments, students with low prior knowledge encounter more problems than students with high prior knowledge (Chen, Fan & Marcredie, 2006). Chen, Fan, and Marcredie (2006) found that students with high prior knowledge often used directed searches for the information they sought, examined the big picture the environment tried to present, and then moved more into the deep structures of a computer-based learning environment to find the best solution for a task. Novices, however, often followed links sequentially from the starting page, looked at the surface features of a page such as headings and bolded terms, examined the environment as pages of separate topics, and primarily sought to find any solution to their task. Such behaviors lead to worse learning outcomes than those used by students with high prior knowledge, and an incomplete understanding of the material presenting in the task. What these students need most are skills that will help them deal with the cognitive demands of high information-learning environments, as well as tools to navigate through nonlinear systems. One set of skills that has been associated with better navigation in computer-based learning environments is self-regulated learning (Duffy & Azevedo, 2015; Greene et al., 2015; Zhou & Winne, 2012).

Theoretical Models of Self-Regulated Learning

Given there are many conceptualizations and definitions of self-regulated learning (Boekaerts, 1996), a simple definition of the construct is as follows: “self-regulated learners are generally characterized as active, efficiently managing their own learning through monitoring and strategy use” (Greene & Azevedo, 2007a, p. 334). Zimmerman (2001) found that a common conceptualization of the construct is the degree to which learners are metacognitively, motivationally, and behaviorally active in their own learning. He found three features that were key to self-regulated learning: first, self-regulated learners actively use metacognitive, motivational, and behavioral strategies; second, they employ a feedback loop to monitor their use of strategies and react to this feedback; and third, self-regulated learners have the motivation to engage in self-regulatory activity. Taking these three features together, self-regulated learners set goals and select and employ self-regulated strategies to achieve these goals using feedback to guide them (Zimmerman, 2001, 2013).

One view on self-regulated learning is that it is the intersection between cognitive and metacognitive theory (Dinsmore & Zoellner, 2018). From cognitive theory comes strategy usage. Mayer defines *cognitive strategies* as “cognitive processes that the learner intentionally performs to influence learning and cognition” (2001, p. 86). These processes can be simple things such as general learning methods, including taking notes, to comparing the different narratives presented in a historical task (Greene et al., 2015).

If strategy use is the *how* of self-regulated learning, *metacognition* is the *when, where*, and *why* to use these strategies. Metacognition research began with Flavell's (1971) and Brown's (1977) studies into learners' knowledge of cognition and regulation of cognition during learning. From this base, metacognition researchers quickly become focused on two major areas, *metacognitive knowledge* and *regulation of metacognition* (Veenman, van Hout-Wolters, & Afflerbach, 2006). There are three areas of metacognitive knowledge often discussed. The first is *declarative knowledge*. This is knowledge about cognition as well as cognitive and metacognitive strategies (Schraw & Dennison, 1994), as well as factual knowledge (Schraw, Crippen, & Hartley, 2006). *Procedural knowledge* is knowledge about how to do things (Schraw, 2006). In metacognition, the knowledge of how to implement strategies for breaking up and solving problems is called procedural knowledge (Pressley & Harris, 2006). Finally, *conditional knowledge* represents the *when, where*, and *why* of using procedural knowledge and declarative knowledge (Schraw, 2006). Conditional knowledge is part of the broader knowledge construct of self-regulatory knowledge, which is composed of not only knowing when and where to apply strategies and declarative knowledge but also knowledge related to the regulation of metacognition, such as planning, monitoring, and evaluation (Mayer & Wittrock, 2006).

Another area of metacognition is *metacognitive experiences*. Flavell (1979) saw metacognitive experiences as overlapping with metacognitive knowledge but still being distinct. Efklides (2009) defined metacognitive experiences as "manifestations of online monitoring of cognition as the person comes across a task and processes the information related to it" (Efklides, 2009, p.78). These experiences include feelings of not understanding information presented as well as when learners feel they have encountered information previously. These experiences can lead to strategy use as well as the creation and modification of goals. One of the most common metacognitive experiences that occurs in everyday life is the "tip of the tongue" state where a learner is sure that they have knowledge of something but cannot recall what it is (A. S. Brown, 1991). For educators, two of the most important metacognitive experiences are feelings of knowledge and judgments of learning (Schwartz, 1994). A *feeling of knowing* is the belief that a piece of information can be recalled from memory. These feelings occur before a learner tries to retrieve knowledge from memory. *Judgments of learning* are cognitive judgments that learners will remember what they have learned at a future point in time (Narens, Jameson, & Lee, 1994). Learners can use these estimates to determine how effective their learning has been and whether they need to change or modify their strategies. Within the context of self-regulated learning, negative judgments of learning, in which a learner feels they do not understand what they just encountered, can prompt the learner to engage in adaptive metacognition and change their strategies for solving a task (Binbasaran Tuysuzoglu & Greene, 2015).

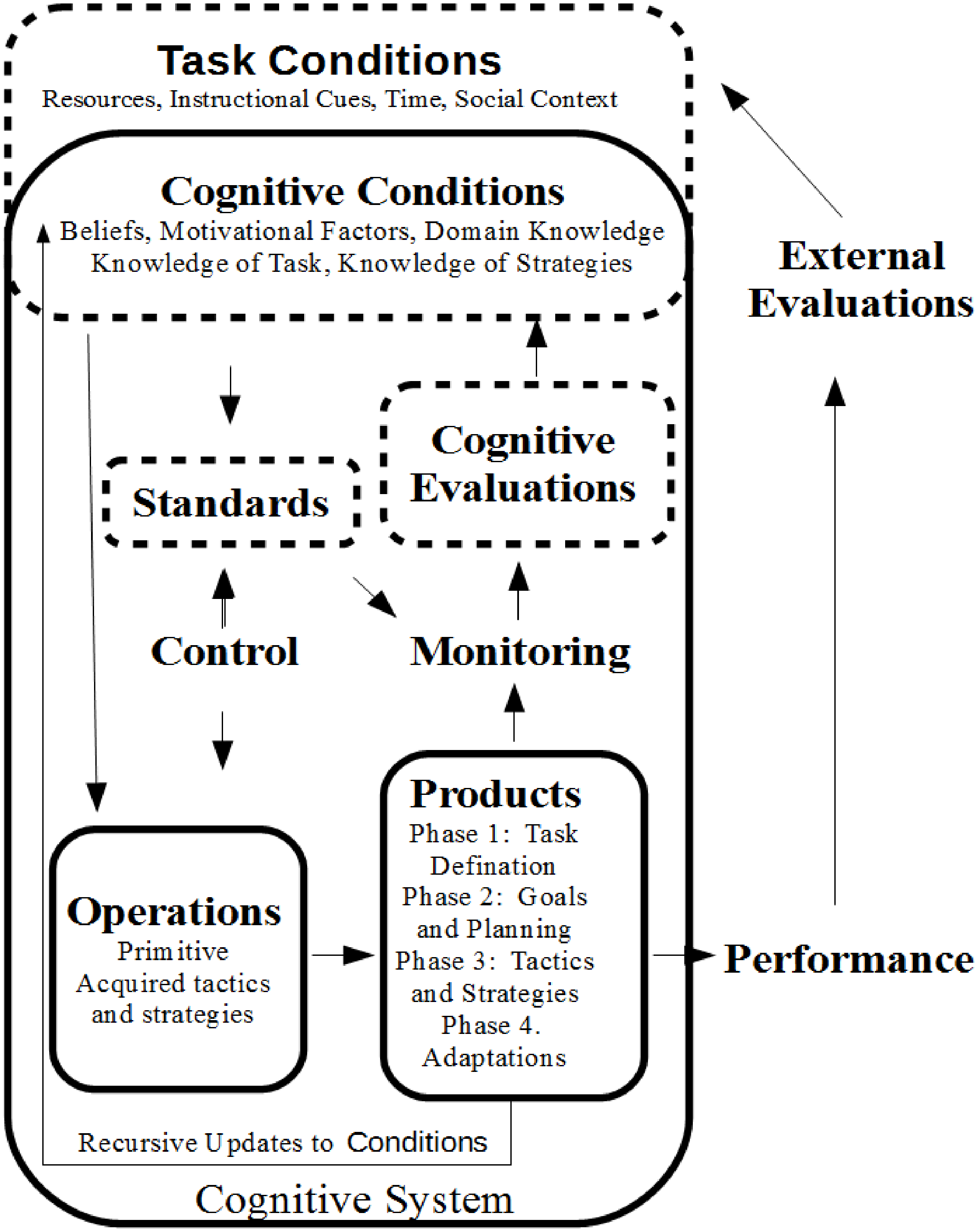
One area of self-regulated learning that is not so easy to pin down is motivation. While cognition and metacognition remain key parts of self-regulated learning theory, the role of motivation varies from model to model. In some models, motivation is infused throughout. This was the case with Pintrich's model where the motivation factor interacts with contextual, cognitive, and behavioral factors (Schunk, 2005). Boekaerts' (1996) model of self-regulated learning separates out cognitive and motivational aspects in self-regulated learning into two distinct paths. Zimmerman's (2013) model present motivation as belief or factors that occur in the forethought phase. These differences influence how researchers study motivation's role in self-regulated learning (McDuffy & Azevedo, 2013), as well as different empirical conclusions on the role it plays in each model. Whereas strategy use and metacognition are well defined concepts in self-regulated learning theory, motivation is less defined in terms of what it empirically means. Greene and Azevedo (2007) defined motivation in self-regulated learning theory as including goal orientation, self-efficacy, expectancy-value theory, self-determination theory, and interest whereas Boeakarts (1996) went further, classifying most of what Greene and Azevedo focused on into one area, motivational beliefs, and adding to them motivational strategies and motivational self-regulation, noting that overlap and broad descriptions of motivation made studying the distinct effects of motivation extremely difficult.

The Winne and Hadwin Model of Self-Regulated Learning

The Winne and Hadwin (1998) model of self-regulated learning is based on information processing theory, and has been used by researchers to study self-regulated learning in computer-based learning environments (Duffy & Azevedo, 2015; Greene et al., 2015; Zhou & Winne, 2012). For this study, it will be the model used to produce the targets of measurement for analysis that define what skills modern learners need. The model consists of five aspects referred to as COPES, an acronym for conditions, operations, products, evaluations, and standards (Winne, 2001). *Conditions* are the resources and constraints the learner has to work within as they complete a task. These conditions include declarative, procedural, and metacognitive knowledge and can include goal orientation, time constraints, background knowledge, and knowledge of tactics and strategies. *Operations* consist of the different cognitive and metacognitive strategies a learner employs during the task and include the procedural knowledge-based strategies that can range from primitive cognitive abilities to complex multilevel strategies. *Products* are units of information generated by the operations and include plans for engaging in operations, such as changes to the learner's knowledge or external products such as notes or test answers. *Evaluations* are knowledge products produced by metacognitive monitoring that are used to compare standards to products of operations. Evaluations include judgments made about learning, utility of tactics, efficacy, and attributions about products. *Standards* are the qualities that good products consist of, by which knowledge products are evaluated. Standards can be created by prior knowledge, such as past performance in graded tasks, and are influenced by effort and utility thresholds and motivational orientations (Winne, 2001).

The four phases of the Winne and Hadwin Model are (a) defining the task, (b) goal setting and planning, (c) enacting tactics, and (d) adapting metacognition (Winne, 2001). The phases are defined by the products they produce. During the task definition stage, students use their prior knowledge and the information that prompts the task to create a personal understanding of the task. Cognitive conditions in this stage are beliefs the learner has and their metacognitive, procedural, and declarative knowledge about the task or similar tasks. These conditions are used in the second phase to define goals that will guide task completion. These goals are updated as the task goes on and new knowledge is constructed. In the third phase, learners enact strategies from their procedural knowledge and limited by the conditions to solve the task. Here, products are compared to the goals by evaluating the products and standards to see if the products have met the standards defined in Phase 2 to complete the goal. Internal feedback drives this stage, as metacognitive monitoring is used to determine if the methods used to complete the task are effective or not, and metacognitive control is used if the learner needs to change tactics. As the learner gains more knowledge about the task, they may go back to Phase 2 and redefine the goals. The final stage is adaptive metacognition. In this stage, the learner's procedural and metacognitive knowledge may be used to allow the learner to change which operations they use to solve a problem, re-conceptualize operations, or change their cognitive conditions by changing their beliefs or knowledge. These stages are not linear, and after the task begins the learner may moves between the first three phases freely (Winne, 2001). This model is presented visually in Figure 1.

Figure 1: Winne and Hadwin Model of Self-Regulation


Task Conditions
Resources, Instructional Cues, Time, Social Context
Cognitive Conditions
Beliefs, Motivational Factors, Domain Knowledge
Knowledge of Task, Knowledge of Strategies
External Evaluations
Standards
Cognitive Evaluations
Control
Monitoring
Products
Phase 1: Task Defination
Phase 2: Goals and Planning
Phase 3: Tactics and Strategies
Phase 4. Adaptations
Operations
Primitive
Acquired tactics and strategies
Performance
Recursive Updates to Conditions
Cognitive System

Of the model's four stages, it is the first three that are of interest to this study. These stages are (a) task definition, (b) planning, and (c) the enacting of tactics and strategies. The model also provides a framework for the targets of measurements. The COPES framework's operations, products, and evaluations will be focused on for this study. While this model has a heavy information-processing focus, it shares many aspects of other cognitive models (Winne & Perry, 2000; Zimmerman, Heart, & Mellins, 1989); metacognitive monitoring and control, planning, evaluation, and active use of strategies are not unique to the Winne and Hadwin model. Therefore, the use of this model is to be a guide to examine self-regulated learning from not just an information processing view but a general cognitive one as well.

Effectiveness of self-regulated learning. Researchers have demonstrated the effectiveness of self-regulated learning in several meta-analyses. Dignath, Buettner and Langfieldt (2008) found an overall mean effect size of $g = .62$ for self-regulated learning interventions on learning outcomes in primary educational tasks, and a follow-up meta-analysis by Dignath and Buettner (2008) found a similar result for the effects of self-regulated learning interventions on performance outcomes of $d = .68$ for primary school learners and $d = .71$ for secondary school learners. Donker, Boer, Konstons, Van Ewvik, and van der Werf (2014) found that across domains and grade levels effect size for self-regulated learning interventions was $d = .66$. Examining these results, the estimated range for the effectiveness of self-regulated learning interventions is between $d = .62$ and $d = .71$, which Cohen (1992) defined as a medium to large effect size. Hattie (2009) found in education the average effect size for any intervention was $d = .4$; therefore, interventions based on self-regulated learning are above average in terms of effectiveness. The evidence shows that not only are self-regulated learning interventions effective, but they produce positive outcomes over numerous studies.

Measurement Methods for Studying Self-Regulated Learning

There have been many methods suggested to measure self-regulated learning, ranging from surveys and interviews to behavioral traces and direct observation (Winne & Perry, 2000). This section starts with a brief review of self-regulated learning measurement methods, first as an aptitude, then as an event. Aptitude measures of self-regulated learning assume that self-regulated learning is a set of stable abilities or predispositions (Winne & Perry, 2000). For instance, a student may score high on a survey measuring motivation and planning, meaning in a future task they may be highly motivated and engage in many plans. However, researchers have started to move away from this assumption, as self-regulated learning has begun to be reconceptualized as a dynamic, contextual system of events (Ben-Eliyahu & Bernacki, 2015).

Aptitude measures of self-regulated learning. During the late 1980s and early 1990s, researchers assessed self-regulated learning as an aptitude that was defined as a metacognitive, motivational, and behavioral construct (Zimmerman, 2008). The Learning and Strategies Inventory or LASSI, (Weinstein et al., 1987) is an 80-item self-report inventory of strategies that students use during their studying that employs a 5-point Likert scale measuring how true the questions were of the students studying patterns. Subscales produced by this measure include concentration, selecting main ideas, information processing, motivation, attitude, anxiety, time management, study aids, self-testing, and test strategies (Zimmerman, 2008).

A second survey developed to measure self-regulated learning, and the most common way of measuring self-regulated learning as an aptitude in a computer-based learning environment context (Saks & Leijen, 2014), is the Motivated Strategies for Learning Questionnaire, or MLSQ, (Pintrich et al., 1993). This is an 81-item self-report survey that has students respond to questions on a 7-point Likert scale describing how true the questions are of them. This measure has three main sections: a motivation scale, a cognitive scale, and a resource management scale. Subscales of the motivation scale include measures of intrinsic goal orientation, extrinsic goal orientation, task value, control of learning beliefs, self-efficacy for learning and performance, and text anxiety. The learning strategies scale has measures of rehearsal, elaboration, organization, critical thinking, and metacognitive self-regulation. The resource management scale includes measures of time and study environment management, effort regulation, peer learning, and help seeking. Researchers have used this survey to study many different aspects of motivation and self-regulated learning in a variety of domains such as middle school physical education students, female engineering students, and gifted high school students. Researchers in different fields also have employed this survey, such as in research on course structure, cooperative learning, multimedia design, and video teleconferencing (Duncan & Mckeachie, 2005 Ben-Eliyahu & Bernacki). The strongest aspect of this scale is that it is so widely employed: at the time of one study's publication, Duncan and Mckeachie (2005) found on Google hundreds, if not thousands, of results for the use of this survey.

Researchers found these tools were effective at predicting achievement (Pintrich et al., 1993; Zimmerman & Bandura, 1994; Zimmerman & Martinez-Pons, 1986). A meta-analysis of studies using the MSLQ found that it did predict student achievement with a low to moderate effect size (Credé & Phillips, 2011). These studies using self-report data helped establish the field of self-regulated learning by first showing that there was a link between self-regulated learning and achievement, improved the construct formation of what is self-regulated learning, and identified areas for future study (Duncan & Mckeachie, 2005).

Problems of attribute-based measures of self-regulated learning. By the late 1990s, some doubts had arisen regarding the measurement of self-regulated learning as an aptitude. The main critique of these methods was that they are not accurate (Winne, 2010; Winne & Perry, 2000), do not capture the conceptual nature of self-regulated learning (Greene, Robertson, & Costa, 2011), and rely on retrospective recall of events (Veenman, 2011).

Winne and Jamieson-Noel (2003) argued that students' accounts of their use of study tactics may not be accurate. They asked students to self-report strategy use while partaking in a computer task that also directly measured whether that strategy occurred using trace data, such as making a note, highlighting text, or reviewing information. The results showed that, on average, the correlation between the self-reports and students' actions was extremely low. Some students reported using strategies they did not actually use. Students could not determine accurately whether the they took a note, copied text to a note, or highlighted text. The students' self-report measures and actual event measures correlations ranged from r = .0 to r = .44. For taking notes, the relationship was r = .72; for copying text to a note, r = .67; and for highlighting, r = .54. For some areas, self-reports on strategy use were incorrect, and at best they were inaccurate, overreporting most strategies and underreporting note taking (Winne & Jamieson-Noel, 2003). This finding should not be unexpected. Students with poor metacognitive skills are expected to have trouble being able to accurately monitor their own actions; this raises serious questions about the validity of self-report measures.

Furthermore, these measures gather information after the fact, having students make judgments about how often they think that an event has occurred. However, memory is not perfect, and distortion can occur when researcher ask students what skills they employed and they remember engaging in actions they did not actually perform or that they later have no memory of (Nisbett & Wilson, 1977). Again, students with poor metacognitive skills may simply not know how often they engage in a behavior or certain types of cognitive processes, as they do not actively monitor their own cognition or behavior. However, even students with good metacognitive skills may report memories that are not accurate accounts of a past event (Veenman, 2011).

Researchers have also raised questions as to whether a one-time measurement can accurately access a learners self-regulated learning aptitude (Winne & Perry, 2000), as aptitude measures could differ over domains, across tasks, and even within a learning task due to the dynamic nature of self-regulated learning processing (Winne, 2010). For instance, a learner with high prior knowledge on a certain time-limited task may engage in a lot of metacognitive monitoring early in the task to review what prior knowledge they have about the task. Then they may not use much monitoring over the rest of the task, as their early use of monitoring would enable them to set up a detailed plan for the completion of the task, along with the strategies they already know from prior knowledge will be effective. However, if this student was in a situation where they had low prior knowledge, they may initially focus on a plan, then engage in more metacognition as they monitor if that plan is effective. Strategy use will vary far more as the learner adapts to how well they believe their strategy is working. In these two situations, the learner will show two different aptitudes for self-regulated learning.

Measuring Self-regulated as an Event with Think-Aloud Protocols.

To deal with problems that arose from measuring self-regulated learning as an attribute, some researchers conceptualized self-regulated learning as a series of events that occur while students work on a task (Azevedo & Cromley, 2003; Azevedo, Cromley, & Seibert, 2004; Greene, Robertson, & Costa, 2011; Winne & Perry, 2000). Whereas several methods were created to do this, including eye tracking, behavioral traces, computer log files, and discussion turns (Azevedo, 2014), one method that has been used often with computer-based learning environments is think-aloud protocols (Azevedo, Moos, Johnson, & Chauncey, 2010; Greene, Costa, Robertson, Pan, & Deekens, 2010; Schraw, 2010; Zimmerman, 2008). A think-aloud protocol is a method of gathering data that consists of having a learner complete a task while verbalizing what they are thinking at the time (Greene, Robertson, & Costa, 2011). In a review of event-based measures of self-regulated learning, Schraw (2010) found that think-aloud protocols were the only method that measured effort, plans, strategy use, and monitoring. The sole factor think-aloud protocols were not able to measure was pre-study factors (Schraw, 2010). While various trace elements like hyperlink choice, eye tracking, and palette choices could be used to measure some of the same cognitive processes as think-aloud protocols, these alone cannot measure the same breath of cognitive and metacognitive processes that a think-aloud protocol can.

The think-aloud methodology has emerged from an information-processing framework (Ericsson & Simon, 1993, 1980). During a think-aloud protocol, researchers ask learners to verbalize their thoughts and actions as they work on a task. These verbalizations are the internal speech that occurs during problem solving (Vygotsky, 1986). It is important when using this method to only capture the verbalizations and not ask for explanations. Learners can describe their cognitive processes, as the process of describing them will not alter them. However, attempting to explain them will alter them and increase cognitive load. In a review of literature, Ericsson and Simon (1993) found no evidence that thinking aloud decreases performance, although it may increase the task duration. More recent studies have demonstrated that a learner’s strategy use and metacognition do not change in quantity during a think-aloud task (Bannert & Mengelkamp, 2008; Veenman et al., 2006)

Verbal analysis. Measuring self-regulated learning using a think-aloud protocol requires two things, a task and a measurement model (i.e., method protocol), for turning what the learners say into usable data. There are many different methods of analyzing verbal data, but the methodology used most commonly to measure self-regulated learning with a think-aloud protocol is based on Chi's (1997) methodology of verbal analysis. With traditional think-aloud protocol analysis, often there is an ideal model for solving a problem, and the goal is to test that model. For example, Ericsson and Simon (1993) were interested in learning how people played chess to create a logical model of chess play that would become the foundation for a chess AI system. With verbal analysis, Chi's focus was more exploratory, in that she wanted to see what learners' models for problem solving and conceptual change were. In protocol analysis, the focus is on the sequence of events, whereas in verbal analysis sequence does not matter as much, as all utterances, regardless of position, reflect the underlying process use that is of interest to the researchers. Using this shift in focus allows researchers to apply a mixture of qualitative methodology and quantitative statistical analysis to think-aloud protocols, whereas the traditional method presented by Ericsson and Simon focused on only measuring differences between a learner completing a task and an ideal solution. An example would be what chess piece does a player move, compared to the best logical move (Ericsson & Simon, 1993).

Azevedo and colleagues (Azevedo et al., 2002; Azevedo & Cromley, 2004b; Greene & Azevedo, 2009; Greene, Robertson, & Costa, 2011) have developed a method of using think-aloud data for analysis of self-regulated learning while performing a computer based learning enviroment task, which consists of four key parts: capturing think-aloud verbalizations during a learning task along with video data of their actions, transcribing the data from audio of a participant's speech into a text document, segmenting the transcripts into codable units, and applying meaningful labels to segments, if possible. A sample task that is used often in this research is for students to use a digital encyclopedia to learn about the human heart while thinking aloud (Azevedo & Cromley, 2004b; Greene & Azevedo, 2007b; Greene, Costa, et al., 2010).

Capturing think-aloud protocol data and transcription. Capturing think-aloud protocol is relatively straightforward in tasks using computer based learning enviroment. Learners engaging in a task are recorded, generally with both audio and video devices (Azevedo & Cromley, 2004b; Greene, Costa, et al., 2010; Hofer, 2004). Other measures of recording, such as eye tracking, screen-capture software, computer logs, or galvanic skin response sensors, can be employed concurrently (Azevedo, 2014). Following the task, the audio is transcribed to aid in the rest of the data preparation.

Segmentation. Segmentation is the process of breaking apart transcriptions into coding chunks. While it is possible to code an entire transcription as a whole, it is generally more useful in self-regulated learning research to break it into smaller chunks (Chi, 1997; Greene, Robertson, & Costa, 2011). These segments can be based on units of time, pauses in speaking, navigation points in the enviroment, or other various breaking points. The protocol that is most commonly used in self-regulated learning research is breaking the contents of the verbalization into codable units (Ericsson & Simon, 1993). Greene et al. (2011) defined the size of this unit as “segments that contain the fewest number of words while still being interpretable as an indicator of a cognitive process, even outside of context” (p. 328). An example segmented protocol can be found in Appendix A.

Coding the protocols. Once the data are segmented, they can be coded. Codes can be either emergent from the data or determined a priori (Chi, 1997). Researchers have used

Azevedo's codebook as starting point for the study of self-regulated learning within a computer-based learning environment using think-aloud methods (see Appendix B; Azevedo et al., 2002; Azevedo & Cromley, 2004a). This list contains 35 codes, each representing a cognitive, metacognitive, or behavior process, and has been employed directly or employed with additions in numerous studies (Azevedo, 2005; Azevedo & Cromley, 2004b; Greene et al., 2015, 2014; Greene & Azevedo, 2009; Greene, Costa, et al., 2010; Moos & Miller, 2015). An example of a transcript coding using this system is listed in Appendix A. This coding system was created with the intent to capture aspects of self-regulated learning using Pintrich's (2000), Winne and Hadwin's (1998), and Zimmerman's (2002) prior work in self-regulated learning. Codes were made to capture aspects of self-regulated learning within four broad areas: planning, monitoring, strategy use, and task difficulty and demands (Azevedo & Cromley, 2004b). The original coding system comprised four broad areas, with a fifth one regarding interest added later (Greene & Azevedo, 2009), the codes themselves are from all aspects of self-regulated learning theory, including knowledge activation, goal creation, metacognition, self-questioning, study skills, knowledge elaborations, coordinating information sources, controlling the environment, time management, and evaluation content. Whereas the original codes form the core of this coding system, codes are added, removed, or refined based on the task and on previous findings. The codebook has expanded to as many as 50 in later studies (Greene et al., 2015), and in some studies (Greene et al., 2014) a second set of codes to examine epistemic cognition was added as well, which brought the code count to over 80.

The code book has been refined by researchers as they complete new studies. The largest changes to coding came from the addition of valence to some of the metacognitive monitoring processes (Azevedo, 2009). Before the addition of valence, feelings of knowledge, judgments of learning, and content evaluations were coded the same whether they were positive or negative experiences. For instance, a segment where a learner stated they found their current learning content useful was coded the same as one where the learner did not find the content useful. Likewise, learners stating that they understood something they just read would have been coded the same as if the learner stated they did not understand what they just read. Positive and negative events, however, can often lead to different events following their occurrence. For instance, a learner who feels they did not understand what they just read may re-read the section, whereas one who feels they understood the content may self-test, summarize what they read, or move on to a new topic. Therefore, it seems important to add valence to certain micro-level codes in the coding scheme (Greene & Azevedo, 2009). With judgment of learning it was found that learners frequently changed their strategy use following negative judgments of learning (Binbasaran Tuysuzoglu & Greene, 2014).

Data products. The final product of this measurement protocol is a list of counts for what self-regulated learning events occurred while the learner completed a task. Researchers refer to these as micro-level codes. These micro-level codes have several properties. First as count variables they can never be negative. Second, count variables are usually distributed with Poisson or negative binominal distributions. In a Poisson distribution the mean of the distribution is equal to its variance. If the variance is higher than the mean, then the variable is considered over-dispersed and may be better modeled using a negative binominal distribution (Cohen, Cohen, West & Aiken, 2003). Second, due to a limited sampling window, learners who do not use a process can fall into two groups. The first are those who would never use a certain process at all. For instance, a student may never take notes when they complete a task. The second group are students who would use a strategy if the task was longer but instead either did not have time to use the strategy or used a different strategy instead. This creates a zero-inflated distribution, where the number of subjects not using a self-regulated learning process may be artificially inflated (Long, 2001). Examining data from a think-aloud protocol analysis using self-regulated learning data, Greene and colleagues (2011) found that the negative binominal distribution was the best distribution to model the data. They also found that no further benefit from gained in using a zero-inflated version of the negative binominal distribution. Similar analyses are necessary whenever researchers analyze TAP data. For an example of micro-level codes and the corresponding macro-level codes, see Appendix B.

Generally, also presented are the macro-level codes. Macro-level codes are linear combinations that represent a theoretical higher order that these codes come from. To create the macro-level codes, one would simply add up the micro-level codes within that macro-level code (Greene et al., 2013).

The macro-level variables are useful for several reasons. The micro-levels can show exactly what processes students used to complete tasks, but sometimes the exact process is not as important as the macro-level processes used (Greene & Azevedo, 2009). Individual differences in how learners complete a task depend on prior knowledge and internal and external conditions. Small changes in these conditions could lead to one student taking notes, whereas another student making a verbal summary. Both are examples of strategy use, and whereas the individual process may not be predictive of learning, the use of either of the processes may. Analyses at the macro-level can account for idiosyncratic differences in micro-level processing that can arise across individuals (Greene et al., 2013).

There are other benefits of using these macro-level codes. Models of self-regulated learning tend to deal in these theoretical higher order ideas and not more specific processes (i.e., planning, monitoring, and strategy use rather than reviewing subgoals, judgments of learning, and elaboration), so analyzing macro-level self-regulated learning processing can allow for better understanding of how the overall self-regulated learning model works in practice (Greene et al, 2013). Finally, these higher ordered variables reduce the amount of information presented, providing not only a more manageable set of processes, but one that results in a model that is better suited to quantitative analyses. Quantitative analysis requires, in general, more subjects than predictors, and in some studies, that is not possible, due to the high resource demands of gathering data from participants (i.e., capturing, transcribing, and coding think-aloud protocol data; Greene et al., 2014). There is also the benefit of working with more normalized data. Despite the individual micro-level processes often being best modeled with some kind of count distribution, the sum of these variables often produces a variable with a normal distribution (Greene & Azevedo, 2009).

Summary. When measuring self-regulated learning as an event, think-aloud protocols employing the Azevedo, Greene, Moos, and colleagues' methods and codes allow researchers to capture the cognitive, metacognitive, and behavioral aspects of self-regulated learning as a learner works through a task. Through the process of recording, transcribing, segmenting, and coding the data, the data transform from open-ended verbalizations from the learner that occur naturally as they engage with a task to counts of how often a learner has enacted particular processes. These processes can be modeled as count data and used for quantitative analysis and building statistical models, either modeled at the micro-level or specific skills, or the macro or broader theoretical skill level. Then these processes can be used to answer the question of what processes are needed by 21st century learners. One further step is required to answer this broad question however. That is what subset of these skills best predict learning. For this, statistical models are required.

Statistical Models for Self-Regulated Learning

When the coding is finished, what is left is a count of the self-regulated learning processes used by learners. Then the issue is how to determine which self-regulated learning processes are indicators of learning and which are not. With 35 self-regulated learning processes, the chances of finding a significant relationship between any one of these and the learning outcome becomes 83.39% due to a very high inflation of the Type I error rate that comes with doing multiple hypothesis tests (Bender & Lange, 2001). If one were to use the conservative Bonferroni adjustment to limit the probability of a false positive result, the corrected *p* value at which to test hypotheses would become $p < .001$ (Shaffer & Saffer, 1995).

Normally, when faced with a high rate of a Type I error, one can increase the power with a larger sample (Cohen, 1992). Using G*Power 3.1 (Faul, Erdfelder & Land, 2009) to determine the number of subjects needed to find a small and medium effect size with a multiple regression design, 35 predictors and a $p < .05$ significant level with a 95% statistical power to find a true effect, 1906 participants would be needed to find a small effect size (adjusted $R^2 = .1$), 277 participants for a medium (adjusted $R^2 = .3$) effect size and 135 for a large (adjusted $R^2 = .5$). This would mean, to find a model with an adjusted R^2 of over .3, a medium effect size, one should have 277 participants, or one risks getting back a false negative result where the independent variables were not found to have a significant relationship with the dependent variables, but a true relationship did exist.

Sample sizes of 1900 and even 300 are unheard of in self-regulated learning research using a think aloud protocol due to how time intense it is to gather data from participants. With an estimate of seven hours per participant needed to transform the raw data into data that are suitable for quantitative analysis (Greene et al, 2013), high sample studies are prohibitively expensive to run. Lack of statistical power coupled with a high false-positive rate when doing multiple comparisons has led to researchers finding other ways to make sense of the patterns of processes found using protocol analysis.

In the early studies, researchers used multiple chi squares to study learning outcomes. In Azevedo and Cromley (2004), Azevedo, Cromley, and Seibert (Azevedo & Cromley, 2004), and Greene and Azevedo (2005), these comparisons were done without correcting for multiple comparisons. Azevedo, Gurhrie, and Seibert (2004) grouped processes comparing individual processes within the higher order category they comprise, such as planning, using a two by four chi square design. Then to examine differences between two groups regarding specific methods of planning use, the researchers broke planning back down into the four parts that comprised it; creating multiple goal plans, creating sub-goals, activating prior knowledge, or recycling the goal of the task into working memory.

Much of the recent work into statistically analyzing think-aloud protocol data for self-regulated learning has employed ordinary least squares (OLS) regression models (Azevedo, Moos, et al., 2010; Greene et al., 2015; Greene & Azevedo, 2007b, 2009; Greene, Costa, et al., 2010; Moos & Miller, 2015). These studies involved a pre-test/post-test format to measure the relationship between self-regulatory learning processing and knowledge gain. Knowledge gain was defined as either the change in pre-test to post-test scores or the post-test score with the pre-test included as an independent variable. Due to a low ratio of subjects to variables, creating a model consisting of all micro-levels processes is not practical, either from an analysis standpoint or a utility one. When there are more predictors than cases, standard OLS regression no longer works as intended, as no singular solution will be found (Freedman, 2009).

Due to problems with the number of indicators increasing to sizes nearly equal to that or greater than the sample size of a study, variable reduction methods are needed. Two methods have been used: one which creates linear combinations of the micro-level processes into a larger grain size, and one that attempts to reduce the variables into a two subset of variables. One of these subsets is positively associated with learning gains and the other is negatively associated with learning gains. In the rest of this section these methods will be discussed, including an overview of other OLS regression methods to provide an overview on the current state of how this protocol has been and can be analyzed. This examination will conclude with an overview of an exhaustive search method that will find the micro-level variables that best predict learning using all-subset regression.

Ordinary least squares regression. Before moving futher into this discussion, there are some aspect of OLS regression that are worth noting. First, OLS is considered a BLUE solution. *BLUE* stands for *best linear unbiased estimator*. *Best* means that the equation for a regression model with the lowest distance between the regression line, or plane with three variables and hyperplane with four or more, and the data points will always be $Y = bX + e$. Y is the dependent variable of interest. X is a matrix of cases by variables. Modifying X is b, which is the regression coeffienct for X. A one-unit increase in X, with all other things held constant, results in a b increase in Y. The final term, e, is a is a vector of unknown error for each case. Using the Gauss-Markov theorem, it can be demonstrated that this produces the model with the least amount of unexplained variance (Freedman, 2009). For a complete discussion on the Guass-Markov theorum, consult Faraday (2014). The next term in *BLUE* is *linear*. OLS regression produces a linear solution, or one that finds a hyperplane that reduces the distance between the value for *y* at point *x* for all data points. The third term is *unbiased*. This means the estimates the model are the same as the parameters in the population The degree to which the population value and the sample estimates vary is the degree of bias that a system has. OLS methods of regression are extremely useful, as when the assumptions of regression are met, these estimates are considered to be equal to the population values. If bias exists in a regression model, then that model's results will differ from the true values of the population. The last term is *estimator*. This term means that, like most statistics, regression results are estimates of population values (Faraway, 2014).

While no two sources are in full agreement on what the assumptions of OLS regression are (Williams, Grajales & Kurkiewicz, 2013), there several issues that can come up in analyses that can cause problems with interreptations that are generally agreed upon and these include: proper *specification*, *normality of residuals*, *independence of residuals*, *homoscedasticity* of variance, lack of *perfect multicollinearity* and *outliers* and *influential cases* (Berry 1993; Cohen, Cohen, Cohen, & Aiken, 2003, Gelman & Hill 2007; Faraday 2014; Field 2013; Thompson 2005) Specification means that the model has the proper predictor variables included and the relationship is being the predictor variables and the dependent variable is modeled properly. For a linear model, one assumes the data actually are best described using linear methods (Thompson, 2005). The second issue is that the residuals, or distance between the regression line and each data point, are normally distributed (Thompson, 2005). Next, is that the residuals are independent. In some situations the residuals can be become correlated, properly known as autocorrelation. This is an issue primarily in time series designs. The next is homoscedasticity of variance, which is that the residual scores have equal variance for all values of the predictor variables. In other words, the residuals are roughly equally distributed along the regression line or plane / hyperplane. When this is not true, values may be accurate for some values of the predictor but not for others (Field, 2013). This usually presents in the ends of a distribution, so scores at the low end or high end of a distribution will show more variance than variables near the mean, leading to problems when one wants to make predictions from the dataset, or estimate values for data points beyond what was observed in the study.

Lack of perfect multicolinearity means that no two predictor variables are identical, and no variables are linear composites of other variables (Faraday, 2014). If this is true, then OLS regression cannot be performed resulting in most computer programs returning an error, reporting a non-positive definate matrix being produced (Crawley, 2007). A final issue in regression models is that there are no outliers or influencal cases (Navarro, 2013). In cases with outliers or influencal cases the regression model is being heavily influenced by a small number that if removed would greatly change the model. When present, these issues create problems ranging from not being able to complete a regression model at all, to distorted parameter estimates, extremely large standard errors and confidence intervals, improper p-values, a model that will fail to replicate, or even different parameter estimates when applied to different subsets of the dataset. Simply put, models with these issues run the risk of being unreliable and invalid.

Theory-driven model (full aggregation). Noting some of these assumptions may not be met in their studies, Greene and Azevedo (2009) reconceptualized the way they analyzed the results from protocols. In Greene and Azevedo (2007b), chi squares were used to determine which micro-level processes were associated with learning gains, by comparing counts of self-regulated learning processes used by students with different mental models, defined as low, medium or high, following a conceptual learning task. They found that control of context, coordinating informaton sources, expecting the adequacy of information, feeling of knowing, inferences, and knowledge elaboration were associated with developing more complex mental models. However, these results may not have been the full story behind the data. Greene and Azevedo (2009) suggested that perhaps by using only micro-levels to analyze the self-regulated learning data, they might have missed other findings. What if self-regulated learning at the micro-level was personalized? In this instance, two learners may have the same outcome of learning while using two different strategies. One could draw a picture, while the other took notes about a repesentation they both looked at. These may be equalevent strategies that are representative of a larger macro-level self-regulated learning process of strategy use. Given that the Azevedo, Greene, Moos, and colleagues method of analysis had these micro-level strategies already nested into higher orders (Greene et al., 2013), it was easy to reconceptualize the codes into macro-level processes. They would just be the sum of counts for the related micro-level codes.

In the past, these researchers did not find micro-level monitoring processes, such as feelings of knowledge or judgments of learnings, to be predictive of learning, but when the monitoring codes were aggregated into a the macro-level, they did find that as a whole they became predictive of learning (Greene and Azevedo, 2009). This may mean that, unlike strategy use, which tends to show specific strategies are important to learning and others less so, monitoring may be best taught on a higher level (Greene et al., 2013). In a computer based learning enviroment to study history, with no embedded scaffolding, planning became predictive of learning gains. However the use of higher-levels strategies that were found to be predictive in the past (Greene & Azevedo, 2007a; Greene, Costa, et al., 2010) were not predictive of learning in this study. This may have been due to students lacking prior knowledge in the history learning task. By using the macro-levels of planning, strategy use and monitoring, the Full Aggregation Model provided a glimpse at the big picture of self-regulated learning usage, while still allowing for the micro-levels to be recorded and used to examine finer grain issues that can appear (Greene et al., 2013).

Data-driven model (data-based aggregation). Having different results at the macro-level and micro-level for self-regulated processes associated with learning created a new problem. If some of the micro-level codes within a macro-level process were predictive of learning, whereas others were not, using the Full Aggregation Model would result in the macro-level variables being comprised of micro-level variables of varying type, with some that may be associated with negative learning gains and others with positive learning gains. When added together, this could lead to unclear results. Reviewing the early analyses of micro-level codes that were predictive of learning, it was found that only a small number of the total codes were predictive of learning gains, even without controlling for experiment-wise Type I error rates (Greene et al., 2015). Examining the correlation matrix of the association between micro-level self-regulated learning variables and learning gains showed that some variables were positively correlated with learning gains whereas others were negatively correlated with learning gains. Thus, the data-driven aggregation method was created. This method was designed to find the subsets that were positively associated with learning gains and combine them and also find the processes that were associated with negative learning gains and combine them. Aggregrating these two sets of micro-levels together into separate variables based on their positive or negative correlations became the basis for the current data-driven aggregation model (Greene et al., 2015, 2014). These regression model demostrated higher adjusted R^2 values than the full aggregation and thus fit the data better, as adjusted R^2 (henceforth adj R^2) is a common method of assessing fit in the social sciences in that it tells you the amount of variance the model can explain (Cohen et al., 2003). While R^2 is commonly used for this purpose as well, R^2 will increase as predictor variables are added to a model, whereas adj R^2 controls for the number of variables added by penalizing models for each additional variable they add.

There is a problem with this method, however. This method produces a model based on the data, but it is still does not answer the question of what specific skills produce the best results in a task. If the weakest correlated variables in these models were removed, would the adj R^2 drop or increase? What about if other variables were added? By only looking at two models out of millions, there is a good chance that the best model, or the model with the highest adj R^2 value, is not represented here. This limitation leaves only one option to find the the regression that best fits the data, examining all possible models.

Step and stage regression methods. The main way variable selection in ordinary least squares regression is performed is using stagewise or stepwise regression. These methods start with either all the variables entered in a model (i.e., backwards selection) or a model only containing the intercept term (i.e., forward selection). Then in the case of forward selection, variables are added until adding no other variables will increase fit beyond a threshold, whereas in the case of backwards selection they are removed until no change will reach the threshold. A third method, iterative selection, starts with either forwards or backwards methods then adds or removes variables by evaluating both additions and removal of variables from the equation (Miller, 2002).

In stepwise regression the regression equation is evaluated to determine which variable added to or removed from the regression equation would have the largest impact on the fit measure. If this impact is greater than a specified amount, for example adj $R^2 = .01$, then the variable is added or removed, and the equation updated. This repeats until no addition or deletion can be made to produce the requested change (Miller, 2002). Stagewise regressions uses the process, but variables can be added in groups. This process is sometimes done when variables are thought to have some degree of collinearity, or shared correlations with one or another variable (Hastie et al, 2009).

The starting state of the model (i.e., all variables or no variables entered) and the rules for adding and subtracting make up the algorithm for model creation (Hattie, 2009). The most common implementation of these methods is starting with a model with all variables entered it (i.e., a one-step forward stagewise insertion) that removes variables with statistically non-significant t-tests (i.e., backwards stagewise and stepwise deletion). This is repeated until all variables significantly contribute to the model. Then the first and last models are reported as full and reduced models (Field, 2013).

Thompson (1995) noted three problems of stepwise regression. First, they do not report the true degrees of freedom. What is reported is the degrees of freedom for the model that the procedure stopped with and this model is treated as if it was the sole analysis preformed in regard to calculating degrees of freedom, and thus, associated p-values. Second, stepwise regression models may not even find the best model for a model with a given number of variables. For example, if the best relationship between variables A, B, C and D, can be represented in the relationship, $y = ax + bx + cx$, it is possible for stepwise regression to pick $y = ax + cx + dx$ then stop trying to find better solutions, incorrectly presenting this as the best one three variable solution. Finally, stepwise regression may capitalize on meaningless differences between scores. As variables are added and dropped from the model, small changes guide the selection, and the differences between why one model was rejected by the algorithm and why another model was selection may become trivial.

These methods are heavily critiqued and nearly every statistic book that mentions them does so with a word of warnings that they should not be used uncritically (Cohen, Cohen, Cohen, & Aiken, 2003; Faraday 2014; Field 2013; Freedman, 2009; Hastie et al., 2009; Miller 2002; Thompson, 2005). Cliff (1987) offered perhaps the most colorful critique of these methods stating, "most computer programs for multiple regression are positively satanic in their temptation toward Type I errors in this context (p. 185)." It may seem strange to see so much hatred for a method, but the history of stepwise regression was not very pleasant. In the late 1960's two major studies were conducted in education. The first was Equality of Educational Opportunity report (also known as the Coleman report; (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966). This report followed the passage Civil Rights Act of 1964. Coleman and his team were tasked with determining the extent in which racial belonging effected achievement and opportunity in education. While the Coleman report is historic for many reasons, for this thesis, it's most infamous conclusion will be the sole focus. The Coleman report found:

> "Taking all these results together, one implication stands out above all: That schools bring little influence to bear on a child's achievement that is independent of his background and general social context; and that this very lack of an independent effect means that the inequalities imposed on children by their home, neighborhood, and peer environment are carried along to become the inequalities with which they confront adult life at the end of school (Coleman et. al, 1966, p. 335)."

The fallout was swift. Educational budgets saw cuts and educational spending slowed and spending froze until the early-1980's when the Nation at Risk (Gardner, 1983) report stimulated it again. Problems were quick to be found with the report. A full critique is beyond this article, but it was found that Coleman used a forward stepwise regression model to generate his conclusions about achievement. In his model, he entered in family background, then allowed other variables to be inserted. Bowles and Levin (1968) quickly pointed out that in the event of collinearity, without further analysis, if family characteristics were entered in the model first, and it was highly correlated with spending, then spending would have very little unique predictive value after the common variance was accounted for. Since, family characteristics were forced to be the first variable into this regression equation, this finding was not unexpected and was the more of the result of the regression methodology than any true effect of educational spending's relationship with achievement. If educational spending was entered first, the results would have become that family characteristics had little predictive value.

This study, and others such as the early study that found Head Start was not effective (see Westinghouse Learning Corporation and Ohio University 1969; Smith & Bissel 1970) and other studies that had their results called into question through secondary analyses, promoted a change in the standards for the way people used statistics (Glass, 1976). The strongest of these changes was creating an environment where high impact policy based on conclusions reached from using stepwise regression that was later found to be misleading. Researchers following these high-profile cases would start to use cautious when they see stepwise methods implemented and refrain from them themselves (Thompson, 1995).

All-subset regression. Whereas in the past they were routinely dismissed as being too computationally intensive to be of viable use (Hastie et al., 2009), calculating all possible regression subsets for up to 100 variables is now part of common statistical software packages, with no limits at all on the number of variables (Gunes, 2015; Lumley, 2015; Miller, 2002; Raschka, 2015; Yang, 2013). This method determines all possible combinations of variables in a regression model then runs every combination as a seperate regression analysis. In the case where there are more predictors than participants, the maximize size of subsets is reduced to examining only the n-1 predictors. For instance, if one has 35 variables and 12 subjects, only models with a maximum of 12-1, or 11, variables can be predicted. Models that would contain 12 to 35 variables cannot be estimated using OLS regression. This is a hard limitation of ordinary least squares regression due to the properties of matrix inversion (Fieller, 2015). This limitation results in only a subset of all models being able to be calculated.

Two pieces of information are required for understanding all-subset regression. The first is the variables that are in the model themselves. The second piece of information that matters is a measure of how well the variables selected fit the data. In this study the value of model fit will be the adj R^2 value. Adj R^2 is the unbiased coefficent of determination and represents the variance that the model can explain. R^2 is ratio of explained variance to total variance. Simply calculated it is the sum squares of the regressor divided by the total sum of squares, or $R^2 = SS_{Reg}/SS_{Totol}$ (Freedman, 2009). It can be surmised however that as one adds more information to a regression equation, more the unexplained variance will become explained thus biasing R^2 towards larger models as each variable added will decrease the total sum of squares of the error. Therefore, adj R^2 is often used to compare regression models of differing sizes, with the following correction being applied to remove the bias. Adj $R^2 = R^2 – (1- R^2)(p/n-p-1)$, where p is the number of predictor variables and n is the sample size (Freedman, 2009). This change, while minor, adds a threshold for the amount of variance that needs to become explained before a variable adds a positive impact to the adj R^2 value, penalizing models with large numbers of variables that may only add trival increases to the explained variance. Employing this method to the self-regulated learning micro-level processes makes it possible to find the combination of variables that produces the best linear unbiased estimator and, thus, the model with the highest adj R^2 value within the OLS regression framework.

Once all possible regression equations have an associated adj R^2 value, they are ranked and the top 5 to 10 are retained (Yang, 2013). The total number of models that are needed to be run are equal to 2^p, where p is the number of predictor values (Freedman, 2009). Each of these models is tested using regression, retaining information that is needed to report the results. Despite the goal being to find the best single model in this study, many times the models may not be all that much different in terms of predictors or adj R^2, so having more than just the top single best model can show how the top models differ and if they are more unique than alike (Breiman, 2001).

All-subsets regression is a common method of statistical analysis in some fields, this type of data mining can be considered a questionable research practice in education and the social sciences (Ioannidis, 2005). This is a notable concern. However, regression can be used for description as well as inference (Berk, 2004). When the goal of a study is to find what processes are most associated with a learning outcome based on the data presented, then one is looking for a descriptive model and not an inferential one. Given the contextual nature of self-regulated learning, generalizing the results of a single study with a specific task to other studies may be problematic, and even replications may yield different results (Greene et al., 2015). Knowing this, data-driven methods such as all subset regression present an opportunity to embrace an exploratory data analysis methodology.

Within this context, regression is used to describe a relationship between the variables and is guided by data (Berk, 2004; Ratner, 2010). Inference and generalization beyond the data should be avoided, as using a method like all-subset regression means that many of the assumptions of OLS regression may be violated (Berk, 2004). Also when performing millions of tests, test statistics such as *F*, *t* and *p* lose their meanings. All subset regression results in models with the most extreme scores, which means the procedure results in ones with the largest *F* values and largest *t* values within the model. Attempting to correct for the Type I error rate with a millions of models would leave alpha levels (i.e., *p* values) approaching zero. This is the cost of using a method like all-subset regression to use the data itself to determine what model fits it best. However, this is percisely the method that machine learning and statistical learning employ to routinely solve complex problems (Breiman, 2001b).

Comparison of statistical models. Each of the three proposed methods has advantages and drawbacks. The Full Aggregation Model has the advantage of being the simplest to understand as it reports the macro-level processes of planning, strategy use, and monitoring. The drawback is this model cannot answer the question of which skills best contribute to learning outside of the broadest sense of planning, strategy use and monitoring skills. Because the specific skills, or micro-level processes, are summed, no estimates of the effectiveness of the skills can be made at the micro-level. The beta weights or correlation coefficients of the individual micro-level processes cannot be calculated with this method. This also proves problematic when some of the skills are predictive of learning (i.e., positively correlated with learning) and some are not predictive of learning (i.e., low correlations with learning), or associated with negative learning outcomes (i.e., negatively correlated with learning). In these cases, the overall predictive value for the macro-level processes will be reduced by inclusion of these skills (Greene and Azevedo, 2009b).

To deal with these limitations, the data driven aggregation method was created. This classifies self-regulated learning processes into two broad areas, those predictive of positive learning outcomes (i.e., positively correlated with learning) and those predictive of negative learning outcomes (i.e., negatively correlated with learning). This also allows for students to vary the skills they use, assuming they are in the same learning outcome measure of positive learning outcomes or negative learning outcomes. Like the full aggregation method however, this method does not allow for direct assessment of the skills predictive of learning or negatively associated with learning (Greene et al., 2013). Skills are all considered to be equally predictive of learning or not associated with learning and not included in the model at all. The beta weights or correlation coefficients of the individual micro-level processes cannot be calculated here either. This model can answer the question of what processes are associated with learning, but assumes all processes are equally associated with learning because it is the sums of the counts for micro-level processes. When examining correlations from the studies using this method however, some processes have higher correlations with learning outcomes, which would indicate that they are more effective at promoting learning however.

Finally, the models produced by this method only will be the one with the highest adj R^2, or the one that fits the data best, if the variables in the positive learning outcome and negative learning outcome are truly independent of each other. Regression examines the relationship between a variable and learning outcome when all other variables are held constant. In the event of covariation or multicollinearity, a variable has a positive correlation with an outcome variable when examined alone may have a reduced, or even negative correlation (i.e., in the case of suppression) with the outcome variable as the variance that the variable is explaining may also be explained by existing variables. This may result in this method while picking the micro-level processes with the highest positive and negative correlations with learning but producing a model that is not the best one that could be made from all possible combinations regressions between the micro-level processes and the learning outcome. Finally, to answer the question which skills best contribute to learning, one would assume to be looking for the skills that uniquely contribute to learning. The data-driven aggregation method cannot answer this, and only will give a list of skills that contribute to learning when used together, allowing for skills that may not be contributing to learning to still be included in the model. Whether a micro-level process is statistically contributing to the model after all others are held constant cannot be evaluated with t-tests in this case as it is done when employing multiple regression without linear composites (i.e., a variable made from adding other variables together; Field, 2013).

All subset regression is the method that should be employed to answer the question of which skills best contribute to learning, with estimates the unique contribution of the skills to the learning outcome. Instead of reducing the question to a broader level or producing a set of skills that examines the best set of skills when evaluated independently, this method will produce the model that gives the set of skills, in this study defined as micro-level processes, and that also best describe the learning outcome, as defined by the adj R^2 level or measure of how well the model fits the data. While the two prior methods attempt to estimate the best model, all subset regression will directly determine which micro-level processes when examined together produces the best fitting model. This method will do so while retaining the impact of the individual processes as each process will have its own regression coefficient, or impact on the learning outcome when all others are held constant, something that the two prior methods cannot produce. As a result of having the regression coefficients, this method will allow for more detailed comparisons of studies using this methodology. Currently, only the beta weights or regression coefficients for macro-levels in the full aggregation method can be compared across studies. The individual beta weights or correlation coefficients micro-level processes cannot be compared across studies and these variables must solely be evaluated in terms of whether they are present or excluded in the data-driven aggregation's positive or negative learning outcomes groups. All subset regression will produce equations that have beta weights and regression coefficients associated with the micro-level processes, so the unique contribution of a micro-level process can be evaluated. This will answer not only what skills are best associated with the learning outcomes, but also give an estimate of how much they relate to the learning outcome. It will also evaluate the micro-level processes when all else is held equal, whereas the micro-level processes when using the full aggregation and data driven aggregation can only be evaluated as a group and not examined for their unique effects.

This method also can be extended to situations where the number of predictors is larger than the number of cases. By examining only models up to a maximize size, this method gives the benefits of the full aggregation method and data driven aggregation methods of being able to work with data sets with more variables than cases, without the loss of information that using composites requires. Instead of losing information in terms of what can be evaluated in a statistical model, one would just have to limit the maximum size of the statistical model to be evaluated. This tradeoff may not be a practical limitation if the size of the statistical models this method selects are small, however. If the best models this method selects when used with SRL process data contain only models of 8 to 12 variables, then a limit of only comparing models of 15 or less variables may have no real effect on the analysis in most cases (Miller, 2002).

These advantages should make the all subset method of model selection the best one to compare models as it provides a way to describe the results without losing information about the micro-level processes. However, in practice all subset regression does have some problems that may not make this the best overall method. With other data, it has been found that all subset regression can overfit the model to the data, which causes issues when replicating the results. This is particularly problematic when the research questions move beyond summarizing the results of one to study, to a predictive format where one seeks to create a model that will best predict the results of a future dataset (Hastie et al., 2009). For this thesis, this is not a problem as the goal is describe to the data that is already gathered, but eventually, research on self-regulated learning should move away from description and inference to that of prediction, and for a model to have predictive validity, it needs to be able to reliably predict future performance (Beck, 2009). All subset regression may be found to lack predictive validity when compared to non-OLS regression methods such as regularized regression or regression forests due to issues it has with overfitting (Hastie et al., 2009). Also, with any variable subset selection method, the results of one study should not be evaluated in isolation, and future research will be needed to determine if this method is reliable. Replication is needed here, as it is with any other study, to determine if the model is one that best fits the true relationship between the micro-level variables and the learning outcome or simply is one that only fits the data it is given. This cannot reliably be estimated within the context of one study (Thompson, 2005).

Summary

Dent and Hoyle (2015) stated there was a need for alignment between the theoritical model, measurement methods, and statistical analysis method within self-regulated learning research. In an attempt to demostrate how this alignment would occur in this thesis, three models were presented. The theoretical model is the Winne and Hadwin (1998) model of self-regulated learning that conceptualizes self-regulated learning occuring as a series of processes within the context of information processing theory. The measurement model is the think-aloud protocol methodology of Azevedo, Greene, Moos and colleagues (Greene et al., 2011). This model measures self-regulated learning processes that occur in the Winne and Hadwin model by having students verbalize their cognitive and metacognitive processes and experiences. Finally, the statistical model uses all-subset regression to find which of these processes best predict learning within a certain computer based learning enviroment. By using these three models the broad question of what skills does the modern learner need and the more precise question that is presented in this study become aligned.

Research Question:

This thesis will examine the following research question:

1. Which variable selection method (full aggregation, data driven aggregation or best all subset regression) best fits the relationship between self-regulated learning and knowledge gain in the examined dataset defined as the one that maximizes adjusted R^2.

CHAPTER 3: METHODS

Participants

For this thesis, I conducted a secondary analysis of data from the study published by Greene et al. (2010). This study was conducted during the 2007-2008 school year at a large public university in the Southeastern United States. One hundred and seventy undergraduate participants were recruited from education classes, receiving extra credit for their participation. The gender breakdown of the students was 103 females and 67 males, and the mean age was 20 years with a standard deviation of 2.14 years. After reviewing the procedures for the study, one participant decided not to participate further and withdrew from the study without penalty. Of the remaining participants, video data were lost for 10 of them and therefore the researchers did not have their audio transcribed. These students will be excluded from this thesis. Furthermore, one more participant was excluded from the study because their handwriting could not be read, making scoring their pre-test and post-test impossible. After these participants were removed from the study, the final participant count was 153.

Materials

Participants completed an informed consent form, a pre-test, a post-test, and a demographic form. The demographic form included questions related to grade point average (GPA), major of study, age, and gender as well as information about their coursework and work experience. The pre-test and post-test were identical and were used to assess declarative and contextual knowledge of the human circulatory system. The tests had been used previously in published studies using this methodology (Azevedo, 2005; Azevedo & Cromley, 2004a). The test was composed of two sections: the first was a matching and labeling section that was used to assess declarative knowledge, and the second was an essay part used to assess conceptual knowledge. The matching section contained 13 item pairs and the labeling section contained 14 items that were related to parts of the human heart. The essay prompt was, "Please write down everything you can about the circulatory system. Be sure to include all the parts and their purpose, explain how they work both individually and together, and explain how they contribute to the healthy functioning of the body" (Greene, Costa, et al., 2010, p. 1033). The pre-test internal reliability was 0.79 and post-test internal reliability was 0.81 using SEM methodology (Greene, Costa, et al., 2010). The test packet is included as Appendix C; however, I will only be using the conceptual knowledge measure for this thesis.

Computer-Based Learning Environment

The computer-based learning environment used for this study was a commercially available edition of Microsoft Encarta (Microsoft Corporation, 2007). Researchers showed participants three articles that they deemed the most useful to learning about the circulatory system. These articles, titled "The Heart," "Blood," and "Circulatory System," consisted of 18 sections, 256 hyperlinks, 40 illustrations, and 1 video. The text from these articles totaled 41,380 words. Each of the primary articles had a hyperlink outline allowing learners to navigate to topics within the article. Embedded in the articles were hyperlinks to videos and photographs. Participants were able to navigate to any part of the Microsoft Encarta environment, but were asked not to leave it or use the built-in dictionary function.

The Learning Task

Participants were asked to learn as much as they could about the circulatory system. They were given a printed copy of instructions that stated:

> You are being presented with a hypermedia encyclopedia, which contains textual information, static diagrams, and a digitized video clip of the circulatory system. We are trying to learn more about how participants use hypermedia environments to learn about the circulatory system. Your task is to learn all you can about the circulatory system in 30 minutes. Make sure you learn about the different parts and their purpose, how they work both individually and together, and how they support the human body. We ask you to "think aloud" continuously while you use the hypermedia environment to learn about the circulatory system. I'll be here in case anything goes wrong with the computer or equipment. Please remember that it is very important to say everything that you are thinking while you are working on this task. (Greene, Costa, et al., 2010).

Learning Task Procedure

The procedure for this study was similar to one used in previous studies by Azevedo and colleagues (Azevedo et al., 2002; Azevedo & Cromley, 2004a). Sessions were conducted in a one-to-one setting with the participant and researcher meeting at a prearranged time. Due to the highly involved nature of the procedure for the first part of this experiment, researchers used a script to standardize the participants' experience as much as possible. After participants were greeted, they were told the study could take up to 90 minutes and that they were free to leave without penalty at any time. If they agreed to continue with the study, they were given an informed consent form to sign, after which they were given a demographic survey, and told they had as much time as they needed to complete it. Next, they were given instructions on how to complete the pre-test. Participants were asked to read the essay prompt aloud and told that they would have 20 minutes to complete the pre-test without the use of materials. If they finished early, they were to tell the researcher. They were asked to complete one page at a time without flipping back and forth. The researcher remained in the room while participants completed the pre-test and answered any questions not related to the contents of the test.

Next, the participants were introduced to the Microsoft Encarta environment. They went to the three main articles the researcher picked out as the most useful to start with while learning how to navigate within the environment. The controls the participants were trained to use were the forward and back buttons, hyperlinks to navigate within or between articles, controls to access and control a video on the heart, and the built-in Encarta search feature. A script was used by the researchers to standardize how this material was presented.

Once a participant was comfortable using the learning environment, they were instructed how to think aloud; specifically, they were asked to verbalize everything they were thinking. In addition, they were asked to read aloud, state any action they were taking (e.g., clicking on a hyperlink), and state when they were taking notes. To help them understand this process, participants were told to think aloud while reading an Encarta article on Michael Jordan for a minute or two until they were comfortable with the process and the researcher could make sure they were correctly thinking aloud. After this, participants were asked if they had any questions.

Once the participant was ready to move on, they were given the learning task in the previous section. The researcher read the task aloud and posted a written copy of it in view, so the participants could refer to it throughout the experiment. The participants were given 30 minutes to navigate the Encarta environment to complete as much of the task as possible, all while verbalizing their thoughts and actions. Participants were instructed that they could take notes if they wished but they could not be used in the post-test. Researchers stayed in the room during this time to help with procedural questions, deal with any technological issues that arose, and provide time prompts at 10 minutes, 20 minutes, and 28 minutes into the study. A tape recorder captured the audio and a video camera was used to capture what actions the participants took in the learning environment as well as other actions such as taking notes. During the task, if a participant fell silent, the researcher would prompt them by saying, "Please say out loud what you are thinking." After 30 minutes, the audio and video recorders were turned off, the notes were removed and placed in the participant's file, and the Encarta environment was closed.

Then participants were given a post-test, which was identical to the pre-test. They had up to 20 minutes to complete it with the same instructions as during the pre-test. After the post-test was finished, the time it took to complete was recorded and participants were de-briefed and asked not to share details of the experiment with any of their classmates who were likely to be participants.

Scoring Knowledge Measures

Once the participant was finished and left, the quizzes were scored by a team of two graduate students and the principal investigator. One point was added for each correct answer in the matching and labeling section. Zero points were given for a wrong answer. The essays were scored using a rubric (see Appendix D), which has been used in several other studies (Azevedo & Cromley, 2004a; Azevedo, Cromley, Winters, & Moos, 2004; Azevedo, Cromley, Winters, Moos, & Greene, 2005; Azevedo, Johnson, Chauncey, & Burkett, 2010; Greene & Azevedo, 2007b; Greene, Moos, Azevedo, & Winters, 2008). The interrater agreement was .994 (334/336 essays). The principal investigator resolved the two disputes that did occur.

Coding Micro-Level Self-Regulated Learning Processes

The audio tapes were given to researchers to transcribe. When the transcriptions were completed, the data was segmented into sequences of words that could be reasonably interpreted as evidence of a process being studied (Greene & Azevedo, 2009). These segments were coded using a codebook that was created prior to the start of the experiment and not altered during the experiment or coding. This codebook is presented in summary form in Appendix B. The codebook is very similar to the one employed in previous studies (Azevedo & Cromley, 2004a; Azevedo, Cromley, Winters, et al., 2004; Azevedo et al., 2005; Azevedo, Johnson, et al., 2010; Greene & Azevedo, 2007b; Greene et al., 2008). The coding process was first done by one of the researchers trained in this coding scheme, then coded again by a second researcher. Ambiguous statements that were not defined by a code in the codebook were given a label of no-code, as were sections that were being read aloud. After both sets of coding were completed, the two researchers who coded the text compared their results to resolve any differences they had in their coding. There were no statistical calculations of interrater agreement reported for this study, as all disputes were resolved through review of the codebook and discussion. In this study, 17,111 segments were coded, and the principal investigator was consulted less than 10 times to settle disputes. The interrater agreement for similar studies was .90 (Azevedo & Cromley, 2004b) and .95 (Azevedo et al., 2008).

Data Preparation

In this thesis, the data selected for analysis were the micro-level counts for each self-regulated learning process and the scores of the pre-test and post-test knowledge. The variable of learning gain was created by subtracting the post-test essay score from the pre-test essay score. The micro-level counts and this learning gain estimate were retained and all other variables in the dataset were dropped. Due to a fundamental difference in the way selecting new information source (SNIS) and control video (CV) were coded from the other micro-level variables, by counting clicks of a mouse and not cognitive processes, they were not used in this study. Whereas Greene, Costa and collegues (2010) dropped several micro-level processes from their analysis (i.e., Recycle Goal in Working Memory, Time Monitoring, and Task Difficulty) this study retained them for analysis.

The goal of this analysis was to see what subset of variables was most predictive of learning. To do this, the self-regulated learning variables were modeled as continuous variables and not made into composites. While they could be modeled as composites (Greene & Azevedo, 2009), since the goal of this thesis was to attempt to find a set of micro level variables that predicts learning, creating composites before the analysis would have been counterproductive. The micro-level processes could also be modeled as dichotomous (Greene, et al., 2011; MacCallum, Zhang, Preacher, & Rucker, 2002). Dichotomous variables are transformations of quantitative variables into two separate groups above and below a cutoff point. Because the full aggregation method and data driven aggregation method use counts of data, dichotomization was not appropriate in this case.

Data Analysis

Data analysis and visualization was completed primarily using IBM's SPSS Statistics software, version 25 (IBM Corp, 2017) with additional analyses done in R 3.5.1, Feather Spray (R Core Team, 2018). Descriptive statistics were calculated and presented in tabular form for knowledge measures, knowledge gain, and self-regulated learning skill processes to obtain the mean, standard deviation, levels of skew, and kurtosis along with quantile scores. Bivariate correlations were calculated for all variables. Visualizations of these statistics were conducted in the form of histograms for the knowledge measures, boxplots for the self-regulated learning process variables, and scatterplots for the bivariate associations. The variables were examined for outliers at this stage. In addition, some variables appeared to be problematic in their distributions.

Then the following models were constructed: The Full Aggregation Model, the Data-Driven Aggregation Model (i.e., Data-Driven Model) and 10 Best All Subset Regression Model, the best of which is further referred to as the Best Subset Model. The Full Aggregation Model was created by first summing the micro-level self-regulated learning variables for all macro-level processes. In simple terms, all the planning micro-level variables were added together to create the planning macro-level variable. This was repeated for strategy use and monitoring as well. These three variables were used as independent variables in a multiple regression with learning gains as a predictor variable.

The Data-Driven Models were created by first defining a positive outcome macro-level variable and a negative outcome macro-level variable. These macro-level variables were comprised of all self-regulated learning micro-level processes that correlated positively or negative with learning gains at three threshold points, .1, .05, .01 (Greene et al., 2014, 2015). Variables that have correlations greater than the threshold were added to the model for that threshold. For example, all self-regulated learning processes with correlations higher than .01 were added to the positive correlation macro-level for the .01 level model, which them summed the counts for each variable and obtained a total count score. Then all correlations that were less than .01 were added to the negative correlation macro-level for the .01 model, and these were also summed. This was repeated for the .05 and .1 levels. Then for each of the three models, these two macro variables of the total count for SRL above or below the correlational strength threshold were used as independent variables, and learning gains as a dependent variable.

The Best Subset Model was created with SPSS's LINEAR function (Yang, 2013). For this, the Best Subset Model was the one with the greatest adj R^2 to allow for comparisons with this model and others that use this general think-aloud protocol methodology. The LINEAR function returned the 10 best fitting models, as defined by those with the 10 highest adj R^2 values. The variables in these models were then used to re-create these as separate multiple regression models to obtain the proper model information that is not produced by the LINEAR function, such as the *F* statistic and related *p*-value for the model, *p*-value of the model, *df* of the model, and beta for the individual predictor variables within the model, along with *t*-scores and associated *p*-values. It must be noted that these are the 10 best results, and approximately 263 million other results will not be reported; therefore, *p*-values associated with the *f* statistics to assess the overall fit of the regression should be interrupted with extreme caution. Typically, when statistics are reported in the context of a regression, such as *p* or *f*, they are reported with the assumption that they were the only analyses done. In the case of this regression method, that is not true, so the meanings associated with these values are no longer accurate (Freedman, 2009). However, convention still requires they be reported (Kelley & Maxwell, 2010). Since all 10 models shared a core set of variables, a Core Model was also created to see how this core compared with the other models generated. Finally, the overall fit (as defined by adj R^2) of the best of the 10 best All Subset Models was compared to the Full and Data-Driven models. The items of interest to the study at this point were the variables each model selected as well as the associated adj R^2 value and patterns that appeared.

Diagnostic tests were then performed on the Full, Data-Driven, and Best Subset Models selected by the LINEAR function and best subset core model. The Breusch-Pagan test for non-constant variance (1979) test was used to test for heteroskedasticity in each model. Variable inflation factors were calculated to examine the degree of multicollinearity. Values higher than 4 indicate increasing amounts of multicollinearity (Field, 2017). Outliers were detected using z-scores with a criterion of three standard deviations being used to determine outliers. Finally, influential cases were examined using Cook’s distance, a score of 1 or greater being considered influential. Residual and diagnostics plots were also conducted to search for problems that are more easily noticed visually (Field, 2013).

Second-Pass Analyses

To examine the impact that influential cases, outliers, and variables with little variance may have on the results of the main analyses, a second set of analyses will be performed. Data was examined first for variables that lacked variance, formally called near-zero variance (Kuhn, 2013). In these cases, most participants did not enact that self-regulated learning processes. This would make the variable non-linear in nature and attempts to model it as linear would produce bias results (Field, 2017). Variables that were found to match this criterion were removed. Second, bivariate scatterplots of self-regulated learning processes and learning gain were examined for cases that seem to be outliers. These cases were removed. Finally, residuals and impacts of influential measures of the models constructed in the first pass were examined to see if any of these variables were in fact problematic. Variables with high residual values $z > 3$ or $z < -3$ or high influence values (Cooks > 1) were removed from the study. After this, the study was repeated, and a reduced version of the results presented to estimate the impact of non-zero variance variables, outliers, and influential cases.

CHAPTER 4: RESULTS

In this section, the models for full aggregation, data-driven aggregation, and best all subset regression will be presented and assessed. First descriptive statistics for the dataset will be presented. The data will be examined with a focus on properties of the distributions and abnormalities. The properties of the models selected will then be discussed, focusing on which variables were included. After this, the models themselves will be presented in terms of the fit of regression model. Next, the models will be examined for problems that often arise when preforming multiple regression, after which the analyses will be repeated without problematic cases and variables. Finally, a summary will be presented.

Descriptive Statistics

The descriptive univariate statistics for the variables included in this study are presented in Table 1.

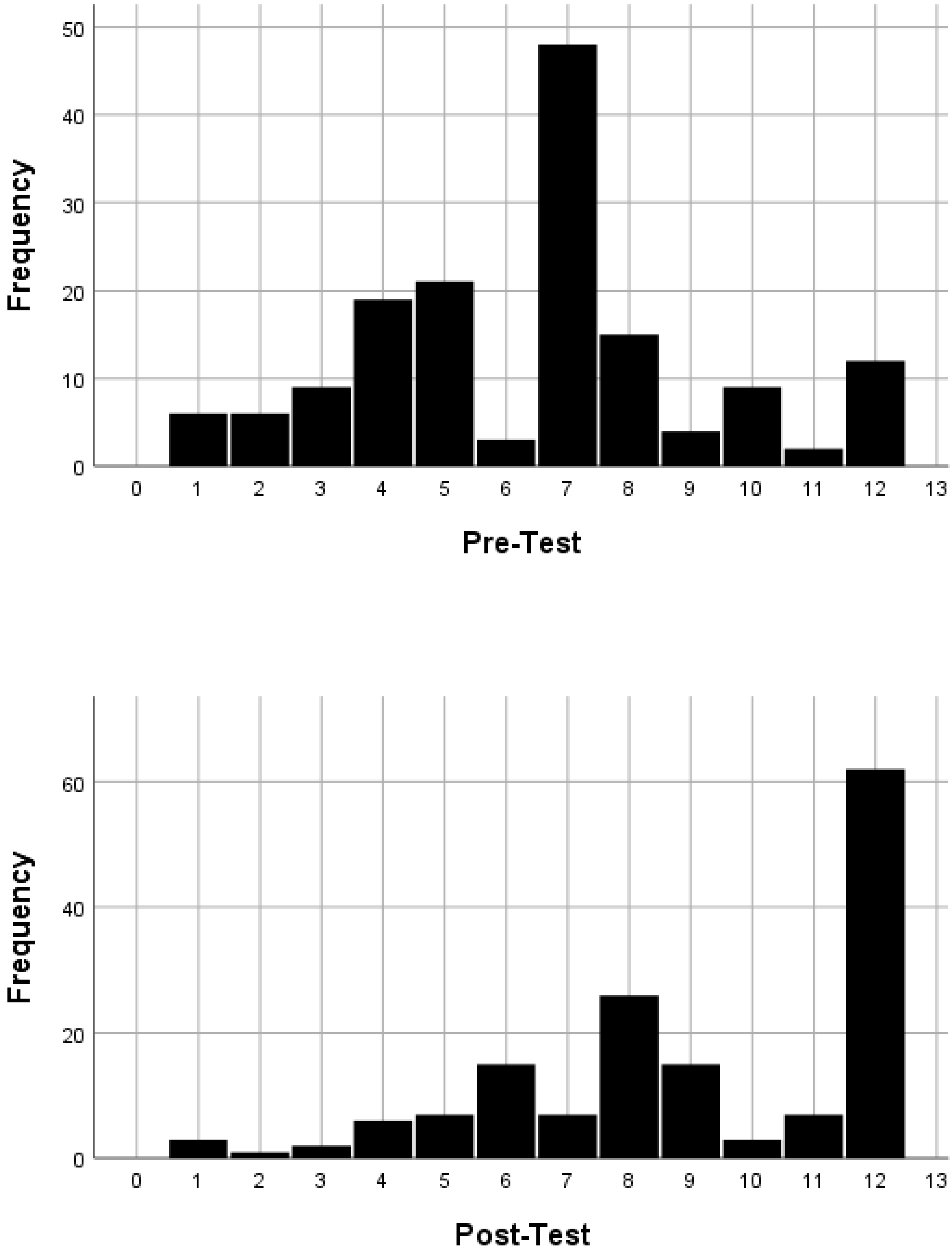
Table 1: *Descriptive Statistics*

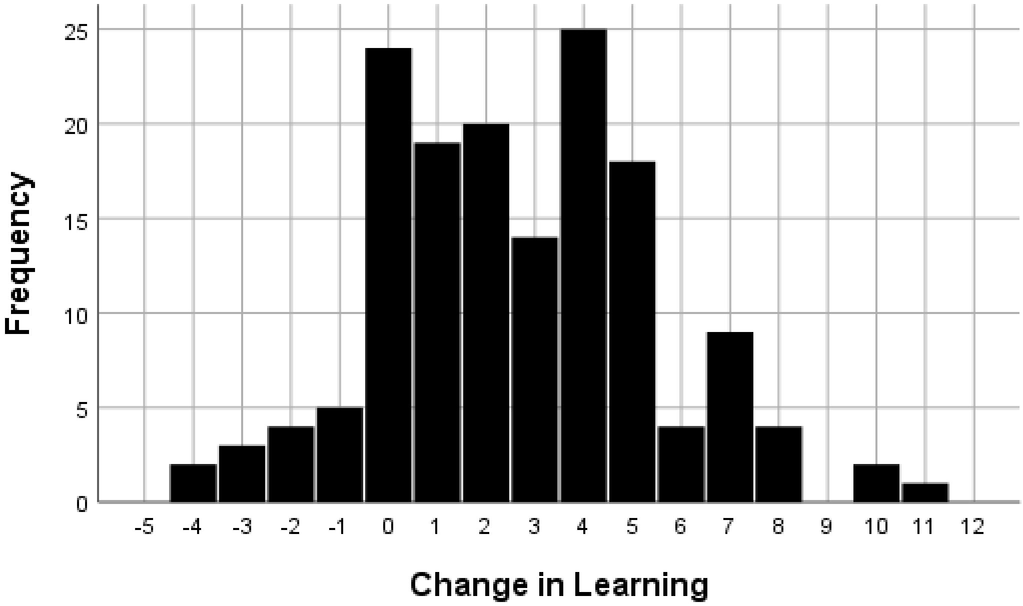
Variable	Mean	Standard Deviation	Skewness SE .186	Kurtosis SE .389	Q1	Median	Q3
Pretest Score	6.44	2.76	.19	-.28	4.00	7.00	8.00
Posttest Score	9.11	2.97	-.68	-.41	7.00	9.00	12.00
Posttest / Pretest Difference	2.67	2.82	.23	.07	.75	2.50	4.25
Content Evaluation Plus	1.01	1.53	2.76	12.31	.00	.00	2.00
Content Evaluation Minus	1.81	2.43	2.05	4.77	.00	1.00	3.00
Coordinating Informational Sources	1.95	3.09	2.58	9.25	.00	1.00	3.00
Draw	.83	2.22	3.77	17.19	.00	.00	.00
Expectation of Adequacy of Content Plus	.75	1.49	3.55	17.55	.00	.00	1.00
Expectation of Adequacy of Content Plus	1.31	2.06	3.24	14.38	.00	1.00	2.00
Feeling of Knowing Plus	5.90	6.46	4.08	24.78	2.00	5.00	8.00
Feeling of Knowing Minus	2.56	2.87	1.50	2.07	.00	2.00	4.00
Help Seeking Behavior	.03	.18	5.33	26.73	.00	.00	.00
Inferences	1.53	1.82	1.84	4.21	.00	1.00	2.00
Interest Plus	2.69	3.33	1.80	3.32	.00	1.00	4.00
Interest Minus	.66	1.70	4.38	24.07	.00	.00	1.00
Judgement of Learning Plus	2.03	2.48	1.52	2.18	.00	1.00	3.00
Judgement of Learning Minus	1.42	2.11	2.40	7.50	.00	1.00	2.00
Knowledge Elaboration	4.04	4.17	1.56	2.80	1.00	3.00	6.00
Memorization	1.66	2.42	2.40	7.35	.00	1.00	2.00
Monitor Progress Toward Goals	.16	.71	7.11	61.00	.00	.00	.00
Monitor Use of Strategies	.09	.31	3.52	12.71	.00	.00	.00
Prior Knowledge Activation	4.45	4.39	1.11	1.21	1.00	3.00	7.25
Planning	.66	1.40	5.18	39.88	.00	.00	1.00
Recycle Goal in Working Memory	.55	1.23	4.10	22.50	.00	.00	1.00

Read Notes	2.34	3.91	2.62	8.50	.00	.00	4.00
Re-Reading	8.75	7.58	1.67	3.50	3.00	7.00	12.00
Search	2.29	2.73	1.56	1.99	.00	1.00	3.00
Sub-Goal	3.74	3.64	1.30	1.10	1.00	3.00	5.00
Summarization	9.19	7.75	1.13	.87	3.00	7.00	13.25
Task Difficulty	.26	.61	3.05	11.58	.00	.00	.00
Time Monitoring	1.18	1.80	1.99	3.92	.00	.00	2.00
Taking Notes	9.76	11.82	1.07	.11	.00	5.00	18.00
Monitor Macro	18.47	14.69	1.30	1.83	7.75	14.50	26.00
Strategy Use Macro	45.12	23.45	.25	-.72	26.00	42.50	64.00
Planning Macro	4.95	4.40	1.04	.34	1.75	4.00	7.00
Positive Variable Macro (>.1)	21.41	13.55	.89	.88	11.00	19.00	30.25
Positive Variable Macro (>.05)	37.29	20.78	.43	-.23	20.00	37.50	50.25
Positive Variable Macro (>.01)	46.83	24.99	.42	-.48	26.00	46.00	65.25
Negative Variable Macro (>.1)	11.56	9.06	1.88	5.34	5.00	9.00	16.00
Negative Variable Macro (>.05)	12.87	10.16	1.79	4.44	6.00	10.00	17.25
Negative Variable Macro (>.01)	25.57	17.00	1.08	1.48	13.00	21.00	36.00

The statistics show that, on average, students came into this task with an average score on the knowledge measure of 6.44 (*SD* = 2.76) and that they gained 2.67 (*SD* = 2.82) points on average when completing the post-test. Histograms for these variables can be seen in Figure 2.

Figure 2: *Histograms of Knowledge Measures*

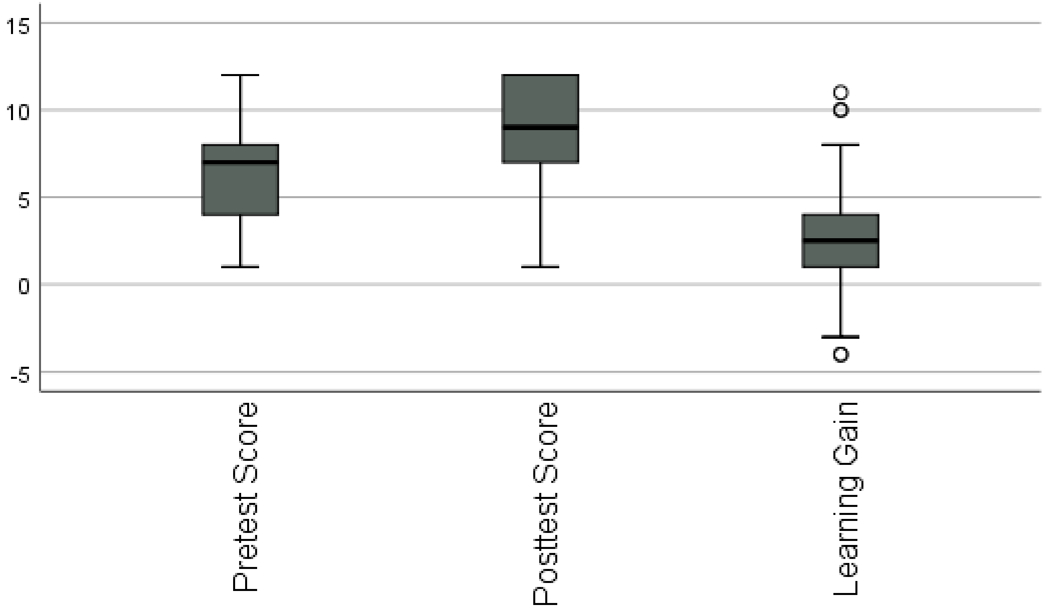




These tables and figures show that not everyone learned during the learning task. Whereas pre-test scores were roughly normally distributed with a center of 7, the post-test scores show extreme negative skew, with 62 out of 154 cases having a perfect score. This resulted in a ceiling effect, where the knowledge measure in the task was limited by the maximum score of the test that measured it (Uttl, 2005). The impact of ceiling effects will be expanded upon in the discussion. The difference between the pre- and post-test showed that whereas scores increased for 75% of the participants, they remained the same for 16% participants and decreased for another 9%. Concerning an increase in score, 87% gained 5 or fewer points, with only 13% people gaining more than 5. Box plots of these scores (Figure 3) show that there may be outliers present, with students decreasing 4 or more and gaining 10 or more points on the on the post-test being problematic.

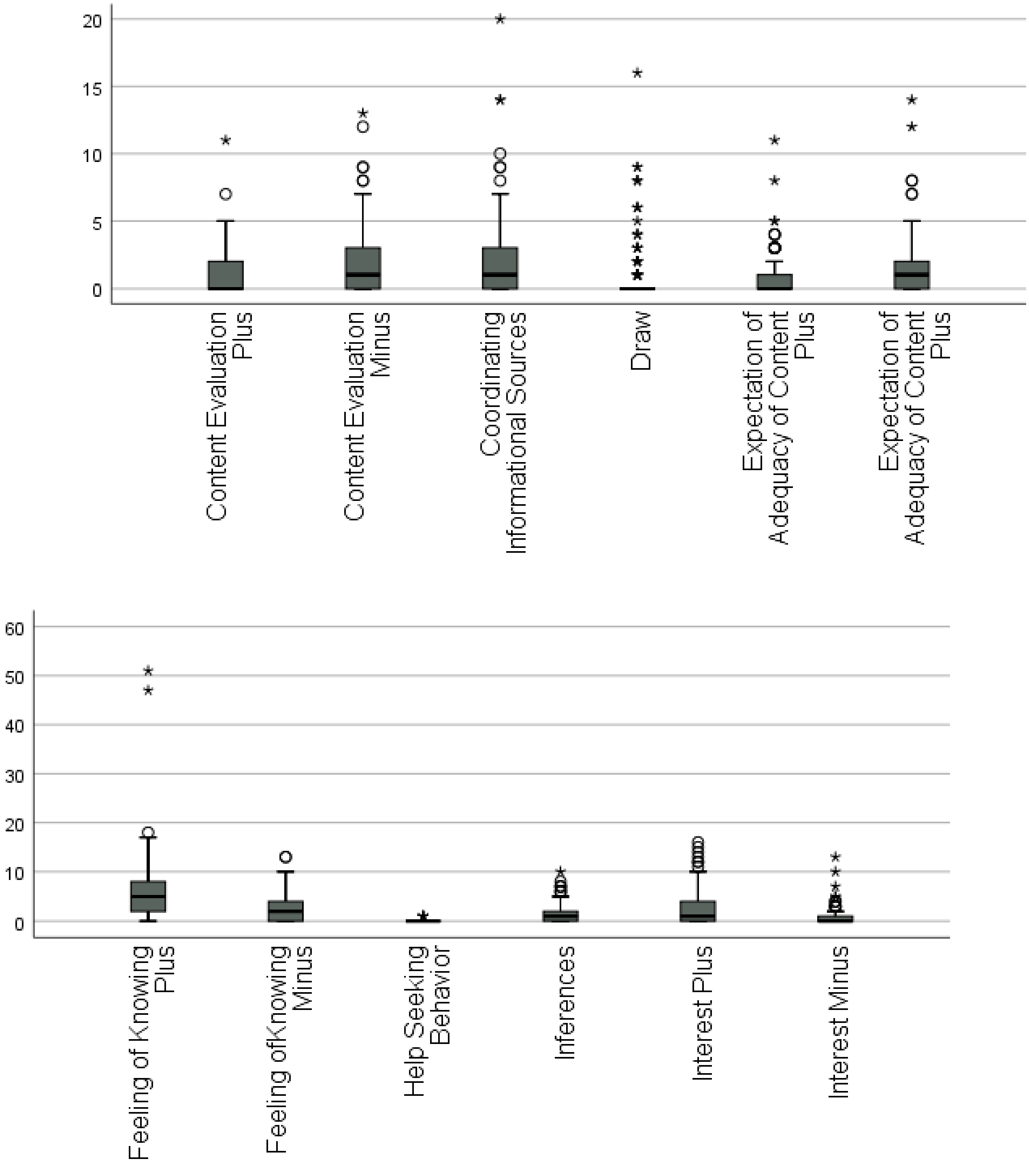
Figure 3:

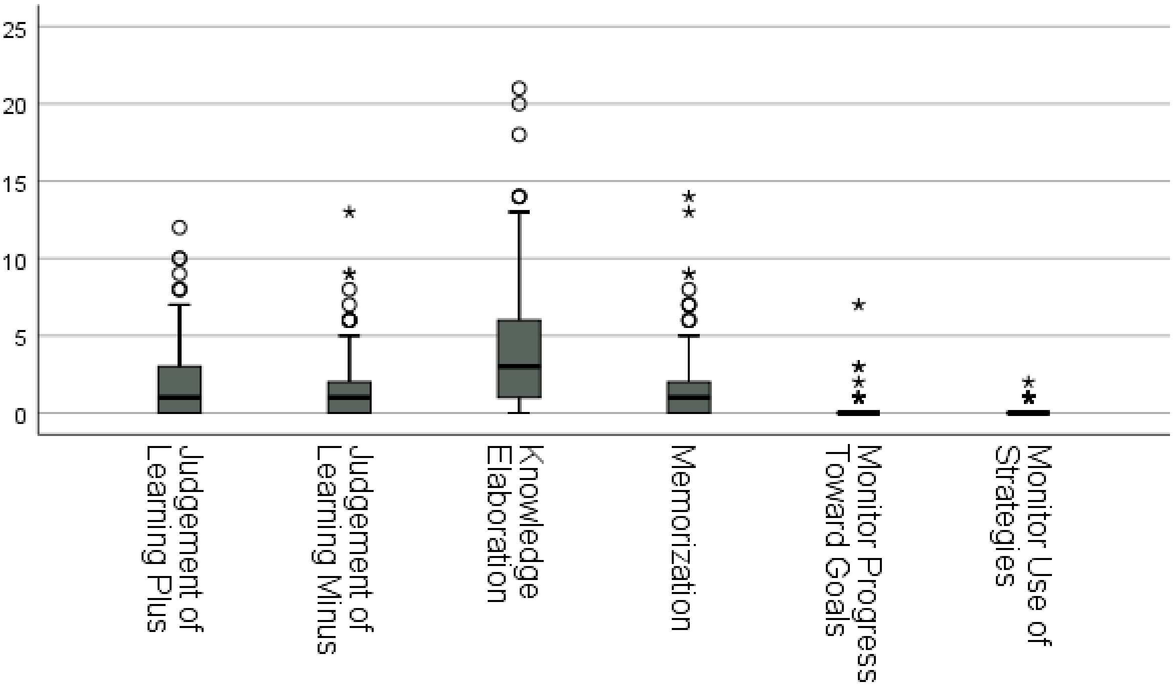
Box Plot of Learning Measures.

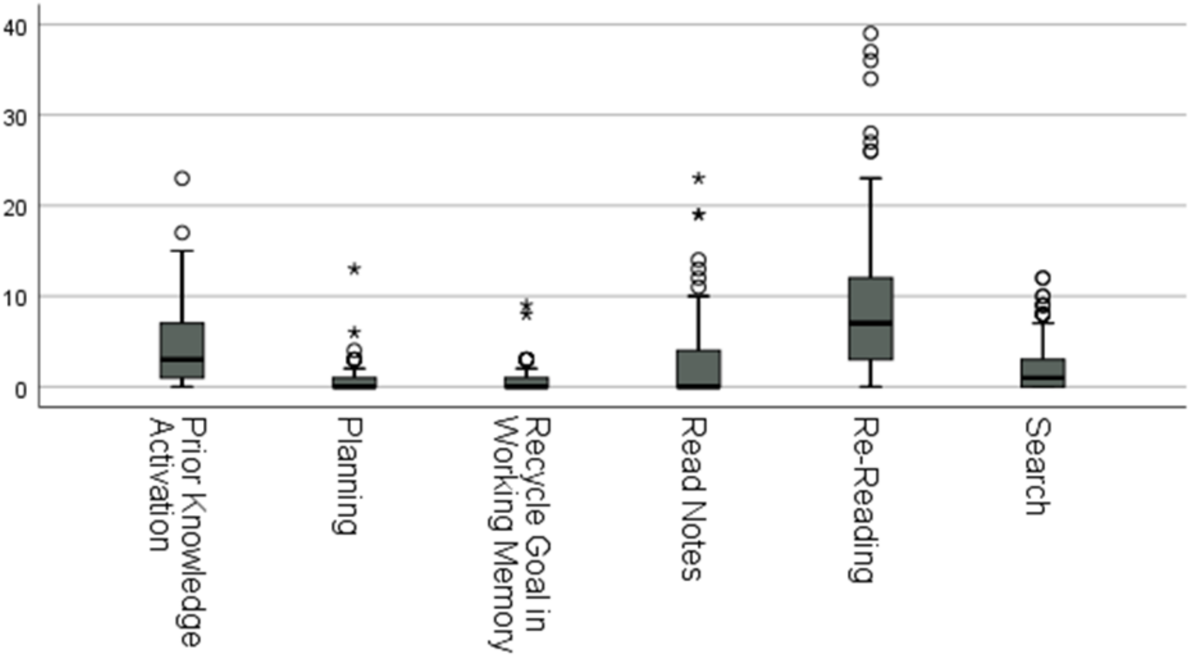


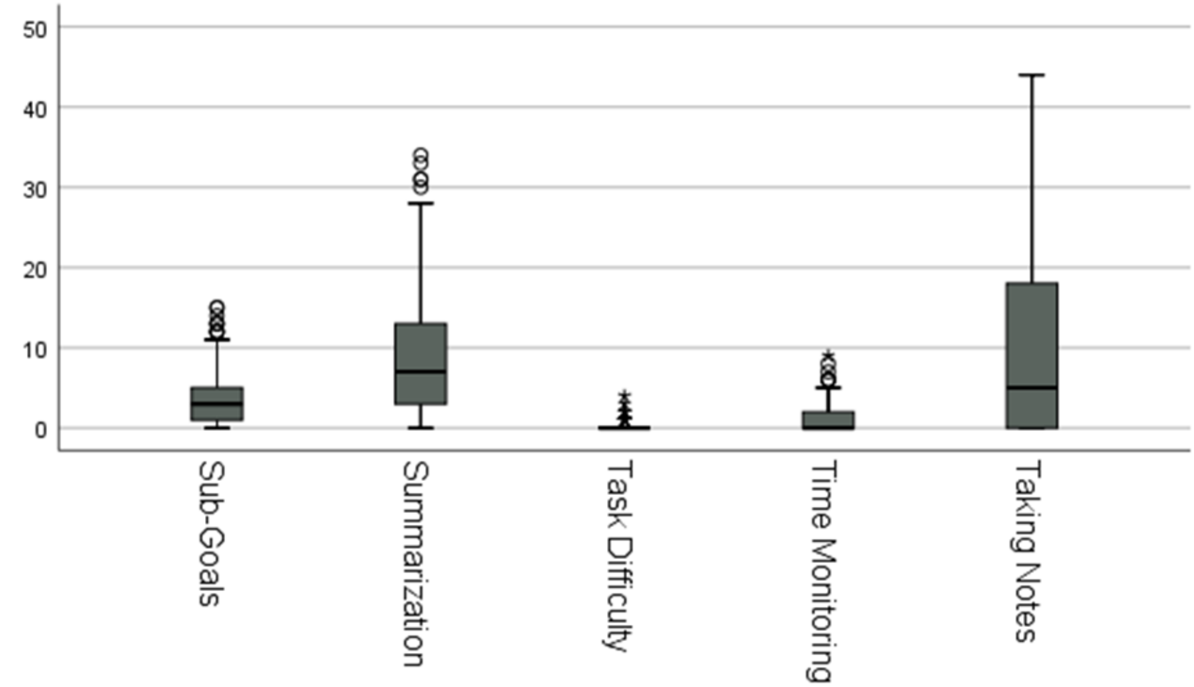
Self-regulated learning processes were varied in use and consisted of positive skew of varying degrees. Box plots for the self-regulated learning processes are illustrated in Figure 4. Stars note the presence of outliers (i.e., extremes greater than three times the interquartile range) in all boxplots.

Figure 4: *Box Plots of Self-Regulated Learning Processes*




25
20
15
10
5
0
Judgement of Learning Plus
Judgement of Learning Minus
Knowledge Elaboration
Memorization
Monitor Progress Toward Goals
Monitor Use of Strategies


40
30
20
10
0
Prior Knowledge Activation
Planning
Recycle Goal in Working Memory
Read Notes
Re-Reading
Search



The boxplots show some problems with the data. Most variables are centered near-zero, with a few having centers between 2 and 5. With the positive skew present in most variables, traditional boxplots will label these as outliers, having greater than 1.5 times the interquartile range (or box size), and the higher values that are three times the interquartile range would be labeled as extremes (Field, 2017). It is clear as well that most of these variables do not have a normal distribution. Most lack whiskers and appear to be compressed into the bottom of the graph. Feeling of Knowledge Plus shows two scores that clearly differ from the rest of the data, as they are more than 20 points higher than the scores below them.

The means of four measures, Task Difficulty, Monitoring Use of Strategies, Monitoring Progress Towards Goals, and Help Seeking Behavior were extremely low. Task Difficulty was only used by 20% of the sample, whereas Help Seeking Behavior was only coded for 5 participants, and Monitoring Use of Strategies and Monitoring Progress Towards Goal both were enacted by 13 participants. Since less than 20% of the participants used these processes, these cases would be given undo weight in the regression analysis. For example, only five students determined the relationship between Help Seeking Behavior and learning. Most datapoints for Task Difficulty, Monitoring Use of Strategies, Monitoring Progress Towards Goals, and Help Seeking Behavior were labeled as outliers on the boxplots as well. Variables in this pattern are referred to as near-zero variance variables (Kuhn, 2013) and are problematic because they can result in a few cases having an undue influence on the models and they lack linearity (Field, 2017).

Correlations are show below in Table 2.

Table 2: *Correlations*

	Post to Pretest Change	Content Evaluation Plus	Content Evaluation Minus	Coordinating Informational Sources	Draw	Expectation of Adequacy of Content Plus
Pre To Posttest Score Change	1					
Content Evaluation Plus	.026	1				
Content Evaluation Minus	-.124	.241	1			
Coordinating Informational Sources	.316	.019	-.021	1		
Draw	-.100	.227	.049	.125	1	
Expectation of Adequacy of Content Plus	.027	.504	.383	.051	.242	1
Expectation of Adequacy of Content Plus	-.078	.186	.621	-.046	.026	.261
Feeling of Knowing Plus	-.111	.172	.121	.046	.032	.157

Feeling of Knowing Minus	-.044	.258	.069	-.018	.070	.295
Help Seeking Behavior	.048	-.025	-.031	.074	-.036	.006
Inferences	.260	.200	.181	.214	.206	.365
Interest Plus	-.033	.258	.122	.135	-.011	-.033
Interest Minus	-.113	.023	.201	-.105	-.017	.255
Judgement of Learning Plus	.059	.320	.265	.149	.177	.458
Judgement of Learning Minus	.046	.228	.287	.119	.219	.297
Memorization	-.104	.335	.118	-.031	.065	.205
Monitor Progress Toward Goals	-.210	.235	.178	-.077	-.021	.248
Monitor Use of Strategies	.057	.150	.214	.025	.165	.289
Prior Knowledge Activation	-.036	.123	.128	.039	.049	.144
Planning	-.020	.211	.082	-.072	.053	.294
Recycle Goal in Working Memory	-.100	.203	.332	-.030	.049	.332
Read Notes	-.019	.200	.082	.016	.229	.095
Re-Reading	.191	.010	.067	.251	.004	.090
Search	.043	.237	.160	-.053	.063	.181
Sub-Goal	.090	.377	.307	.070	.072	.383
Summarization	.141	.141	.150	.121	-.050	.297
Task Difficulty	.073	.089	.130	-.027	.037	.128
Time Monitoring	-.001	.165	.214	-.026	.112	.301
Taking Notes	.063	.235	.026	-.001	.236	.142

	Expectation of Adequacy of Content Minus	Feeling of Knowing Plus	Feeling of Knowing Minus	Help Seeking Behavior	Inferences	Interest Plus
Expectation of Adequacy of Content Plus	1					
Feeling of Knowing Plus	.269	1				
Feeling of Knowing Minus	.158	.506	1			
Help Seeking Behavior	-.027	-.043	-.100	1		
Inferences	.086	.168	.278	-.013	1	
Interest Plus	.073	.301	.376	-.027	.096	1
Interest Minus	.097	.118	.164	.058	.070	.115
Judgement of Learning Plus	.172	.288	.436	.087	.405	.367
Judgement of Learning Minus	.147	.079	.227	.016	.378	.127
Knowledge Elaboration	.157	.354	.330	-.019	.276	.393
Memorization	-.008	-.051	.028	-.050	.128	.045
Monitor Progress Toward Goals	.287	.111	.185	-.041	-.003	.090
Monitor Use of Strategies	.151	.256	.170	-.054	.308	.110
Prior Knowledge Activation	.133	.385	.415	.040	.186	.261
Planning	.227	.181	.197	-.034	.047	-.016
Recycle Goal in Working Memory	.394	.210	.215	-.083	.103	.122
Read Notes	.027	.051	.071	.285	.233	-.036
Re-Reading	-.041	-.078	-.072	.055	.139	.026
Search	.117	.055	.176	.048	.237	.093
Sub-Goal	.186	.122	.148	-.037	.351	.148
Summarization	.138	.200	.252	-.114	.318	.170
Task Difficulty	.113	.106	.109	-.078	.181	-.018

	Interest Minus	Judgement of Learning Plus	Judgement of Learning Minus	Knowledge Elaboration	Memorization	Monitor Progress Toward Goals
Interest Minus	1					
Judgement of Learning Plus	.170	1				
Judgement of Learning Minus	.161	.523	1			
Knowledge Elaboration	-.019	.328	.249	1		
Memorization	.083	.183	.154	.011	1	
Monitor Progress Toward Goals	-.043	.117	.022	.116	-.038	1
Monitor Use of Strategies	-.003	.192	.272	.255	.067	.024
Prior Knowledge Activation	-.050	.323	.182	.526	.113	.159
Planning	.040	.197	.085	.112	.078	.589
Recycle Goal in Working Memory	.177	.269	-.041	.126	.048	.210
Read Notes	-.015	.386	.303	.034	.074	.014
Re-Reading	.095	.142	.193	.013	.182	-.106
Search	.194	.115	.141	.042	.146	-.023
Sub-Goal	.133	.309	.254	.260	.226	.100
Summarization	.027	.236	.224	.457	.104	.198
Task Difficulty	.166	.192	.270	.070	.179	.027
Time Monitoring	.303	.332	.162	.319	.187	.142
Taking Notes	.015	.236	.106	-.043	-.059	.099

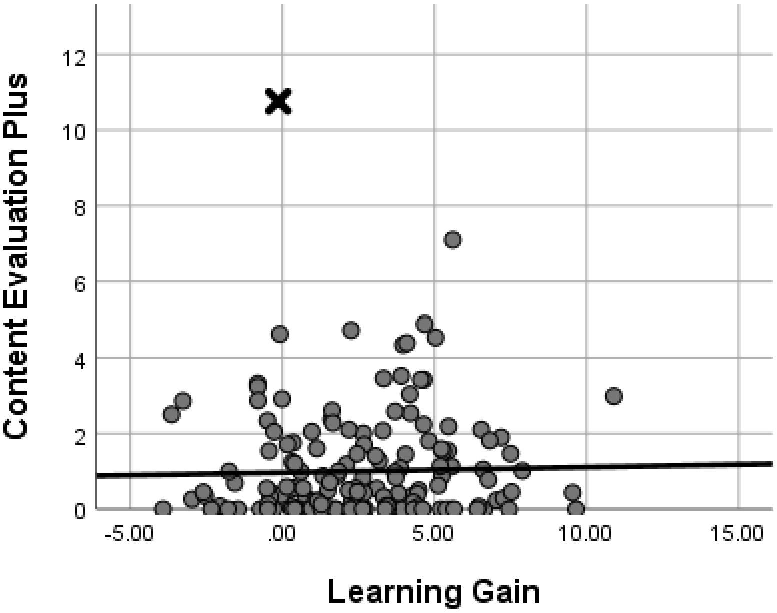
	Monitor Use of Strategies	Prior Knowledge Activation	Planning	Recycle Goal in Working Memory	Read Notes	Re-Reading	Search
Monitor Progress Toward Goals	1						
Prior Knowledge Activation	.032	1					
Planning	.086	.167	1				
Recycle Goal in Working Memory	.073	.202	.109	1			
Read Notes	.120	.071	.100	.090	1		
Re-Reading	.054	-.090	.031	-.001	.085	1	
Search	.077	.146	-.016	.175	-.006	-.048	1
Sub-Goal	.293	.200	.065	.176	.170	.086	.275
Summarization	.052	.356	.220	.132	-.031	.176	.046
Task Difficulty	.184	.194	-.027	.060	.080	.003	.119
Time Monitoring	.204	.303	.278	.333	.182	.201	.323
Taking Notes	.090	-.114	.185	.176	.558	.055	.053

	Sub-Goal	Summarization	Task Difficulty	Time Monitoring	Taking Notes
Sub-Goal	1	.360	.285	.310	.019
Summarization	.360	1	.218	.171	-.039
Task Difficulty	.285	.218	1	.093	.069
Time Monitoring	.310	.171	.093	1	.144
Taking Notes	.019	-.039	.069	.144	1

There are a few things to note in Table 2. First, the highest correlation was between Monitoring Progress Towards Goal and Planning, which was .589. This indicated that problematic multicollinearity should not be present in this dataset (Field, 2017). The second issue to note were the correlations between learning gain and Task Difficulty (.073), Monitoring Use of Strategies (.057), Monitoring Progress Towards Goals (-.210), and Help Seeking Behavior (.048). These four variables showed very little variation (see Table 1), and three of them had near zero correlation scores with learning gain. Monitoring Progress Towards Goals was also problematic, as all but 13 scores were 0, with 12 at 1 and 1 at 2. The correlation with learning gain was entirely based on these 13 scores.

A full scatterplot matrix of all 553 bivariate comparisons was too large to include in this paper (see supplemental material for this). Instead, a much-reduced set of scatterplots are discussed where problems of note were observed. Figures 5 to 12 present various problems that were observed with descriptions of what the problem is and why it is a problem. The scatterplots have been jittered to prevent overlapping points.

Figure 5: *Scatterplot Content Evaluation Plus and Learning Gain*



In this figure, one value stands apart. The participant used content evaluations 11 times, and although the participant indicated that the material was useful, his/her score on the post-test did not increase. Notably, this participant had the highest rate for showing interest in this task, but pre-scores were low (4) and no gain in knowledge was made.

Figure 6: *Scatterplot Content Evaluation Plus and Draw*

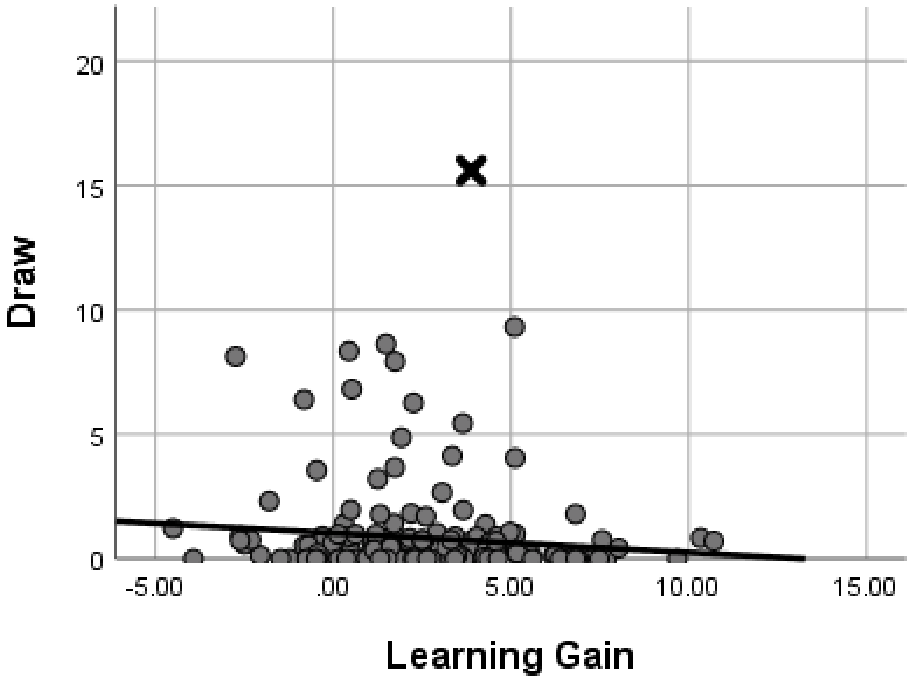
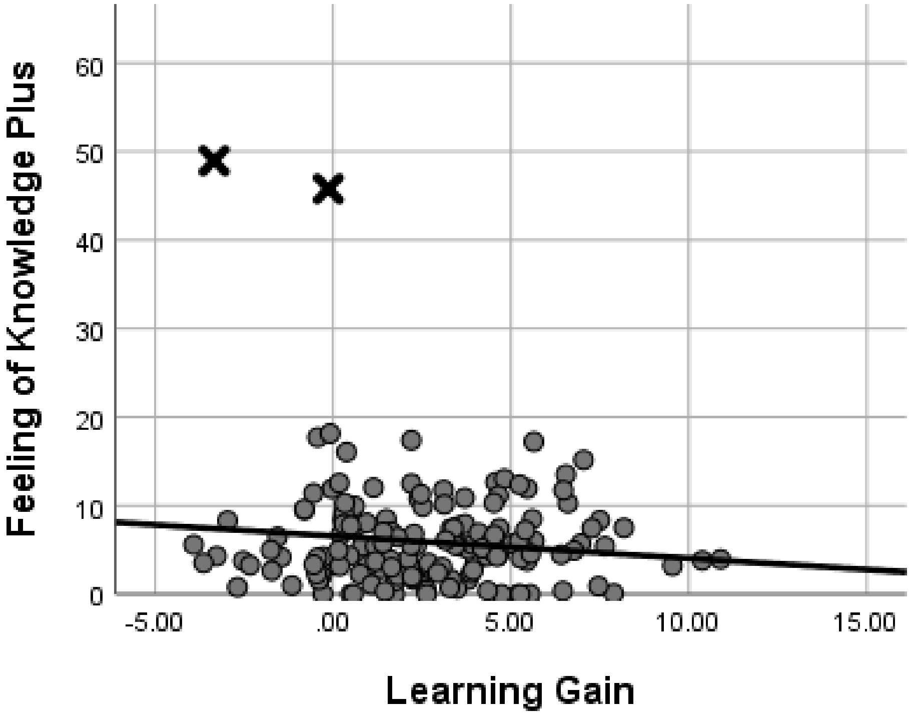


Figure 6 seems to show a problem with the draw strategy. Analysis of the outlier case here shows that while the participant did draw more than any other participant, there are no other abnormalities in his self-regulated learning usage. The participant scored a perfect score on the post-test and an eight on the pre-test, and the rest of their self-regulated learning use was in line with what others had used.

Figure 7: *Scatterplot Content Evaluation Plus and Learning Gain*



As stated previously, Feeling of Knowledge Plus had two scores that seemed to be set apart from the others (Figure 7). Looking at these two cases in a bivariate setting, they clearly are problematic. One of the participants had a perfect score on the pre-test and post-test. However, this participant noted taking several biology classes in the past, which may explain this value. The other participant had several other concerning values for self-regulated learning processes. First, they had the highest use of knowledge elaborations. In addition, he/she was recorded using 214 self-regulated learning processes; only one person was higher at 222. Finally, the participant started the task with a perfect pre-test score and lost three points on the post-test.

Figure 8: *Scatterplot Help Seeking Behavior and Learning Gain*

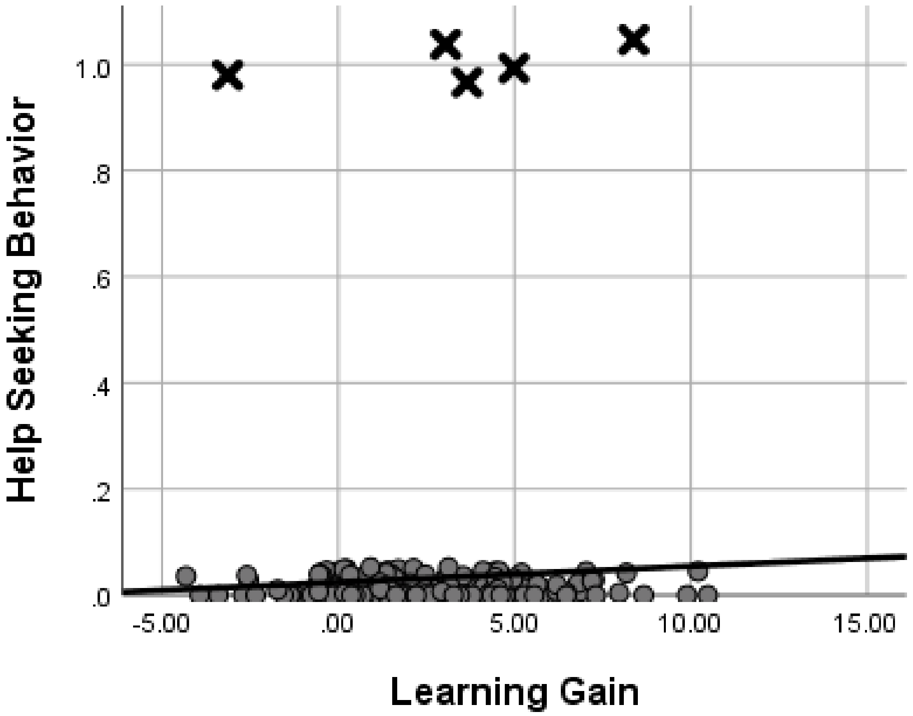


Figure 8 shows little variance as most participants did not use help seeking behavior. Notably, it also shows that this measure should not be categorized as an interval variable, as scores were or either 0 or 1, meaning this would be best recoded as a categorical or binary variable, and the assumption of a linearity may not be met here. Furthermore, since there is almost no variation in scores what can be learned from this variable is limited using a method that relies on variance like ordinary least squares regression.

Figure 9: *Scatterplot Monitoring Progress Towards Goals by Learning Gain*

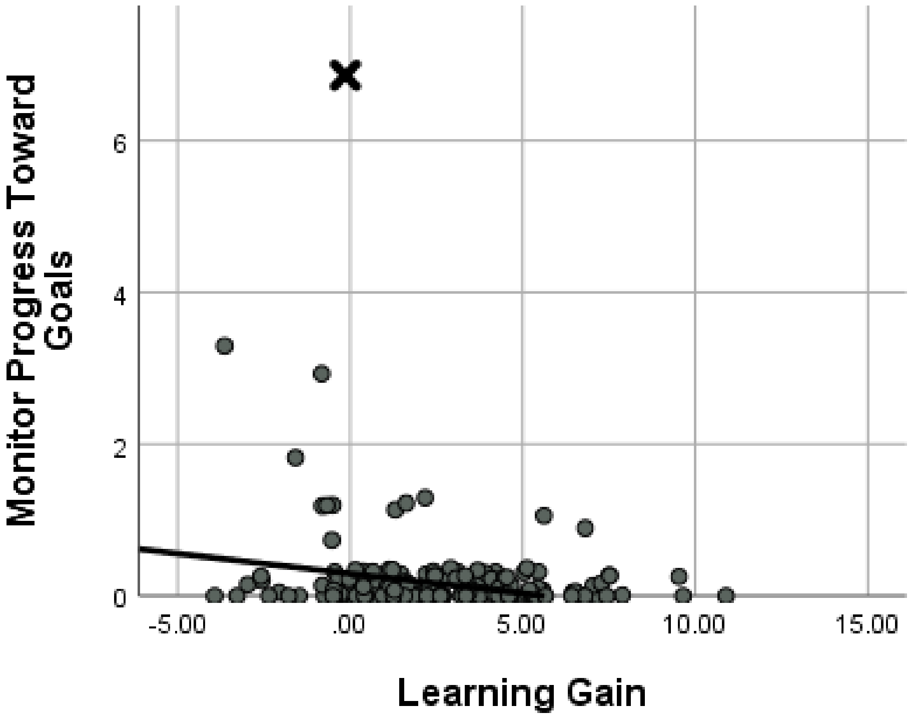
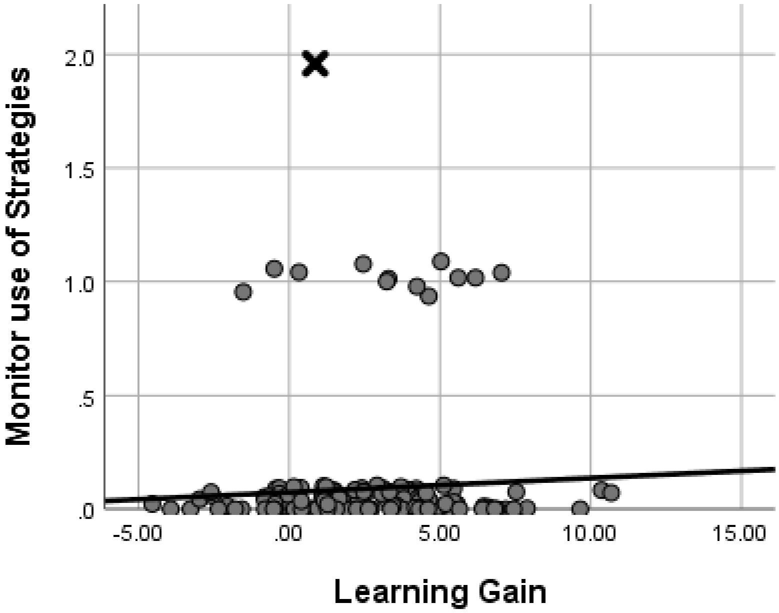


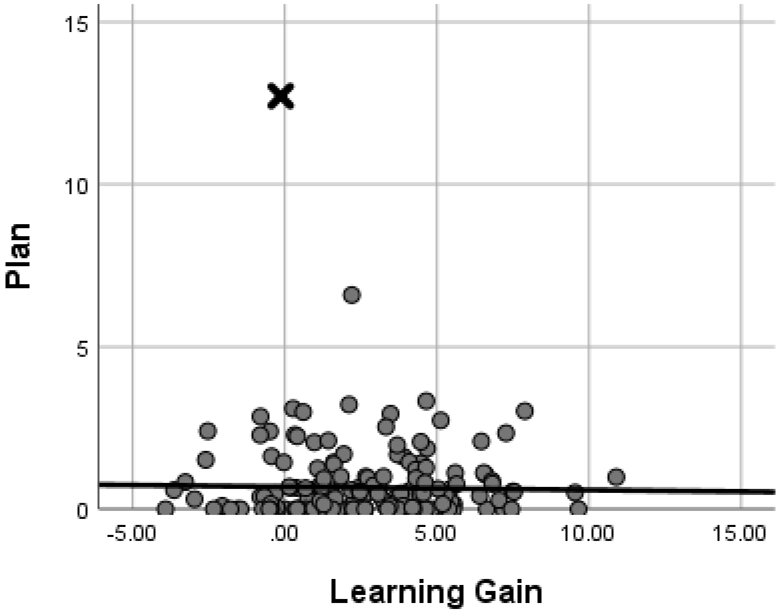
Figure 9 shows a clear outlier in the bivariate relationship between Monitoring Progress Towards Goals and learning gain. Most participants did not engage in this self-regulated learning process, whereas the nine cases above 1 are roughly uniformly distributed across learning gain. There is one case at 2, two cases at 3, and one case at 7. One of the cases at 3 and the case at 7 had a learning gain of 0. The cases at 2 had a learning gain of -1. The second case at 3 had a learning gain of -4. Removing these two cases with negative learning gains subsequently reduced the correlation to -.139. Dropping the other three cases further reduced the correlation to -.111. This illustrates how near-zero variance can create instability and biased results, as 3% of the sample nearly doubled the correlation alone.

Figure 10: *Scatterplot of Monitoring use of Strategies and Learning Gain*



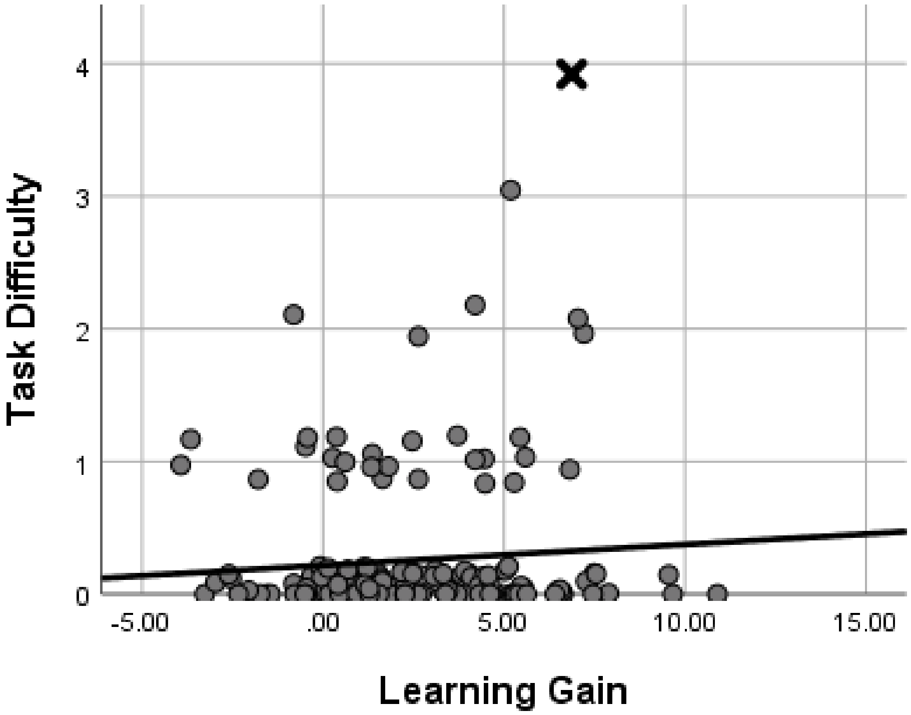
In Figure 10, like Help Seeking Behavior, Monitor Use of Strategies appears to be more categorical in nature than an interval variable. Here, most points are at 0, a few at 1, and a single point at 2. Thus, the usefulness of this variable seems limited.

Figure 11: *Scatterplot of Planning and Learning Gain*



In Figure 11, a single point stands out alone again, with a participant having made 13 plans while completing the learning task, but not showing any learning gain. This participant also received a perfect score on the pre-test and repeated this performance on the post-test.

Figure 12*: Scatterplot of Task Difficulty and Learning Gain*



Finally, Figure 12 shows the relationship between Task Difficulty and learning gain. As can be seen, most values are equal to zero, with an additional 23 cases at 1, 5 at two, and 1 at both 3 and 4. This variable had a low correlation with learning gain and approached a near uniform distribution when the seven highest Task Difficulty cases were removed.

Creation of the Data-Driven Models

To create the Data-Driven Models, macro-level variables were created of composites of items that negatively or positively correlated to learning gain. Table 3 shows the variables that were included in each level. At the most restrictive level, the .1 level, the positive macro-level variable consisted of Coordinating Sources of Information, Inference, Summarization, and Re-Reading. At the .05 level, Monitoring use of Strategies, Judgments of Learning Plus, Sub-Goals, Task Difficulty entered the model. At the least restrictive level (.01), Content Evaluation Plus and Help Seeking Behavior entered the model. For the negative macro, at the most restrictive level (.1), Drawing, Content Evaluation Minus, Feeling of Knowledge Plus, Interest Minus, Memorization, Monitoring Progress Towards Goals, and Recycle Goal in Working Memory entered the model. At the next level (.05), Expectation of Adequacy of Content Minus entered the model. At the least restrictive level (.1), Feeling of Knowledge Minus, Interest Plus, Prior Knowledge Activation, Plan, and Read Notes entered the model.

Table 3: Data-Driven Aggregation Variable Map

.1	.05	.01	Variable	-.01	-.05	-.1
■	■	■	Coordinating Sources of Information			
■	■	■	Inference			
■	■	■	Summarization			
■	■	■	Re-Reading			
	■	■	Monitoring Use of Strategies			
	■	■	Judgements of Learning Plus			
	■	■	Sub-Goal			
	■	■	Task Difficulty			
	■	■	Taking Notes			
		■	Content Evaluation Plus			
		■	Expecting Adequacy of Content Plus			
		■	Help Seeking Behavior			
			Draw	■	■	■
			Content Evaluation Minus	■	■	■
			Feeling of Knowledge Plus	■	■	■
			Interest Minus	■	■	■
			Memorization	■	■	■
			Monitor Progress Toward Goals	■	■	■
			Recycle Goal in Working Memory	■	■	■
			Expectation of Adequacy of Content Minus	■	■	
			Feeling of Knowledge Minus	■		
			Interest Plus	■		
			Prior Knowledge Activation	■		
			Plan	■		
			Read Notes	■		

Best All Subset Model

After running the best all subset regression, the top 10 models, as determined by those having the highest adj R^2 value, were selected. They are presented Table 4.

Table 4: *Variable Inclusion Chart*

Variable	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	Core
Adj R^2	.280	.278	.277	.276	.275	.274	.274	.273	.273	.273	.269
Content Evaluation Plus	■	■	■	■	■	■	■	■	■	■	■
Coordinating Information Sources	■	■	■	■	■	■	■	■	■	■	■
Draw	■	■	■	■	■	■	■	■	■	■	■
Feeling of Knowledge Plus	■	■	■	■	■	■	■	■	■	■	■
Inference	■	■	■	■	■	■	■	■	■	■	■
Interest Minus	■	■	■	■	■	■	■	■	■	■	■
Memorize	■	■	■	■	■	■	■	■	■	■	■
Monitor Progress Towards Goal	■	■	■	■	■	■	■	■	■	■	■
Plan	■	■	■	■	■	■	■	■	■	■	■
Task Difficulty	■	■	■	■	■	■	■	■	■	■	■
Content Evaluation Minus	■	■	■	■		■	■	■			
Read Notes	■	■		■	■		■		■	■	
Re-Reading	■	■	■		■			■		■	
Take Notes	■			■	■			■	■		

Since these 10 models shared all but four variables, a core model was created from these shared variables. The Core Model would allow for determining the unvarying part of these models directly and provide a contrast point for the other models. Beta weights for the variables within these models are provided in Table 5.

Table 5: *Betas for Best Subset Models*

Variable	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	Core
Content Evaluation Plus	.173	.188	.178	.171	.155	.173	.183	.152	.150	.160	.157
Coordinating Information Sources	.240	.237	.242	.244	.244	.268	.264	.249	.268	.247	.270
Draw	-.219	-.210	-.223	-.228	-220	-228	-216	-229	-226	-.223	-.227
Feeling of Knowledge Plus	-.186	-.185	-.184	-.185	-191	-195	-196	-190	-201	-.190	-.200
Inference	.274	.271	.258	.256	.263	.265	.278	.243	.270	.245	.253
Interest Minus	-.112	-.109	-.105	-.105	-130	-095	-099	-123	-120	-.124	-.114
Memorize	-.219	-.236	-.233	-.227	-221	-217	-219	-230	-204	-.237	-.221
Monitor Progress Towards Goal	-.340	-.340	-.334	-.332	-360	-348	-354	-354	-372	-.356	-.368
Plan	.218	.229	.221	.216	.227	.235	.243	.226	.238	.232	.244
Task Difficulty	.131	.136	.132	.130	.126	.129	.132	.125	.123	.127	.125
Content Evaluation Minus	-.101	-.107	-.109	-.107	---	-102	-100	---	---	---	---
Read Notes	-.130	-.080	---	---	-136	---	-073	---	-133	---	---
Re-Reading	.098	.103	.097	.094	.091	---	---	.087	---	.090	---
Take Notes	.099	---	---	.030	.107	---	---	.035	.113	---	---

It is worth noting that all the variables in the Best All Subset Model also appeared in the Data-Driven Model, and the beta weight signs align with the macro-level variable to which the variable was added.

Model results. After the variables included in the different models were determined, standard multiple regressions were run on the full aggregation, the three data-driven aggregation models, the best all subset model and the core model from the best all subset regression. The results are in Table 6.

Table 6: Model Comparisons

Model	Variable	β	*t*	VIF	*f*	*Df*(*f*)	Adj R^2
Macro-level					4.041***	3,153	.056
	Planning	.045	.477	1.468			
	Monitoring	-.243	-2.445**	1.604			
	Strategy Use	.292	3.140***	1.403			
Data Driven Model .1					13.421***	2,153	.140
	Positive	.418	4.760***	1.372			
	Negative	-.371	-4.224***	1.372			
Data Driven Model .05					13.997***	2,153	.145
	Positive	.355	4.504***	1.111			
	Negative	-.320	-4.058***	1.111			
Data Driven Model .01					14.699***	2,153	.152
	Positive	.348	4.588***	1.037			
	Negative	-.281	-3.796***	1.037			
Best All Subset Model					5.241***	14,153	.280
	Content Evaluation Minus	.098	-1.359	1.176			
	Coordinating Information Sources	.240	3.224**	1.174			
	Draw	-.219	-2.975**	1.156			
	Feeling of Knowledge Plus	-.186	-2.543*	1.137			
	Inference	.274	3.590***	1.239			
	Interest Minus	-.112	-1.538	1.122			
	Memorize	-.219	-2.787**	1.310			

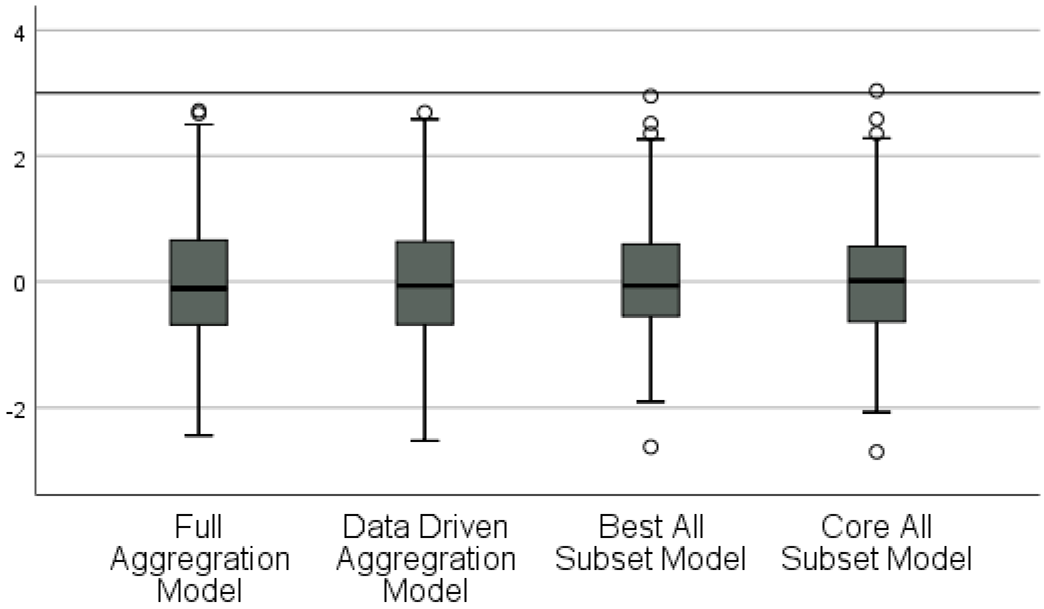
	Monitor Progress Towards Goal	-.340	-3.769***	1.724			
	Plan	.218	2.444*	1.691			
	Task Difficulty	-.130	1.806	1.117			
	Read Notes	.131	-1.522	1.554			
	Re-Reading	.099	1.306	1.184			
	Take Notes	-.101	1.136	1.608			
Core Subset Model					6.634***	10,152	.269
	Content Evaluation Minus	.157	1.965	1.335			
	Coordinating Information Sources	.270	3.744***	1.088			
	Draw	-.227	-3.128**	1.106			
	Feeling of Knowledge Plus	-.200	-2.727**	1.120			
	Inference	.253	3.365***	1.179			
	Interest Minus	-.114	-1.596	1.067			
	Memorize	-.221	-2.896**	1.224			
	Monitor Progress Towards Goal	-.368	-4.150***	1.647			
	Plan	.244	2.771**	1.626			
	Task Difficulty	.125	1.715	1.107			

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

Results from these models show several trends. First, the best fitting model would be the Best Subset Model that has an adj R^2 value of .280 compared to the Full Aggregation Model's .056 and the best of the three Data-Driven Models (i.e., the .1 level model) with an adj R^2 of .152. If the best fitting model was reduced to only variables that occurred in all 10 of the top fitting models, it would achieve an adj R^2 value of .269, a loss of only .011. Furthermore, whereas the Full Aggregation Model and Data-Driven Models cannot show the weight any variable has within the larger macro-level construct, the Best Subset and the Core Model do not have this drawback and allow for direct examination of the effects that each self-regulated learning process has within the model. The Best Subset Model consisted of Coordinating Information Sources, Draw, Feeling of Knowledge Plus, Inference, Interest Minus, Memorize, Monitor Progress Towards Goal, Plan, Task Difficulty, Content Evaluation Minus, Read Notes, Re-Reading, and Take Notes. It was noted that the 10 best all-subset regression models shared a core set of variables. A Core Model was created of these variables consisting of Content Evaluation Minus, Coordinating Information Sources, Draw, Feeling of Knowledge Plus, Inference, Interest Minus, Memorize, Monitor Progress Towards Goal, Plan, Task Difficulty. I observed a pattern at this time that the t-tests from the Best All Subset Model were significant for most of the variables moved to the Core Model.

Regression diagnostic analyses. Boxplots for standardized residuals are presented in Figure 13. Residuals are expected to be normally distributed, with most *z*-scores being between 1 and -1 (Field, 2017). With 153 cases, some are expected between -2 and -1 as well as 1 and 2. Values over 2 or under -2 are generally seen as outliers, with values over 3 or under -3 being seen as potentially problematic (Field, 2017). Both the Best All Subset Model and the Core All Subset Model have a single residual at the $z = 3$ level, which was the same participant in both models. Examining the case, it was a participant who showed a negative learning gain, dropping four points from pre- to post-test. Four other cases may be outliers as well for having values over 2 or under -2.

Figure 13: *Boxplot of Residuals*



To test for normality of residuals, the Shaprio-Wilks test was used. Results are shown in Table 7. Significant results in these tests would indicate that one should reject the assumption that the data is normally distributed. None of the results of this test indicated that the assumption of normality had been violated (Field, 2017).

To assess linearity, scatterplots of predicted versus residual values have been provided in Figure 14. In cases were non-linearity is present, a curved pattern can be seen in the residuals. In this case, the residuals of all four models appear distributed in a circular pattern around a central point (Field, 2017). Heterogeneity of variance can be assessed in these residual plots as well. If the spread of predicted values has large residuals, a funnel or triangular shape will appear in the plots. This was not present in the residual plots for these models. Running the

Breusch-Pagan test supported the hypothesis that the variance is normally distributed (Field, Miles, & Field, 2012).

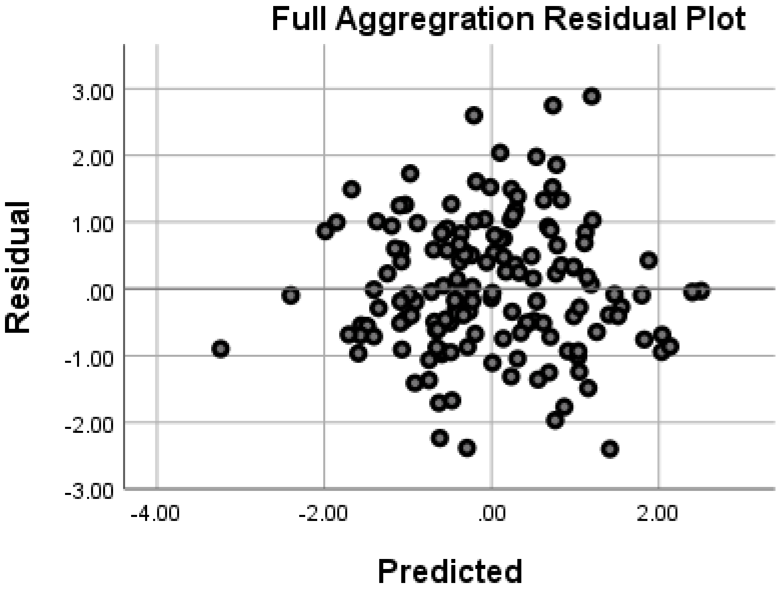
To assess multicollinearity of variance, the variable inflation factor (VIF) for all variables was calculated and presented with the models and max values being reported in Table 6. Values of 10 or higher mean that multicollinearity could be problematic. The highest VIF observed was 1.74, suggesting there may be some multicollinearity present but not enough to cause real problems (Field, 2017).

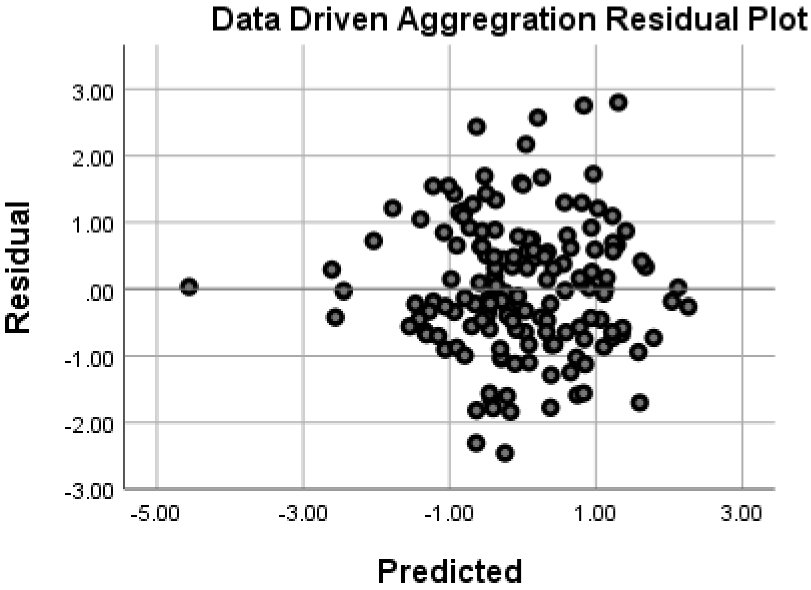
Finally, sometimes cases have a higher influence on the final regression model than they should. To discover this, Cook's distance is calculated, and cases are examined to see if any of them have a greater Cook's distance measure than 1 (Field, 2017). Table 7 reports the max Cook's distance values observed with all models having values well under 1, meaning there should not be influenceable cases, however, Field (2017) notes that many researchers may disagree about what is an influenceable case.

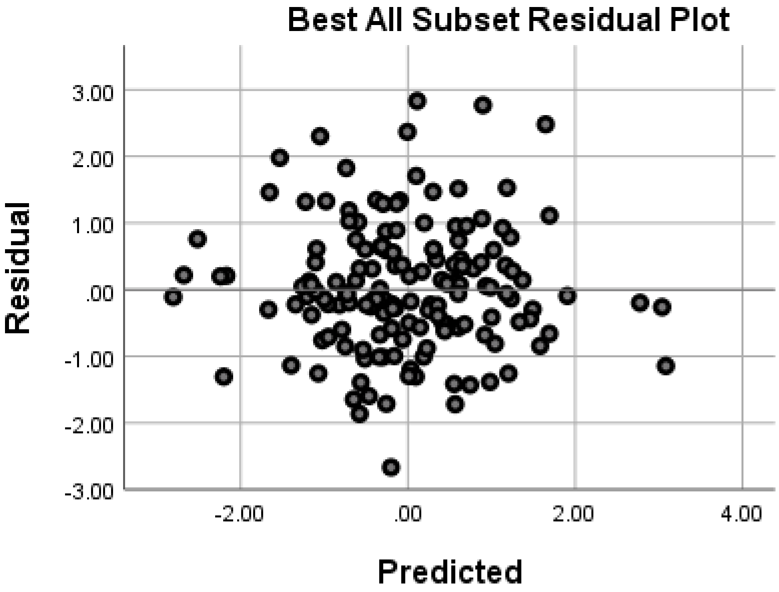
Table 7: Diagnostic Tests

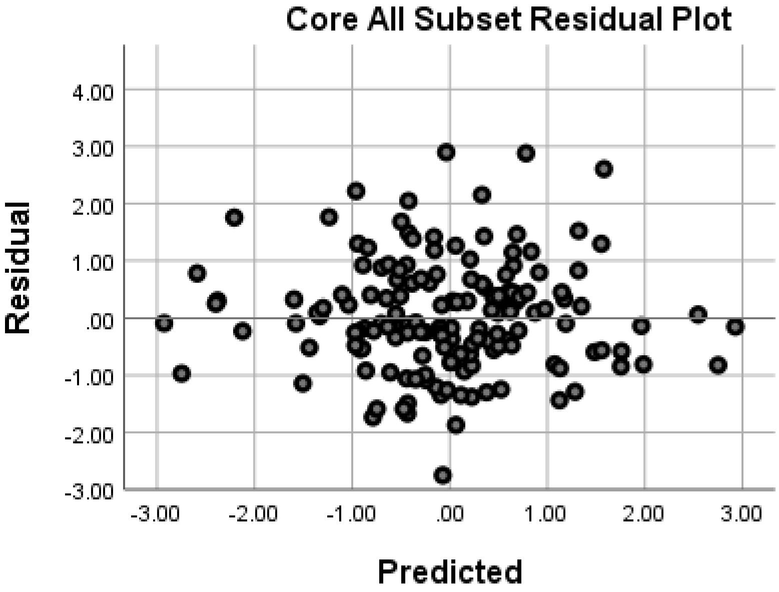
Model	Adj R^2	Max VIF	Max Cook	Outliers > 3 $\|z\|$	Shapiro-Wilks *df* = 154	Breusch-Pagan Test
Full Aggregation	.056	1.604	.057	0	.987	2.114(*df*=3)
Data-Driven Aggregation .01 Level	.152	.900	.062	1	.993	1.248(*df*=2)
Best All Subset Model	.280	1.724	.202	0	.989	5.444(*df*=14)
Core All Subset Model	.269	1.647	.262	1	.990	4.268(*df*=10)

Figure 14: Diagnostic Plots






Best All Subset Residual Plot
Residual
3.00
2.00
1.00
.00
-1.00
-2.00
-3.00
-2.00
.00
2.00
4.00
Predicted


Core All Subset Residual Plot
Residual
4.00
3.00
2.00
1.00
.00
-1.00
-2.00
-3.00
-3.00
-2.00
-1.00
.00
1.00
2.00
3.00
Predicted

Second-Pass Model

To determine the influence of outliers (i.e., the bivariate and residual outliers as well as potentially problematic variables) the data was cleaned and the models re-run. Five cases were removed for being outliers in univariate or bivariate measures and six more were removed for having residuals that were outliers. The variables Task Difficulty, Monitoring Use of Strategies, Monitoring Progress Towards Goals, and Help Seeking Behavior were previously identified as potentially problematic for lack of variation and were removed as well.

The new dataset was used to re-create the models presented previously. The Full Aggregation Model's monitoring macro-level variable was recalculated, whereas planning and strategy use did not change. The Data-Driven Model for the .01 level was created next (Table 8) and every variable left in the data set was used. Finally, the Best Subset Model was re-created. This model dropped to nine variables: Content Evaluation Plus and Minus, Coordinating Information Sources, Drawing, Inference, Memorization, Planning, Recycle Goal in Working Memory, and Summarization. The Core Model consisted of just three variables: Inferences, Drawing, and Coordinating Sources of Information. Changes from the previous Best All Subset models included the removal of Feeling of Knowledge Plus, Interest Minus, Monitoring Progress Towards Goal (removed from study), Planning, Task Difficultly (removed from study), Read Notes, Re-Reading, and Take Notes. Added to the Best Subset Model was Recycle Goal in Working Memory and Summarization. The Core Model saw the removal of Content Evaluation Plus, Feeling of Knowledge Plus, Interest Minus, Memorization, Monitor Progress Towards Goal (removed from study), Planning, and Task Difficulty (removed from study).

Table 8: Variable Inclusion Chart

Variable Name	Positive Macro	Negative Marco	Correlation with Learning Gain	Included in Best Subset Model	Included in Core Model
Coordinating Informational Sources	■		.352	■	■
Inferences	■		.278	■	■
Summarization	■		.224	■	
Re-Reading	■		.180		
Knowledge Elaboration	■		.129		
Judgement of Learning Plus	■		.128		
Feeling of Knowing Plus	■		.117		
Sub-Goal	■		.111		
Prior Knowledge Activation	■		.102		
Content Evaluation Plus	■		.088	■	
Feeling of Knowing Minus	■		.082		
Planning	■		.080	■	
Judgement of Learning Minus	■		.071		
Expectation of Adequacy of Content Plus	■		.071		
Taking Notes	■		.054		
Interest Plus	■		.050		
Time Monitoring	■		.035		
Search		■	-.015		
Expectation of Adequacy of Content Minus		■	-.036		
Read Notes		■	-.043		
Memorization		■	-.071	■	
Content Evaluation Minus		■	-.076	■	
Interest Minus		■	-.089		
Draw		■	-.107	■	■
Recycle Goal in Working Memory		■	-.126	■	

The results of the second-pass analyses are summarized in Table 9. Here, the Full Aggregation Model has an adj R^2 value of .044, a decrease from the .056 it had with the full dataset. A notable change also occurred in the beta weights, with Planning going from a beta weight of -.012 in the full data set to .045 in the second-pass dataset. In addition, monitoring increased from -.243 to .043 in the second-pass dataset. This changed the associated *t*-test from being statistically significant ($p < .01$), to not being statistically significant ($p = .681$). The Data-Driven Model at the .01 level changed from an adj R^2 level of .152 to .138. Beta weights for the positive macro-level variable increased to .418 from .348 and the negative macro-level decreased to -.304 from -.281. The Best All Subset went from 14 variables to 9 and an adj R^2 value of .280 to .221. In the Core Model, the number of variables dropped from 9 to 3 and the adj R^2 dropped from .269 to .186. However, direct comparisons should be done with caution on the Best Subset Model and the Core Model due to differences in variables included.

Table 9: *Data Cleaned Model Results*

Model	Variable	*β*	*t*	*f*	Df(*f*)	Adj R^2
Full Aggregation				3.191*	3,144	.044
	Monitoring	-.412	.681			
	Planning	-.012	-.123			
	Strategy Use	2.838	.005**			
Data-Driven Aggregation (>.01 Level)				12.547***	2,144	.138
	Positive	.418	4.794***			
	Negative	-.304	-3.494***			
Best All Subset				5.544***	9,144	.221
	Content Evaluation Plus	.125	1.530			
	Content Evaluation Minus	-.096	-1.200			
	Coordinating	.303	3.956***			

	Informational Sources					
	Draw	-.213	-2.693**			
	Inferences	.225	2.713**			
	Planning	.105	1.374			
	Recycle Goal in Working Memory	-.124	-1.550			
	Summarization	.119	1.500			
Core All Subset				11.970***	3,144	.186
	Coordinating Informational Sources	.321	4.151***			
	Draw	-.198	-2.573*			
	Inferences	.247	3.149**			

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

The changes in adj R^2 values show the degree to which the removal of 11 cases caused the models to overfit the data to their anomalies (Field, 2017). As would be expected, the Full Aggregation Model dropped the least, changing only -.012. The Data-Driven Model, due to the high number of variables included, also was resistant to large-scale change, dropping only -.014. The Best All Subset Model dropped -.014. This shows was is only slightly less robust to influential cases than the Full Aggregation Model. The Best Subset Model, which has been noted in literature to overfit the data (Hastie et al., 2009), dropped -.059, which is consistent with the idea that this may depend more on extreme cases than the other two methods. This decrease in fit was like the decrease in the fit in the Full Aggregation Model, showing that after removing potential problems from the dataset both had adj R^2 decreases of 21%. The Data-Driven Model seemed to be the most resistant to change, only changing by 10%. The Core Model dropped .083 or 30%. While the original Best Subset Model lost only 4% moving to the Core Model in the full dataset, the move from Best Subset Model to Core Model decreased 15% in the second-pass dataset. This is offset by the Core Model on the second-pass dataset being an extremely parsimonious model, having only three self-regulated learning processes explaining 18.6% of the variability in changes of scores between the pre- and post-test.

Summary

In this thesis, I attempted to answer my research question regarding variable selection method produced the model with the highest adjusted R^2. The best fitting model would be the Best Subset Model that had an adj R^2 value of .280 compared to the Full Aggregation Model's .056 and the best of the three Data-Driven Model's (the .1 level model) .152. The best all-subset method selected the follow variables: Coordinating Information Sources, Draw, Feeling of Knowledge Plus, Inference, Interest Minus, Memorize, Monitor Progress Towards Goal, Planning, Task Difficulty, Content Evaluation Minus, Read Notes, Re-Reading, and Take Notes.

It was noted that a Core Model could be created by looking at the variables that only occurred in all ten of the top fitting models that were produced using all best subset regression. Compared to the Best Subset Model, the Core Model would achieve an adj R^2 value of .269, a loss of only .011. Furthermore, looking at t-tests of the variables in the Best Subset Model, moving from the Best Subset Model to the Core Model dropped many statistically non-significant variables from the Best Subset Model. Of the variables that were non-significant, Task Difficult, and Interest Minus were flagged as potentially having problems.

Regression diagnostics showed no major problems in turns of non-normality of residuals, heterogeneity of variance, problematic multi-collinearity or influential variables, single variate and bi-variate graphs showed that some variables may have been problematic due to having near-zero variance. To determine the effect these would have on the dataset the analyses were rerun with a reduced dataset. Task Difficult, Monitoring Use of Strategies, Monitoring Progress Towards Goals, and Help Seeking Behavior were removed, as were 11 cases, five of which were outliers in univariate or bi-variate measures and 6 of which had high residual value, and the analyses were rerun. Here the results were similar to the original analyses, with lower adj R^2 rates as the artificial inflation caused by outliers was removed. The Best Subset Model consisted of the variables: Content Evaluation Plus and Minus, Coordinating Information Sources, Drawing, Inference, Memorization, Planning, Recycle Goal in Working Memory, and Summarization. The Core Model consisted of just three variables: Inferences, Drawing, and Coordinating Sources of Information. All the variables that lacked significant t-tests for the variables within the Best Subset Model were removed moving to the Core Model. The Best Subset Model adj R^2 decreased from .280 to .221, while the Core model Decreased from .269 to .186.

Conclusion

With these analyses, I examined different ways of constructing models as well as how well they fit the observed data. Data was first summarized and examined for cases that may be problematic during the later analyses, after which the models were created. With the full dataset, the Best Subset fit the best statistically, followed by the Data-Driven Model and then the Full Aggregation Model. It was noted that among the best subset regression models, a core set of processes was used in all 10 best models, so a Core Model was created that preserved the strong fit of the Best Subset Model, while presenting a more parsimonious model. Several cases had outliers, however, and some variables had little variation. A second-pass analysis was performed on a reduced dataset that removed variables that may be problematic due to low variability and removed the outliers found in the data summary and regression diagnostics phases. The fit of all four models decreased but the order of fit remained the same. A Core Model was revealed with three processes: Coordinating Information Sources, Inferences, and Planning.

CHAPTER 5: DISCUSSION

The goal of this thesis was to examine secondary data generated from a self-regulated learning study (Greene et al., 2010) that used micro-level event data with best all subset regression. Many self-regulated learning studies have multiple variables and, in some cases, more variables than participants. This has led to the creation of macro-level variables in the past (Greene & Azevedo, 2009), but these variables do not allow the direct inspection of the effect of the individual self-regulated learning processes on the model. Best all subset regression can deal with datasets with more subjects than variables without experiencing data loss that comes from creating macro-level variables.

My research question was focused upon which model performed the best using adj R^2. The Best Subset Model performed better than the Data-Driven Aggregation model that has been used in the past, as well as the Full Aggregation Methods (Greene et al., 2014, 2015). Because there are often multiple "best" models (Breiman, 2001b), a Core Model was also constructed of the variables that were seen in the 10 best models. Adj R^2 does penalize models as they add variables, nonetheless variables can still be added to model that only slightly increase adj R^2 (Miller, 2002). Using a Core Model was an attempt to control against this.

Although the Core Model fit the data the best statistically and theoretically, some cases in the dataset seemed to exert undue influence on the results and some variables lacked enough variation to be useful. To fully answer my research question, a second pass analysis was performed after removing several problematic cases and variables. In this analysis, it was shown that while all models decreased in terms of adj R^2, the Best Subset Model still had a higher adj R^2 than the Data-Driven Aggregration and Full Aggregration Models. Reducing down to a core model left a model consisting of only three variables: Coordinating Information Sources, Inferences, and Planning. Finally, whereas the Best Subset Model on the cleaned dataset had a slightly higher adj R^2 than the Core Model, a look at the *t*-tests showed Content Evaluation Plus and Minus, Planning, Recycle Goal in Working Memory, and Summarization did not add to the model in terms of statistical significance. When these variables were removed, it reduced the model down to the Core Model. This model appears to be the best in terms of overall statistical fit beyond just adj R^2.

When taken together, best all subset regression appears to be the superior method for analyzing data generated from think-aloud protocol analysis studies, compared to Full or Data-Driven Aggregation. Specifically, this method produces models with higher adj R^2 values with no loss of information that occurs when variables are combined and performs the variable reduction that other studies have sought (Greene, Costa et. al. 2010; Greene et. al., 2015) without a loss of information of evidence or problems in the regression diagnostics.

Theoretical Implications

Examining the best overall model, the two positive predictors (i.e., Coordinating Information Sources and Inferences) are of interest. Examining how these were coded it becomes evident that the codes share a base process (Greene, Costa et al., 2010). Both codes are defined as using multiple pieces of information. Using information from two or more environmental elements, such as personal notes and textual information from the screen or text and a visual diagram would be coded as Coordinating Information Sources. Inferences are defined as using two or more pieces of information to draw a conclusion or make a hypothesis. The ability to use and synthesize multiple pieces of information is one of the core components of digital literacy (Bulger, Mayer, & Metzger, 2014). While Greene, Yu et al. (2014) defined self-regulated learning as a key component of digital literacy, and these results seem to narrow this down for this task and environment, the ability to use multiple pieces of information are the strongest predictors of learning. Greene, Copeland, Deekens, and Yu (2018) also found Inferences and Coordinating Information Sources to be the two most predictive self-regulated learning strategies. In Greene, Bolick, et al., (2015) Inference was found to be predictive of learning gain in a history task and Corroborating Information Sources, a code related to Coordinating Information Sources, was found to be predictive of learning in both a history and a science task. Thus, codes addressing working with multiple pieces of information have been found to be related to learning across studies and across domains. In all studies to date, these strategies of using multiple pieces of information are sub-processes of self-regulated learning strategy use, but it may be prudent to reconceptualize this as their own multiple source use category (Greene, Copeland, Deekins & Freed, 2018) and create a code system to better understand what processes matter the most and to better break up comparing information, combining information, drawing conclusions about multiple pieces of information, or asking questions about multiple pieces of information.

Furthermore, these results indicate that perhaps the current set of micro-level codes could be modeled at different grain sizes. Chi (1997) suggested that after a protocol analysis is finished it should be reperformed at a different grain size to better answer the questions a researcher has. Because two of the three micro level processes in the final reduced core model seem related to a similar higher-level construct and there were problems with other variables at the micro-level, perhaps it would be best to increase the grain level of analysis somewhere between the micro level and the macro level that the full-aggregation method used. This meso-level would benefit from many of the things found in a macro-level analysis (Greene and Azevedo, 2009). Instead of increasing the grain size to the largest size possible, I suggest reducing it to the largest size indicative of a unique self-regulated learning process, guided by the research that defines what areas comprise self-regulated learning theory. Using Pintrich and colleagues' (2000) framework as a base, a proposed set of meso-level processes could be metacognitive knowledge, metacognitive monitoring, metacognitive judgments, planning, strategy use, strategy selection, and volitional control. From this research, I would recommend adding multiple source usage. Most of the current micro-level processes would easily fit into one of these codes, since Azevedo and colleagues used similar framework for the original code lists.

The meso-level of measurement would have several advantages. First, like the macro-level analysis it can account for individual differences in the way students work through a task (Greene & Azevedo, 2009). For instance, it would not matter if a student made a judgment of learning or had a feeling of knowledge as long as they used metacognitive monitoring. Likewise, when a student evaluates a text to determine whether or not it will be helpful to task may not be as important as when they make this metacognitive judgment. The second benefit is because these will be combinations of linear processes, the processes may be more resistant to outliers. Outliers found in several variables in this thesis at the micro-level were not found when combined to the macro-level constructs. It may be found that the meso-level also will help account for extreme variation by taking into account the different ways self-regulated learning processes manifests. For instance, two students may read a paragraph and engage in basic strategy use. One student may take notes, engaging in note taking 10 times. Another may summarize sections as they read, engaging in summarization 10 times. If the average use of note taking and summarization was five each, these two students may be outliers. However, using the meso-level for strategy use, both students would show average levels of strategy use.

Finally, these meso-level codes will enable the study of the sequential nature of self-regulated learning. By moving to a higher level, the meso-level codes become natural targets for process mining, a method that has been used to study the sequential nature of self-regulated learning (Bannert et al, 2015; Sonnenberg & Bannert, 2015).   These studies are situated at a very similar grain level to the meso-level codes proposed here and there is a great deal of overlap between the proposed meso-level codes suggested and the codes already used in process mining research.

Analysis Implications

The results of this study show some of the strengths and weaknesses the Full, Data-Driven, and Best Subset Models. The Full Aggregation Model, although a simple and easy-to-understand model, has problems with the theoretical foundation of how the macros are created. The results of this study show that two micro-level strategies were predictive of learning and one was not. Using the macro-level, if these were the only strategies examined, they would have been added and considered to be positive predictors of learning. Likely due to items with positive and negative correlations with learning being combined, the fit of these models was extremely low. Moving forward to deal with this issue, positive predictors of learning and negative predictors of learning for each macro-level should be separated into different macros for the theoretical constructions (i.e., planning, monitoring, and strategy use). The added value in the application shows which behavior should be encouraged and which should be discouraged.

The Data-Driven Model, although producing better fitting models, has its own issues. It cannot account for variation in the use of a single SRL process when the use of the other SRL processes are held equal, so some included variables may not explain any significantly unique variance. With no penalty for adding variables in the selection part, variables can be added that reduce the adj R^2 values and do not penalize additional variables. Next, the Data-Driven Model masks the contributions of the individual variables in the regression equations. Finally, there is an issue of researcher degrees of freedom (Simmons, Nelson & Simonsohn, 2011). Since the variables selected depend on the correlation threshold for variable inclusion, the results become heavily influenced on the values the researcher selects. I would suggest in the future that the data-driven aggregation be replaced by the Best Subset Model because the Best Subset Model does what this model sought to accomplish without these weaknesses.

The Best Subset Model here seemed to perform better than expected (Miller, 2002; Hastie et al., 2009). Results were produced instantly with modern computing, so the computational complexity is not a factor (Miller, 2002). In addition, by producing a series of top models, core models could be formed that account for the weakness that best all subset regression can produce multiple good models (Berk, 2004). Although it does over fit on outliers (Hastie et al., 2009), the Full and Data-Driven models did as well, as this is a common problem of all ordinary least squares regression methods (Field, 2017). OLS can be biased due to influential cases and variables with near-zero variance, as well. However, removing less than 10% of the sample and four variables seemed to correct for this. While this cleaning of data also introduces researcher degrees of freedom because the researcher determines what to remove and why, there are automatic data preparation solutions in SPSS and most other software that will clean data to predefined thresholds if researcher degrees of freedom are problematic (Kuhn, 2013). Furthermore, least trimmed squares regression will do nearly the same process by removing a set percentage of cases with high residual values (Wilcox, 2002). This will remove outliers and convert many near-zero variance variables into zero variance variables. Although such variables add nothing of value to the study, they will not negatively affect the results. Finally, data cleaning can only do so much. External replication still should be employed when possible (Ioannidis, 2005).

The benefits of this method easily outweigh the potential fit methods. First, it can be used on small sets of data. As the number of predictors in studies has slowly increased, Greene et al. (2018) indicated that having 83 variables analyzed with traditional methods could be extremely problematic. Best all subset regression can easily analyze these datasets. It cannot however give these datasets more statistical power (Field, 2017).

Data Cleaning

Perhaps the most dramatic result of this study was how much change occurred when four variables and 11 cases were removed. This result was not expected, as regression diagnostics and data cleaning have not been frequently discussed in self-regulated learning research. After cleaning the data, the results led to a very simple model with a good degree of fit that aligns with previous research conducted on other datasets and with other tasks. I suggest that moving forward, data be screened for outliers and influential residuals and analyses re-run to see what effect these cases have on the overall results. Likewise, near-zero variance variables should be either removed from the study or discretized to prevent undue influence (Kuhn, 2013).

Data cleaning does have the drawback again of adding researcher degrees of freedom. Extreme data points can be removed to help confirm a hypothesis, and often are (Simmons, et al. 2011). Robust regression can be used to produce similar results to data that was cleaned (Field, 2017), but it does so at the cost of efficacy, as only a portion of the sample will be used in the analysis (Wilcox, 2002). In this study, results were provided before and after data cleaning and problematic points were pointed out in univariate and bivariate graphs of independent variables and residuals. This should be encouraged in future studies to promote full transparency (Simmons et al., 2011). The relationship between the bias produced by problematic points and the statistical power loss from robust regression methods is an area that needs much more research so that researchers can make informed decisions on which methods to use when. Currently, there is a lack of information on when to clean data or when to use robust regression (Van den Broeck, Argeseanu Cunningham, Eeckels, & Herbst, 2005).

Limitations

Although this method of variable selection has its strengths, it also has some drawbacks. Discussed so far were the tendency to overfit data, researcher degrees of freedom in data cleaning, and the knowledge measure having a ceiling effect. There are three areas of limitations for this study: the garden of forking paths, measurement, and analysis issues.

The garden of forking paths. Before discussing the other three areas, to prevent redundancy, the garden of forking paths should be discussed. Previously, the concept of researcher degrees of freedom has been brought up, which are decisions that a researcher can make that change the outcome of a study (Simmons, Nelson & Simonsohn, 2011). The process of researcher choices creating many sets of results are sometimes referred to the garden of forking paths (Gelman & Loken, 2014). Here at each stage, the results split into multiple paths of results based on a researcher choice, creating an exponential increase in solutions as a researcher degrees of freedom increase. In this study, examples of these decisions would be using SPSS to perform the best subset selection procedure forcing the selection to be based on ordinary least squares regression, selecting adj R^2 as the value to maximize around, SPSS offered a limited selection, cleaning the data, defining outliers using bivariate measure and z-scores on the residuals, defining problematic variables by those that lack variation, removing instead of trying to fix problematic cases through winsorizing extreme cases and discretizing the variables, and presenting a core model along with the best all subset regression model. Instead of removing variables with near zero variance, I could have created dummy variables that instead focused simply on whether the participant used the self-regulated process. Finally, rather than removing cases, trimmed least squares regression could have been used. This path I would expect to lead to different results, which would then change the conclusion of this thesis. I tried to present the results from different decision points when possible, but it was not reasonable to present all possible permutations of results that could occur.

This is a serious concern because what appears to be well-made decisions by a researcher can be decisions made post hoc to drive the conclusion of the study towards a certain conclusion. Since the proposal of this thesis included the methods for analysis and stated that this is an exploratory, descriptive study, it was protected from this problem. Nonetheless, peer-reviewed studies do not require preregistration and oftentimes what is exploratory analysis can be presented as if it was confirmatory (Gelman & Loken, 2014).

Although research into the garden of forking paths has focused on unethical research, there is a concern that even well intended, highly justified decisions that are routinely made during a study can alter the results according to a researcher’s bias towards certain analyses and methods. Silberzahn et al. (2018) examined this in an extreme way. They had 29 diverse research teams examine a dataset to determine if a player’s skin tone resulted in receiving more red cards in soccer. It was found that the analyses the researchers used greatly varied in terms of variables, regression type, number of covariates, and the conclusions drawn. Ultimately, two-thirds found some support that skin tone influenced red cards given and one-third did not find any support for this conclusion. These researchers lacked any reason to unethically alter their results, which shows that even in ethical research, results can vary by what decisions are made during analysis (Silberzahn et al., 2018).

Measurement issues. Several issues were found in the measurement aspect of this study that are persistent in other studies as well. The first is a ceiling effect on the post-test scores. As discussed previously, a ceiling effect limits the ability to obtain a true measure of learning. Given the goal of this study was to find what processes contribute to the highest learning gains, this is very problematic. The ceiling effect inhibits the true range of learning gains, meaning participants with high pre-test scores that get a maximum score on the post-test will appear to have learned little. The participants that a learning gain analysis really gives weight to are the ones who started with lower pre-test scores and ended with high or maximum post-test scores.

Second, some of the variables showed near-zero variance. In one case, this led to a small number of cases inflating the correlation between a self-regulated learning process and the learning outcome. Subsequently, these processes would appear in the best all subset models. Variables that contain mostly zeros and less than 10% of the remaining scores higher should be removed from future studies or reconceptualized as discrete variables. When discretized, the variables can be then used as dummy variables for whether the process occurs, which may be more important than how often it occurred in low variance situations (Kuhn, 2013).

Analysis issues. The most serious limitation in this study is the lack of replication. Until this study is replicated these results are limited to a single set of participants, completing a single task in a single learning environment. This is a noted weakness of best all subset regression, and all variable selection methods (Hastie et. al., 2009; Miller, 2002), and a question that a single sample design cannot answer (Hastie et al., 2009). While the data was cleaned to remove influential cases and appeared to be consistent with current theory and past results, it cannot be determined if these results will replicate from the data in this study alone. Replication will be needed using the best all subset method to determine if the data can be generalized to other sets of participants, other learning tasks and other learning environments.

Future Research

Future research should follow four paths, some I have discussed already, and others I have not. These are to employ modern statistical methods, examine resampling methods, examine what cases cause problems and how to fix them and examine the contextual and temporal nature of self-regulated learning.

Modern data analysis. The first path is to adopt a more modern analysis technique found in learning analytics and educational data mining. Traditionally, a high number of variables ranged from what could be done by hand with a calculation machine (Salsburg, 2001) to having simply less variables than subjects (Hattie, et al., 2007), with an optimal solution being a ratio of subjects to predictors that is currently not agreed upon (for a partial list see Miller, 2002). Researchers in other fields, however, routinely work with data sets with thousands of predictors. Stock market analysis involves thousands of stocks to correlate. Microarray analysis often uses tens of thousands of variables. Modern medical imaging can produce 50,000 variables worth of information. Social network analysis using geospatial grids and individual words as variables means the number of variables is unlimited (Fan, Han, & Lui, 2014).

Modern data analysis no longer is limited by sample size as it was in classical analysis. The main features of modern data analysis are cross-validation and multiple methods (Hastie et al., 2007). In cross-validation, the data is broken up by case and one set of data is set aside as training data, while the rest is set aside as testing data. Multiple methods are used on the training data to try to find what can best explain the data, after which the methods are used on the testing sample to see which one best fits novel data (Hastie et al., 2007). I suggest that this is the next direction that event-based measuring of self-regulated learning must go. Some early research has been done in this area (Bannert et al., 2014; Sonnenberg & Barrent, 2015, 2018) but I believe much more research is needed here. This thesis gives the best ordinary least squares solution, which I would use as a comparison measure for these other measures.

Rather than suggest a single method, I would advise using multimethod workflow. Although the belief is that there are few methods to analyze datasets in which there are more predictors than participants, the opposite is true. There are countless modern methods, almost all of which have unique variations. Add to these differences in preparing data, and cross-validation and the garden of forking paths appears again. Since these methods are not like OLS and often produce parameters that are hard to interrupt, comparing studies is becoming progressively harder when modern methods are used. To solve this issue, the use of a multimethod workflow has been proposed that standardizes using several methods, with standardized data preparation and cross-validation (Tsiliki et al., 2015). The R Regression package does just this by applying linear models, generalized linear models, partial least squares models, regularized regression, support vector machines, and random forest regression with the application feature selection and/or elimination methods. This will create a standardized output that can be reported from study to study and cleans the data in a uniform way across studies, greatly reducing researcher degrees of freedom, and allowing much easier comparison of results across studies (Tsiliki et al., 2015). This is in line with Sonnenberg and Barrent (2015) who highlighted the need to run multiple methods on a dataset so comparisons of methods can be made.

Resampling methods. The second area for future research concerns statistical methods for working with small sample sizes. Whereas best all subset regression will produce results on small datasets, it cannot increase statistical power, so the results still run the risk of producing false negative results and limiting the type of analyses that can be done. To solve this problem, resampling methods should be employed (Hastie et al., 2009). Bootstrapping will allow for an approximation of what the data would look like if a larger sample with the same characteristics were drawn. It does this by sampling from the original dataset with replacement to create a larger dataset (Hastie et al., 2009). This is used in many modern regression methods, referred to as bagging, or bootstrap aggregation. With bagging, instead of bootstrapping a single sample, many samples are drawn and analyzed, after which an average of the results of all the models is produced. This method is known to be very effective at reducing overfitting, a problem found in same sample sets (Hastie et al., 2009).

Data cleaning and robust methods. The next areas that should be examined are robust methods and data cleaning. The ordinary least squares methods presently used are not robust to outliers. Methods of regression do exist that can deal with these outliers by using only a portion of the dataset (Wilcox, 2011). Trimmed least squares regression seems to be a promising method. By using a robust method and standardizing a cutoff point, this can allow results to be presented without biasing cases and a low level of researcher degrees of freedom. The other option is to clean the data with full transparency. This does open the garden of forking paths and may encourage unethical results (Gelman & Loken, 2014), but presenting results with outliers will lead to results that cannot be replicated (Ioannidis, 2005). Determining how to best clean this data should be a focus of future research to establish some standard for assessing, dealing with, and reporting problematic cases in datasets that will reduce the number of choices a researcher has to make, while increasing the likelihood of replication.

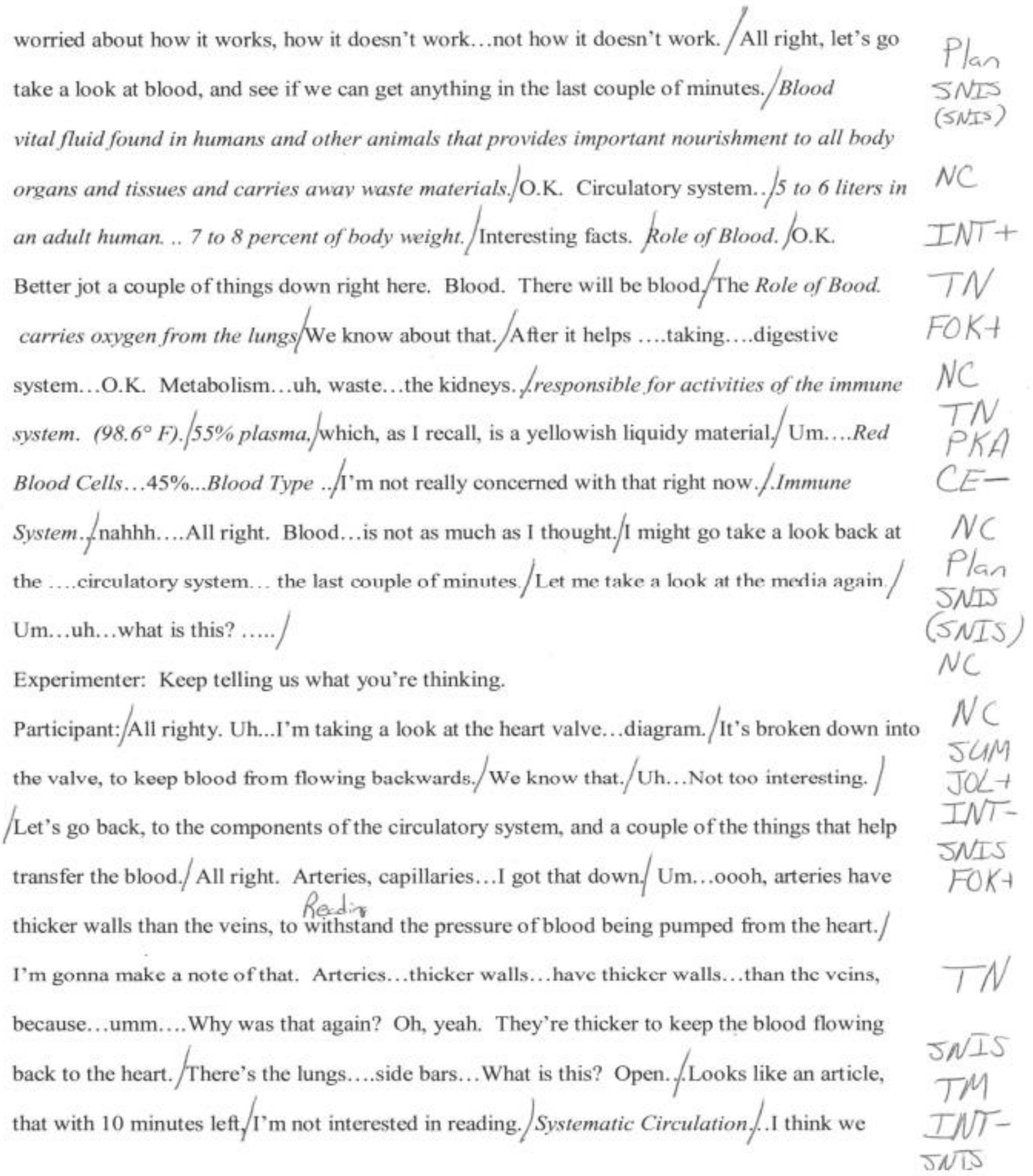
Although cleaning cases improves the models, it does not deal with the issue of what causes these issues. In this dataset, students with negative learning scores appear as outliers in many measures. I tried to examine why some subjects were outliers here, but much more research must be done to determine if these occurrences are part of a larger issue. For instance, would English-as-a-second language learners be unaffected by the increased cognitive demands of a task that involves thinking aloud? What is the effect of doing learning tasks with students who are extremely familiar with the domain? How does attention deficit hyperactivity disorder affect the think-aloud process? These are but three examples of unmeasured variables that could be extreme confounders from a theory standpoint. With this in mind, I believe outliers should be examined separately with a qualitative framework in an attempt to determine their cause.

Temporal and contextual nature of self-regulated learning. Self-regulated learning is said to be a dynamic and conceptual process (Greene & Azevedo, 2010). However, it is often measured as a summary of student actions removed from the context in which they were enacted. Future research using think-aloud protocol analysis should capture information about what elements of the environment were being interacted with when a process was used and information about when in a task the process was used as well as information about the series of processes it was used in. Some studies such as Binbasaran and Greene (2015), Bannert, Reimann, and Sonnenberg (2014) and Sonnenberg and Barrent (2015, 2018) have been doing this already, but much more research is needed.

Conclusion

In this study, I sought to examine different methods in which self-regulated learning processes were identified as the most helpful on average to student learning. Three methods of selecting self-regulated learning processes following a think-aloud protocol analyses were examined. It was found that best all subset regression outperformed the other two variable selection methods without the loss of information that occurs when the other methods are used. It was suggested that this method in the future replace the Data-Driven Aggregation Model and the Full Aggregation Model. It was noted that best all subset regression can sometimes overfit the data, so replications of these findings on a different dataset is vital. Finally based on the results, a way to use this method to compare results across different studies was proposed to answer which self-regulated learning processes are the most important for learning.

APPENDIX 1: CODED PROTOCOL



worried about how it works, how it doesn't work…not how it doesn't work. /All right, let's go Plan
take a look at blood, and see if we can get anything in the last couple of minutes./*Blood* SNIS (SNIS)
vital fluid found in humans and other animals that provides important nourishment to all body
organs and tissues and carries away waste materials./O.K. Circulatory system../*5 to 6 liters in* NC
an adult human. .. 7 to 8 percent of body weight./Interesting facts. /*Role of Blood.* /O.K. INT+
Better jot a couple of things down right here. Blood. There will be blood/The *Role of Bood.* TN
carries oxygen from the lungs/We know about that./After it helps ….taking….digestive FOK+
system…O.K. Metabolism…uh, waste…the kidneys../*responsible for activities of the immune* NC
system. (98.6° F)./*55% plasma,*/which, as I recall, is a yellowish liquidy material/ Um….*Red* TN PKA
Blood Cells…45%...Blood Type ../I'm not really concerned with that right now./.*Immune* CE−
System../nahhh….All right. Blood…is not as much as I thought/I might go take a look back at NC Plan
the ….circulatory system… the last couple of minutes./Let me take a look at the media again./ SNIS (SNIS)
Um…uh…what is this? …../ NC

Experimenter: Keep telling us what you're thinking.

Participant:/All righty. Uh...I'm taking a look at the heart valve…diagram./It's broken down into NC SUM
the valve, to keep blood from flowing backwards./We know that./Uh…Not too interesting. / JOL+ INT−
/Let's go back, to the components of the circulatory system, and a couple of the things that help SNIS
transfer the blood./ All right. Arteries, capillaries…I got that down/ Um…oooh, arteries have FOK+
Reading
thicker walls than the veins, to withstand the pressure of blood being pumped from the heart./
I'm gonna make a note of that. Arteries…thicker walls…have thicker walls…than the veins, TN
because…umm….Why was that again? Oh, yeah. They're thicker to keep the blood flowing
back to the heart./There's the lungs….side bars…What is this? Open../.Looks like an article, SNIS TM
that with 10 minutes left/I'm not interested in reading./*Systematic Circulation*/..I think we INT− SNIS

[1] *From Greene et al. (2010). See Appendix A for a list of codes descriptions. Slash marks separate the segments.*

APPENDIX 2: SELF-REGULATED LEARNING PROCESSES

(based upon Azevedo, Moos, Greene, Winters, & Cromley, 2008)

Macro-Level Process: Planning

Micro-Level Processes	*Description*[1]	*Student Example*
Planning (Plan)	Stating two or more sub-goals simultaneously or stating a sub-goal and combining it with a time requirement.	"First I'll look around to see the structure of environment and then I'll go to specific sections of the circulatory system"
Sub-Goal (SG)	Learner articulates a specific sub-goal that is relevant to the experiment provided overall goal. Must verbalize the goal immediately before taking action.	"I'm looking for something that's going to discuss how things move through the system"
Recycle Goal in Working Memory (RGWM)	Restating the goal (e.g., question or parts of a question) in working memory	"…describe the location and function of the major valves in the heart"

Macro-Level Process: Monitoring

Micro-Level Processes	*Description*	*Student Example*
Content Evaluation (Plus and Minus)[2] (CE+/-)	Monitoring content relative to goals. Learner states content is or is not useful toward reaching the goal.	"I'm reading through the info but it's not specific enough for what I'm looking for"
Expectation of Adequacy of Content (Plus and Minus) (EAC+/-)	Expecting that a certain type of representation will prove either adequate or inadequate given the current goal	"…the video will probably give me the info I need to answer this question" or "I don't think this section on blood pressure will answer my question"

[1] All codes refer to what was recorded in the verbal protocols (i.e., read, seen, or heard in the environment and/or during discussions).

[2] Plus and minus indicates that there are two separate codes. Plus is used when a participant notes the presence of the attribute and minus is used when the participant notes the absence of the attribute i.e., Content Evaluation (-) when the content is deemed not helpful by the participant.

Feeling of Knowing (Plus and Minus) (FOK+/-)	Learner is aware of having read something in the past and having some understanding of it, but not being able to recall it on demand or learner states this is information not seen before	"… I recognize that from the pretest…" or "artherosclerosis – I never heard that word before."
Judgment of Learning (Plus and Minus) (JOL+/-)	Learner makes a statement that they understand what they've read or becomes aware that they don't know or understand everything they read	"I get it" or "I don't know this stuff, it's difficult for me"
Monitor Progress Toward Goals (MPG)	Assessing whether previously-set goal has been met.	"Those were our goals, we accomplished them"
Monitor Use of Strategies (MUS)	Participant comments on how useful a strategy was	"Yeah, drawing it really helped me understand how blood flow throughout the heart"
Time Monitoring (TM)	Participant refers to the number of minutes remaining	"I only have 3 minutes left"
Task Difficulty (TD)	Learner indicates the task is hard or easy.	"This is harder than reading a book."

Macro-Level Process: Strategy Use

Micro-Level Processes	*Description*	*Student Example*
Control Video (CV)	Using pause, start, rewind, or other controls in the digital animation	Clicking pause during the video
Coordinating Informational Sources (COIS)	Coordinating multiple representations, e.g., drawing and notes.	"I'm going to put that [text] with the diagram"
Draw (DRAW)	Making a drawing or diagram to assist in learning	"…I'm trying to imitate the diagram as best as possible"
Inferences (INF)	Making inferences based on what was read, seen, or heard in the hypermedia environment	… [Learner sees the diagram of the heart] and states "so the blood…. through the …then goes from the atrium to the ventricle… and then…"

Knowledge Elaboration (KE)	Elaborating on what was just read, seen, or heard with prior knowledge	[after inspecting a picture of the major valves of the heart] the learner states "so that's how the systemic and pulmonary systems work together"
Memorization (MEM)	Learner tries to memorize text, diagram, etc.	"I'm going to try to memorize this picture"
Prior Knowledge Activation (PKA)	Searching memory for relevant prior knowledge either before beginning performance of a task or during task performance	"It's hard for me to understand, but I vaguely remember learning about the role of blood in high school"
Read Notes (RN)	Reviewing learner's notes.	"Carry blood away. Arteries—away."
Re-reading (RR)	Re-reading or revisiting a section of the hypermedia environment	"I'm reading this again."
Search (SEARCH)	Searching the hypermedia environment with or without the Encarta search feature	"I'm going to type blood pressure in the search box"
Selecting a New Informational Source (SNIS)	The selection and use of various cognitive strategies for memory, learning, reasoning, problem solving, and thinking. May include selecting a new representation, coordinating multiple representations, etc.	[Learner reads about location valves] then switches to watching the video to see their location
Summarization (SUM)	Summarizing what was just read, inspected, or heard in the hypermedia environment	"This says that white blood cells are involved in destroying foreign bodies"
Taking Notes (TN)	Copying text from the hypermedia environment	"I'm going to write that under heart"

Macro-Level Process: Task Difficulty and Demands

Micro-Level Processes	*Description*[2]	*Student Example*
Help Seeking Behavior (HSB)	Learner seeks assistance regarding either the adequateness of their answer or their instructional behavior	"Do you want me to give you a more detailed answer?"

Macro-Level Process: Interest

Micro-Level Processes	*Description*	*Student Example*
Interest Statement (Plus and Minus) (INT+/-)	Learner has a certain level of interest in the task or in the content domain of the task	"Interesting", "This stuff is interesting”

APPENDIX 3: BLANK PRETEST[1]

Pretest
Participant ID: ______________
Date:____________________

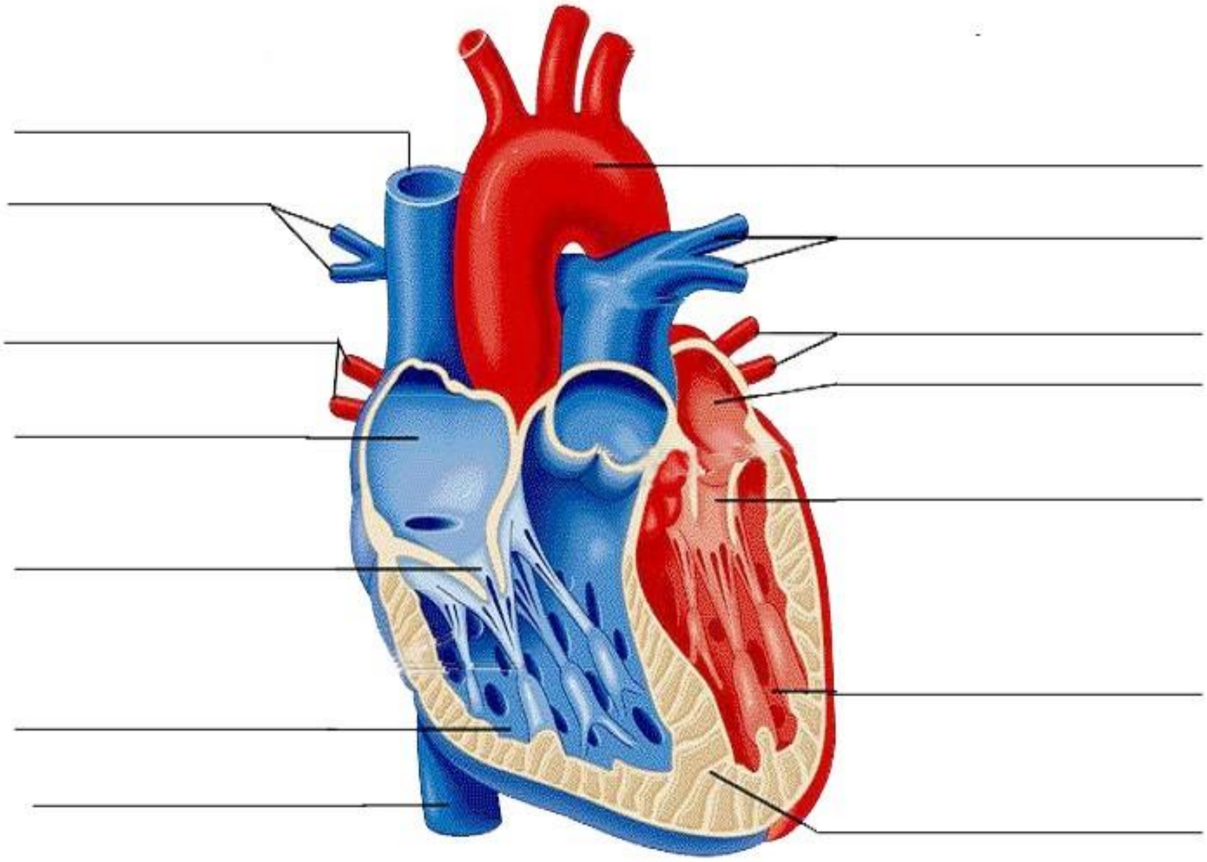
MATCH AS MANY COMPONENTS OF THE HEART AS YOU CAN (13 points)

Valve	A muscular pump that circulates blood throughout the body
Ventricle	The fluid that circulates through the heart and blood vessels
Vein	Pattern of blood flow through the lungs
Heart	The main organ that supplies the blood with oxygen
Lung	A muscular chamber that pumps blood out of the heart
Pulmonary Circulation	A structure which keeps blood from flowing backwards within the circulatory system
Aorta	The impulse-generating tissue located in the right atrium. The normal heartbeat starts here
Atrium	Thin-walled vessel that carries blood back toward the heart
Artery	Smallest blood vessel in the body
Capillary	Largest artery in the body; carries blood from the left ventricle of the heart to the thorax and abdomen
Blood	Thick-walled, elastic vessel that carries blood away from the heart to the arterioles
Pacemaker	Flow of blood from left ventricle through all organs except the lungs
Systemic Circulation	Chamber of the heart that receives blood from veins and pumps it to the ventricle on the same side of the heart

[1] *Modified for presentation.*

Pretest
Participant ID: ______________
Date:____________________

LABEL AS MANY COMPONENTS OF THE HEART AS YOU CAN
(14 in total)



Pretest
Participant ID: ______________
Date:____________________

PLEASE WRITE DOWN EVERYTHING YOU CAN ABOUT THE CIRCULATORY SYSTEM.

Be sure to include all the parts and their purpose, explain how they work both individually and together, and also explain how they contribute to the healthy functioning of the body.

____
____
____
____
____
____
____
____
____
____
____
______Please use the back of this sheet if you need more space….

APPENDIX 4: MENTAL MODELS

Necessary Features for Each Type of Mental Model (From Greene & Azevedo, 2008)

Low Mental Model Category

1. No understanding

2. Basic Global Concepts
blood circulates

3. Global Concepts with Purpose
blood circulates
describes “purpose” - oxygen/nutrient transport

4. Single Loop – Basic
blood circulates
heart as pump
vessels (arteries/veins) transport

5. Single Loop with Purpose
blood circulates
heart as pump
vessels (arteries/veins) transport
describe “purpose” - oxygen/nutrient transport

6. Single Loop - Advanced
blood circulates
heart as pump
vessels (arteries/veins) transport describe “purpose” – oxygen/nutrient transport
mentions one of the following: electrical system, transport functions of blood, details of blood cells

Intermediate Mental Model Category

7. Single Loop with Lungs
blood circulates
heart as pump
vessels (arteries/veins) transport
mentions lungs as a “stop” along the way • describe “purpose” – oxygen/nutrient transport

8. Single Loop with Lungs - Advanced

blood circulates
heart as pump
vessels (arteries/veins) transport
mentions Lungs as a "stop" along the way
describe “purpose” – oxygen/nutrient transport
mentions one of the following: electrical system, transport functions of blood, details of blood cells

<u>High Mental Model Category</u>

9. Double Loop Concept

blood circulates
heart as pump
vessels (arteries/veins) transport
describes “purpose” - oxygen/nutrient transport
mentions separate pulmonary and systemic systems
mentions importance of lungs

10. Double Loop – Basic

blood circulates
heart as pump
vessels (arteries/veins) transport
describe “purpose” - oxygen/nutrient transport • describes loop: heart - body - heart - lungs - heart

11. Double Loop – Detailed

blood circulates
heart as pump
vessels (arteries/veins) transport
describe “purpose” - oxygen/nutrient transport
describes loop: heart - body - heart - lungs – heart
structural details described: names vessels, describes flow through valves

12. Double Loop - Advanced

blood circulates
heart as pump
vessels (arteries/veins) transport
describe “purpose” - oxygen/nutrient transport
describes loop: heart - body - heart - lungs - heart
structural details described: names vessels, describes flow through valves

mentions one of the following: electrical system, transport functions of blood, details of blood cell

REFERENCES

Anderson, E. (1936). The species problem in Iris. *Annals of the Missouri Botanical Garden*, 23 (3), 457–509.

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist*, 40(4), 199–209.

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning,* 4(1), 87–95.

Azevedo, R. (2014). Issues in dealing with sequential and temporal characteristics of self- and socially-regulated learning. *Metacognition and Learning*, 9(2), 217–228.

Azevedo, R., & Cromley, J. G. (2004a). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology*, 96(3), 523–535.

Azevedo, R., & Cromley, J. G. (2004b). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. *Journal of Educational Computing Research*, 30(1–2), 87–111.

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology*, 29(3), 344–370.

Azevedo, R., Cromley, J.G., Winters, F.I., & Moos, D.C. (2004). "Designing adaptive scaffolds in hypermedia to facilitate students? self-regulated learning." Paper presented at Annual Conference of the American Educational Research Association, San Diego, CA.

Azevedo, R., Cromley, J. G., Winters, F., Moos, D., & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science,* 381–412.

Azevedo, R., Cromley, J. G., Winters, F., Moos, D., & Greene, J. A. (2006). Using computers as metacognitive tools to foster students' self-regulated learning. *Technology Instruction Cognition And Learning*, 3(1/2), 97-104

Azevedo, R., Guthrie, J. T., & Seibert, D. (2004). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. *Journal of Educational Computing Research,* 30(1), 87–111.

Azevedo, R., Johnson, A. M., Chauncey, A., & Burkett, C. (2010). Self-regulated learning with metatutor: Advancing the science of learning with metacognitive tools. *New Science of Learning: Computers, Cognition, and Collaboration in Education,* 225–247.

Azevedo, R., Moos, D., Johnson, A. M., & Chauncey, A. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: issues and challenges. *Educational Psychologist,* 45(4), 210–223.

Azevedo, R., Moos, D., Winters, F., Greene, J. A., Cromley, J. G., Olson, E. D., & Chaudhuri, P. (2008). Why is externally-regulated learning more effective than self-regulated learning with hypermedia ? *Educational Technology Research and Development*, 56(1), 45–72.

Azevedo, R., Seibert, D., Guthrie, J. T., Cromley, J. G., Wang, H., & Tron, M. (2002). How do students regulate their learning of complex systems with hypermedia? Paper Presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).

Azevedo, R., Taub, M., & Mudrick, N.V. (2017). Using multi-channel trace data to infer and foster self-regulatedlearning between humans and advanced learning technologies. In Schunk, D & Greene, J.A (Eds.), *Handbook of self regulation of learning and performance (2nd ed.)*. New York,NY: Routledge.

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

Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science*, 40(1), 193–211.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, 9(2), 161–185.

Bannert, M., Sonnenberg, C., Mengelkamp, C., & Pieger, E. (2015). Short-and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. Computers in Human Behavior, 52, 293–306.

Becker, H. J. (2000). Findings from the teaching, learning, and computing survey: Is Larry Cuban right? Henry Jay Becker University of California, Irvine.

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology,* 54, 343–349.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57, 289–300.

Bennett, S., Maton, K., & Kervin, L. (2008). The "digital natives" debate: A critical review of the evidence. *British Journal of Educational Technology*, 39(5), 775–786.

Berk, R. (2004). *Regression analysis: a constructive critique*. Newbury Park: CA: Sage Publications.

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2), 802-837

Berry, W. D. (1993). Understanding regression assumptions. Newbury Park, CA: Sage Publications.

Binbasaran Tuysuzoglu, B., & Greene, J. A. (2014). An investigation of the role of contingent metacognitive behavior in self-regulated learning. *Metacognition and Learning,* 10(1), 77–98.

Boekaerts, M. (1996). Self-regulated learning at the junction of cognition and motivation. *European Psychologist*, 1(2), 100–112.

Bowles, S., & Levin, H. M. The determinants of scholastic achievement. An appraisal of some recent evidence. *Journal of Human Resources*, 1968 (3), 3-24.

Breiman, L. (1995) Better subset selection using the non-negative garrote. *Technometrics*, 37, 738–754.

Breiman, L. (2001a). Random forests. *Journal of Machine Learning*, 45(1), 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–215.

Brown, A. L. (1977). Knowing when, where, and how to remember: a problem of metacognition. In Glaser, R, *Advances in instructional psychology* (pp. 77–165). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin,* 109(2), 204–223.

Bulger, M. E., Mayer, R. E., & Metzger, M. J. (2014). Knowledge and processes that predict proficiency in digital literacy. *Reading and Writing*, 27(9), 1567–1583

Bulman, G., & Fairlie, R. W. (2016). Technology and education: Computers, software, and the internet. In E. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 5, pp. 239–280). North-Holland.

Breusch, T. S.; Pagan, A. R. (1979). "A simple test for heteroskedasticity and random coefficient variation". *Econometrica*. 47 (5): 1287–1294.

Castella, M., Ghosh, M, Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis,* 5(2), 369-411.

Chen, S. Y., Fan, J. P., & MacRedie, R. D. (2006). Navigation in hypermedia learning systems: Experts vs. novices. *Computers in Human Behavior*, 22, 251–266.

Chi, M. (1997). Quantifying qualitative analyses of verbal data: a practical guide. *Journal of the Learning Sciences,* 6(3), 271–315.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research,* 53(4), 445–459.

Clark, R. E., & Sugrue, B. M. (1990). North American disputes about research on learning from media. *International Journal of Educational Research*, 14(6), 507–520.

Cliff, N. (1987). *Analyzing multivariate data.* New York: Harcourt Brace Jovanovich.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences, 3rd ed*. Lawrence Erlbaum Associates Publishers.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, D.C.: U.S. Government Printing Office.

Coley, R. J. J., Cradler, J., & Engel, P. (1997). Computers and classrooms: The status of technology in u.s. schools. *ETS Policy Information Center Report*, 1–67.

Crawley, M. J. (2007). *The R book*. Chichester, England: Wiley.

Credé, M., & Phillips, L. A. (2011). A meta-analytic review of the Motivated Strategies for Learning Questionnaire. *Learning and Individual Differences,* 21(4), 337–346.

Cuban, L. (2001). *Oversold and Underused: Computers in the Classroom*. Cambridge, MA: Harvard University Press.

Cuban, L., & Jandrić, P. (2015). The dubious promise of educational technologies: Historical patterns and future challenges. *E-Learning and Digital Media*, 12(3–4), 425–439.

Dabbagh, N., & Kitsantas, A. (2005). Using web-based pedagogical tools as scaffolds for self-regulated learning. *Instructional Science*, 33(5–6), 513–540.

Darabi, A. A., Nelson, D. W., & Palanki, S. (2007). Acquisition of troubleshooting skills in a computer simulation: Worked example vs. conventional problem solving instructional strategies. *Computers in Human Behavior*, 23(4), 1809–1819.

de Boer, H., Donker-Bergstra, A. S., & Kostons, D. K. (2012). Effective strategies for self-regulated learning: A meta-analysis. *Groningen*: Gronings Instituut voor Onderzoek van Onderwijs; Rijksuniversiteit Groningen.

Deekens, V. M., Greene, J. A., & Lobczowski, N. G. (2017). Monitoring and depth of strategy use in computer-based learning environments for science and history. *British Journal of Educational Psychology*, 88(1), 63-79.

DeMaris, A. (2004). *Regression with social data : modeling continuous and limited response variables.* Wiley-Interscience.

Dent, A. L., & Hoyle, R. H. (2015). A framework for evaluating and enhancing alignment in self-regulated learning research. *Metacognition and Learning*, pp. 165–179. Springer US.

Devolder, A., van Braak, J., & Tondeur, J. (2012). Supporting self regulated learning in computer-based learning environments: Systematic review of effects of scaffolding in the domain of science education. *Journal of Computer Assisted learning*, 28(6), 557–573.

Dignath, C., Buettner, G., & Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively? *Educational Research Review*, 3(2), 101–129.

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, 3(3), 231–264.

Dinsmore, D. L. and Zoellner, B. P. (2018), The relation between cognitive and metacognitive strategic processing during a science simulation. *British Journal of Educational Psycholology*, 88: 95-117.

Donker, A. S. S., de Boer, H., Kostons, D., Dignath, C., & van der Werd, M. P. C. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, 11, 1–26.

Duffy, M., & Azevedo, R. (2015). Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Computers in Human Behavior*, 52(November), 338–348.

Duncan, T. G., & Mckeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational Psychologist*, 40(n2), 117–128.

Efklides, A. (2009). The role of metacognitive experiences in the learning process. *Psicothema*, 21(1), 76–82.

Efron, B (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507), 991–1007.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *The American Psychologist*, 63(7), 591–601.

Ericsson, K. A., & Simon, H. (1993). Protocol analysis: Verbal reports as data (revised edition). Cambridge, Mass: MIT Press.

Ericsson, K. A., & Simon, H. A. (1980). Protocol analysis: Verbal reports as data. *Psychological Review*, 87(3), 443.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review,* 1(2), 293-314.

Fan, J. & Li, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20(1), 101-148.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149-1160.

Faraway, J. J. (2014). Linear models with R (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Field, A. P. (2017). Discovering statistics using IBM SPSS statistics: North American edition. Los Angeles: CA. Sage Publications.

Fieller, N. (2015). Basics of matrix algebra for statistics with R. Boca Raton, FL: Chapman and Hall/CRC.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. 7 (2), 179–188.

Flavell, J. H. (1971). First discussant’s comments: what is memory development the development of? *Human Development*, 14, 272–278.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911.

Fox, John (1991). Regression diagnostics: An introduction. Newbury Park, CA: Sage Publications.

Fox, J & Weisberg, S (2011). An {R} companion to applied regression, second edition. Thousand Oaks CA: Sage

Freedman, D. A. (2009). Statistical models (2nd ed.). Cambridge, MA: Cambridge University Press.

Gardner, D. P. (1983). A Nation at Risk. Washington DC: US Department of Education (ED).

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102, 460–465.

George, E. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452), 1304–1308.

Glass, G.V. Primary, secondary, and meta-analysis of research. *Educational Researcher*. 1976, 5, 3–8.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, 26(3), 499–510.

Greene, J. A., & Azevedo, R. (2005). Adolescents' use of SRL behaviors and their relation to qualitative mental model shifts while using hypermedia. *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, 233–240.

Greene, J. A., & Azevedo, R. (2007a). 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., & Azevedo, R. (2007b). Adolescents' use of self-regulatory processes and their relation to qualitative mental model shifts while using hypermedia. *Journal of Educational Computing Research*, 36(2), 125–148.

Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology*, 34(1), 18–29.

Greene, J. A., & Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist*, 45(4), 203–209.

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., Bolick, C. M., & Robertson, J. (2010). Fostering historical knowledge and thinking skills using hypermedia learning environments: The role of self-regulated learning. *Computers & Education*, 54(1), 230–243.

Greene, J. A., Copeland, D. Z., Deekens, V. M., & Freed, R. (2018). Self-regulated learning processes and multiple source use in and out of school. In J. L. G. Braasch, I. Bråten, & M. T. McCrudden (Eds.). Handbook of Multiple Source Use (pp. 320-338). New York: Routledge.

Greene, J. A., Copeland, D. Z., Deekens, V. M., & Yu, S. B. (2018). Beyond knowledge: Examining digital literacy's role in the acquisition of understanding in science. *Computers & Education*, 117, 141–159.

Greene, J. A., Costa, L., & Dellinger, K. (2011). Analysis of self-regulated learning processing using statistical models for count data. *Metacognition and Learning*, 6(3), 275–301.

Greene, J. A., Costa, L., Robertson, J., Pan, Y., & Deekens, V. M. (2010). Exploring relations among college students' prior knowledge, implicit theories of intelligence, and self-regulated learning in a hypermedia environment. *Computers & Education*, 55(3), 1027–1043.

Greene, J. A., Dellinger, K. R., Tüysüzoğlu, B. B., & Costa, L. (2013). A two-tiered approach to analyzing self-regulated learning data to inform the design of hypermedia learning environments. In R. Azevedo & V. Aleven (Eds.), International handbook of metacognition and learning technologies (pp. 117–128). New York, NY: Springer New York.

Greene, J. A., Moos, D., Azevedo, R., & Winters, F. (2008). Exploring differences between gifted and grade-level students' use of self-regulatory learning processes with hypermedia. *Computers & Education*, 50(3), 1072–1083.

Greene, J. A., Robertson, J., & Costa, L. (2011). Assessing self-regulated learning using think-aloud methods. In B. J. Zimmerman & C. Scharmberg (Eds.), Handbook of self-regulation of learning and performance (pp. 313–328). New York: Routledge: Taylor & Francis Group.

Greene, J. A., Robertson, J., & Costa, L. (2011). Self-Regulated learning using think-aloud methods. In D. H. Schunk & B. J. Zimmerman (Eds.), Handbook of self-regulated learning and preformance (pp. 313–328). New York, NY: Routledge.

Greene, J. A., Yu, S. B., & Copeland, D. Z. (2014). Measuring critical components of digital literacy and their relationships with learning. *Computers and Education,* 76, 55–69.

G'Sell, M., Wager, S., Chouldechova, A. & Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B.*, 78(2), 423-444.

Gunes, F. (2015). Penalized regression methods for linear models in SAS/STAT ®. Cary, NC.

Hastie, T., Tibshirani, R. J., & Friedman, J. H. (2009). The elements of statistical learning. New York, NY: Springer-Verlag New York.

Hattie, J. (2009). Visible learning - Synthesis of over 800 meta-analyses relating to achievement. New York, NY: Routledge.

Hanushek, E. (2016). What matters for achievement: updating coleman on the influence of families and schools. *Education Next*, 16(2), 22-30.

Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: thinking aloud during online searching. *Educational Psychologist*, 39(1), 43–55.

Hoerl, A. E. and Kennard, R. W. (1970b) "Ridge regression: An application to non orthogonal problems", *Technometrics*, 12, 69-82.

IBM Corp (2017). IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8).

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, 2(10).

Kelley, K. & Maxwell, S. E. (2010). Multiple regression. In G. R. Hancock & R. O. Mueller (Eds.) The reviewer's guide to quantitative methods in the social sciences (pp. 281–298). New York, NY: Routledge.

Kirschner, P. A. & van Merriënboer, J. J. G. (2013). Do learners really know best? urban legends in education. *Educational Psychologist*, 48(3), 169–183.

Kramarski, B., & Gutman, M. (2006). How can self-regulated learning be supported in mathematical E-learning environments? *Journal of Computer Assisted Learning*, 22(1), 24–33.

Kulik, C., & Kulik, J. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, 7(1–2), 75–94.

Kulik, J., Kulik, C., & Cohen, P. A. (2010). Effectiveness of Computer-based College Teaching : A Meta-analysis of Findings. *Review of Educational Research*, 50(4), 525–544.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York: Springer.

Lantz, B. (2013). Machine learning with R. Packt Publishing Ltd.

Lichtinger, E., & Kaplan, A. (2015). Employing a case study approach to capture motivation and self-regulation of young students with learning disabilities in authentic educational contexts. *Metacognition and Learning*, 10(1), 119–149.

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

Lockhart, R, Taylor, J, Tibshirani, R & Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, 42(2), 413 – 468.

Lockhart, R, Taylor, J, Tibshirani, R. J. & Tibshirani, R. (2014). Rejoiner: A significance test for the lasso. *Annals of Statistics*, 42(2), 518–531.

Long, J.S. (2001). Regression Models for Categorical Dependent Variables Using STATA. STATA Corporation. College Station, Texas.

Lord, F.M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304-305.

Lumley, T. (2015). Package "leaps": regression subset selection version 2.9.

MacCallum, R., Zhang, S., Preacher, K. & Rucker, D. (2002). On the Practice of Dichotomization of Quantitative Variables. Psychological Methods. 7(1). 19-40

Margaryan, A., Littlejohn, A., & Vojt, G. (2011). Are digital natives a myth or reality? University students' use of digital technologies. *Computers & Education*, 56(2), 429–440.

Margaryan, A., Nicol, D., Littlejohn, A., & Trinder, K. (2008). Students' use of technologies to support formal and informal learning. In World Conference on Educational Multimedia, Hypermedia and Telecommunications 2008 (pp. 4257–4266).

Mayer, R. E. (2001). What good is educational psychology? The case of cognition and instruction. *Educational Psychologist*, 36(2), 83–88.

Mayer, R. E., & Wittrock, R. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), Handbook of educational psychology (2nd ed., pp. 287–304). Mahwah, NJ: Erlbaum.

McCardle, L., & Hadwin, A. F. (2015). Using multiple, contextualized data sources to measure learners perceptions of their self-regulated learning. *Metacognition and Learning*, 10(1), 43–75.

Microsoft Corporation. (2007). Encarta Premium. Redmond, WA.: Microsoft Corporation.

Miller, A. (2002). Subset selection in regression (2nd ed.). CRC Press.

Moos, D., & Miller, A. (2015). The cyclical nature of self-regulated learning phases: stable between learning tasks? *Journal of Cognitive Education and Psychology*, 14(2), 199–219.

Mosteller, F., & Tukey, J. (1977). Data analysis and regression: a second course in statistics. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Reading, Mass: Addison-Wesley.

Muis, K. R., Winne, P. H., & Jamieson-Noel, D. (2007). Using a multitrait-multimethod analysis to examine conceptual similarities of three self-regulated learning inventories. *The British Journal of Educational Psychology*, 77(Pt 1), 177–195.

Müller, N. M., & Seufert, T. (2018). Effects of self-regulation prompts in hypermedia learning on learning performance and self-effica*cy. Learning and Instruction, 58, 1–11.*

Narens, L., Jameson, K. A., & Lee, V. A. (1994). Subthreshold priming and memory monitoring. *Metacognition: Knowing about Knowing*, 71–92.

National Center for Education Statistics. (2015). National Assessment of Educational Progress Data Explorer. Washington, D.C.

Navarro, D. (2013). Learning statistics with R : A tutorial for psychology students and other beginners (Version 0.5).

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.

Norris, C., Sullivan, T., Poirot, J., & Soloway, E. (2003). No access, no use, no impact : snapshot surveys of educational technology in K-12. *Journal of Research on Technology in Education*, 36(1), 15–28.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive Load Theory: Instructional Implications of the Interaction between Information Structures and Cognitive Architecture. *Instructional Science*, 32(1–2), 1–8.

Pietro, A. P., & Murray, S. (2015). U.S. Higher Education Institutions Expected to Spend $6.6 billion on IT In 2015, According to IDC Government Insights. Framingham, MA.

Pintrich, P. R. (2000). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology*, 25(1), 92–104.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (Mslq). *Educational and Psychological Measurement*, 53(3), 801–813.

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, 9(5), 1–6.

Pressley, M., & Harris, K. R. (2006). Cognitive strategies instruction : From basic research to classroom instruction. *The Journal of Education*, 189(1/2), 77–94.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria. https://www.R-project.org

Raschka, S. (2015). Python machine learning. Packt Publishing Ltd.

Ratner, B. (2010). Variable selection methods in regression: Ignorable problem, outing notable solution. *Journal of Targeting, Measurement and Analysis for Marketing*, 18(1), 65–75.

Saks, K., & Leijen, Ä. (2014). Distinguishing self-directed and self-regulated learning and measuring them in the e-learning context. *Procedia - Social and Behavioral Sciences*, 112, 190–198.

Salomon, G. (1998). Novel constructivist learning environments and novel technologies: some issues to be concerned with. *Learning and Instruction*, 8(1), 3–12.

Salsburg, D. (2001). The lady tasting tea: How statistics revolutionized science in the twentieth century. New York: W.H. Freeman.

Scheiter, K., & Gerjets, P. (2007). Learner control in hypermedia environments. *Educational Psychology Review*, 19(3), 285–307. http://doi.org/10.1007/s10648-007-9046-3

Scheiter, K., Scheiter, K., Gerjets, P., Huk, T., Imhof, B., & Kammerer, Y. (2009). The effects of realism in learning with dynamic visualizations. *Learning and Instruction*, 19(6), 481–494.

Schellings, G. L. M., van Hout-Wolters, B., Veenman, M. V. J., & Meijer, J. (2013). Assessing metacognitive activities: the in-depth comparison of a task-specific questionnaire with think-aloud protocols. *European Journal of Psychology of Education*, 28(3), 963–990.

Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., … Woods, J. (2014). The effects of technology use in postsecondary education: A meta-analysis of classroom applications. Computers & Education, 72, 271–291.

Schraw, G. (2006). Knowledge: Structures and processes. In P. A. Alexander & P. H. Winne (Eds.), Handbook of educational psychology (2nd ed., pp. 245–264). Mahwah, NJ: Erlbaum.

Schraw, G. (2010). Measuring self-regulation in computer-based learning environments. *Educational Psychologist*, 45(4), 258–266.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: metacognition as part of a broader perspective on learning. *Research in Science Education*, 36(1), 111–139.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19(4), 460–475.

Schunk, D. H. (2005). Self-regulated learning: the educational legacy of Paul R. Pintrich. *Educational Psychologist*, 40(2), 85–94.

Schwartz, B. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, 1(3), 357–375.

Sclott, J. (2017). Matrix analysis for statistics: 3rd Edition. New Jersey: Hoboken. John Wiley & Sons, Inc.

Selwyn, N. (2009). The digital native: myth and reality. *Aslib Proceedings*, 6(4), 364–379.

Shaffer, J. P., & Saffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46(1), 561–584.

Sheingold, K., Hadley, M., & Thesiar Lieliillan, K. (1990). Accomplished teachers: Integrating computers into classroom practice. New York, NY.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey… Nosek, B. A.(2018). Many analysts, one data set: making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3), 1-20.

Simmons J. P., Nelson L. D., Simonsohn U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*. 22 1359–1366.

Sonnenberg, C., & Bannert, Maria. (2015). Discovering the effects of metacognitive prompts on the sequential structure of srl-processes using process mining techniques. *Journal of Learning Analytics*. 2. 72-100.

Sonnenberg, C., & Bannert, M. (2018). Using Process Mining to examine the sustainability of instructional support: How stable are the effects of metacognitive prompting on self-regulatory behavior? *Computers in Human Behavior*.

Swezller, J., & Sweller, J. (1994). Cognitive load theory, learning difficult, and instructional design. *Learning and Instruction*, 4, 295–312.

Tibshirani, R.(1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

Thompson, B. (1995). "Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial." *Educational and Psychological Measurement*, 55(4), 525-534.

Thompson, B. (2005). Foundations of behavioral statistics an insight-based approach. New York: NY. Guilford Press.

Tsiliki, G.; Munteanu, C.R.; Seoane, J.A.; Fernandez-Lozano, C.; Sarimveis, H.; Willighagen, E.L. Rregrs (2015): An r package for computer-aided model selection with multiple regression models. *Journal of Cheminformatics,* 7, 46.

Uttl, B. (2005). Measurement of individual differences: Lessons from memory assessment in research and clinical practice. *Psychological Science*, 16(6), 460–467

Vandevelde, S., Van Keer, H., Schellings, G. L. M., van Hout-Wolters, B., Keer, H. Van, & Hout-Wolters, B. (2015). Using think-aloud protocol analysis to gain in-depth insights into upper primary school children's self-regulated learning. Learning and Individual Differences, 43(November), 11–30.

Veenman, M. V. J. (2007). The assessment and instruction of self-regulation in computer-based environments: A discussion. *Metacognition and Learning*, 2(2–3), 177–183.

Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instruments: a discussion. *Metacognition and Learning*, 6(2), 205–211.

Veenman, M. V. J., van Hout-Wolters, B., & Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning*, 1(1), 3–14.

Vygotsky, L. S. (1986). Thought and language - Revised Edition. Boston, MA: MIT Press.

Wagner, S., Hastie, T & Efron, B (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1), 1625–1651.

.Wagenmakers, E., Wetzels, W., Borsboom, H., van der Maas, H., Kievet, R (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science,* 7(6), 632–638.

Weinstein, C. E., Schulte, A., & Palmer, D. R. (1987). The Learning and Study Strategies Inventory. Clearwater, FL: H & H Publishing.

Wright, D.B. (2006). Comparing groups in a before-after design: When t test and ANCOVA produce different results. *British Journal of Educational Psychology*, 76, 663– 675.

Westinghouse Learning Corporation and Ohio University. (1969). The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development (Vols. 1 and 2, Report to the Office of Economic Opportunity). Athens.

Williams, Matt N., Grajales, Carlos Alberto Gómez, & Kurkiewicz, Dason (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation,* 18(11).

Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing. Amsterdam: Academic Press.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), Self regulated learning and academic achievement Theoretical perspectives (pp. 153–190). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist,* 45(4), 267–276.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated engagement in learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), Metacognition in Educational Theory and Practice (pp. 277–304). Hillsdale, NJ: Lawrence Erlbaum.

Winne, P. H., & Jamieson-Noel, D. (2003). Self-regulating studying by objectives for learning: Students' reports compared to a model. *Contemporary Educational Psychology*, 28(3), 259–276.

Winne, P. H., Jamieson-Noel, D., & Muis, K. R. (2000). Methodological issues and advances in researching tactics, strategies, and self-regulated learning. In P. R. Pintrich & M. Maehr (Eds.), Advances in motivation and achievement: New directions in measures and methods (pp. 121–155). JAI Press.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 531–566). Orlando, FL: Lawrence Erlbaum Associates.

Winters, F., Greene, J. A., & Costich, C. M. (2008). Self-regulation of learning within computer-based learning environments: A critical analysis. *Educational Psychology Review*, 20(4), 429–444.

Wolters, C., & Pintrich, P. R. (1998). Contextual differences in student motivation and self-regulated learning in mathematics, English, and social studies classrooms. *Instructional Science*, 26(1/2), 27–47.

Yang, H. (2013). The case for being automatic: Introducing the automatic linear modeling (LINEAR) Procedure in SPSS Statistics. *Multiple Linear Regression Viewpoints*, 39(2), 27–37.

Zeidner, M., Boekaerts, M., & Pintrich, P. R. (2005). Self-regulation: Directions and challenges for future research. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of Self-Regulation (pp. 750–768). Burlington: MA: Elsevier.

Zheng, L. (2016). The effectiveness of self-regulated learning scaffolds on academic performance in computer-based learning environments: A meta-analysis *Asia Pacific Education Review,* 17 (2016), 187-202

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

Zimmerman, B. J. (2001). Theories of self-rgulated learning and academic achivement: an overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), Self-regulated learning and academic achievement: Theoretical perspectives. (pp. 1–37).

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166–183.

Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, 48(3), 135–147.

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., Heart, N., & Mellins, R. B. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3), 329–339.

Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, 23(4), 614–628.