ABSTRACT

Title of Dissertation: WRITING TO DISCOVER: ADDING COMPLEXITY TO VIEWS OF WRITING AS AN AGENT OF CHANGE IN UNDERGRADUATES' KNOWLEDGE, INTEREST, CONFIDENCE, AND CALIBRATION

Juliana E. van Meerten, Doctor of Philosophy, 2024

Dissertation directed by: Associate Professor
Donald J. Bolger
Department of Human Development and Quantitative Methodology

The rationale behind the present study consisted of evidence reported to date for Galbraith et al.'s (1999, 2018, 2023) dual-process model of writing, suggesting that not only do writers engage in knowledge transformation, but in the development of new knowledge. An additional aspect of Galbraith et al.'s (2023) work is their proposal and validation of a novel subjective knowledge measure, tailored to those two processes, with potential to be used as a tool for calibration, knowledge activation, and learning. The purpose of the present study was to (a) investigate knowledge development comparing two different writing tasks relative to a comparison task of rereading a text passage, (b) explore patterns in subjective knowledge, confidence, and situational interest ratings throughout engagement with such tasks, (c) examine the predictive power of those ratings for post-intervention knowledge, and (d) compare confidence ratings with evidence of knowledge, that is, calculating calibration scores.

The study used a pretest-posttest repeated measures intervention design, in which 146 undergraduate students, enrolled in human development and psychology research methodology courses, were randomly assigned to experimental or comparison conditions. Students in all conditions started by reading a text on the topic of research design, after which students in the experimental conditions engaged in two writing activities, consisting of a free-write (for both experimental conditions) and either an explanatory or persuasive writing task. Simultaneously, students in the comparison condition reread the initial text twice while being tasked with, first, surface-level strategies and, second, deep-level reading strategies. At least a week after the intervention, students in all conditions completed a transfer test, consisting of an argument writing task. Students rated their subjective knowledge about the topic (using an adapted version of Galbraith et al.'s [2023] instrument), confidence in their knowledge about the topic, and situational interest in the topic at hand multiple times throughout the study. The study occurred in real classrooms, using materials akin to existing course materials, on a topic already part of existing course curricula but not yet covered, which contributed to its high ecological validity.

Exploratory factor analyses indicated that the two subscales of subjective knowledge ratings and the single-item confidence rating needed to be combined into one factor (Subjective Knowledge/Confidence; SKC) and treated as such in all analyses. Further, tests of condition regarding knowledge gains, one of the primary hypotheses, needed to be adjusted because of a failure of randomization between groups that was observed upon analyzing initial between-group equivalence. Despite random assignment to conditions, significant differences between conditions on the primary dependent variable of conceptual knowledge were found at pretest for the comparison (control) group. Because such a difference at pretest would invalidate any causal conclusions drawn from comparisons between the experimental and comparison conditions, further comparisons were made only between the two experimental groups in addressing those research questions that pertained to the effect of condition on changes in knowledge and the subjective factors measured, as well as the predictive value of those subjective factors for post-intervention knowledge levels.

Findings indicated that the writing intervention central to the present study had a positive, significant effect on learning about the topic of research design for students in both experimental conditions (i.e., explanation and persuasion) relative to their pretest knowledge levels. Additionally, students in the persuasion group were significantly better calibrated than students in the rereading group, and SKC ratings at posttest were a significant predictor of transfer-test knowledge scores for both the explanation and the persuasion groups, indicating an improved relationship between confidence and actual knowledge levels.

The findings of this study underscore the importance of providing students with a range of learning strategies, including rereading and writing, to help them acquire knowledge. Educators can use these findings to inform their instructional decisions, recognizing that students' individual needs will vary and that a combination of strategies may be most effective in promoting knowledge development.

WRITING TO DISCOVER: ADDING COMPLEXITY TO VIEWS OF WRITING AS AN AGENT OF CHANGE IN UNDERGRADUATES' KNOWLEDGE, INTEREST, CONFIDENCE, AND CALIBRATION

by

Juliana E. van Meerten

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Associate Professor Donald J. Bolger, Chair
Associate Professor Lucas P. Butler
Professor Daniel L. Dinsmore
Assistant Professor Sarah McGrew
Associate Professor Stephanie Prichard
Professor Min Wang

ACKNOWLEDGMENTS

They say it takes a village to raise a child. I have learned that it certainly takes a village to complete a doctoral program. I am grateful to so many people who have played a role in my doctoral journey over the past six years. Some of them played a bigger role when I was 22-year-old Jules and some became more present when I was approaching becoming Dr. Jules. As I was writing these pages, I realized I could have gone on and on and on, expressing my gratefulness, appreciation, and love to all those people who were and are part of my village. I did my best to keep it contained!

To my 3D-advising team: DJ, Dan, and Diane. Every time I use "I" in this document, I really mean "me with the help of these incredible, knowledgeable scholars who have helped me shape and execute these ideas." Thank you for indulging me in creating a project that was sparked by curiosity and interest and helping me shape it into something I am so proud of. Thank you for motivating me, talking me off ledges, and generally being golden human beings.

To my other committee members, Drs. Luke Butler, Min Wang, Sarah McGrew, and Stephanie Prichard. Thank you for your kindness in providing feedback and propelling me and my study forward. This project would not have been the same without you.

To Pat. You have taught me what scholarship looks like and have forever changed the types of questions I ask. Thank you for inspiring and showing me how to be not just any teacher, but the kind of teacher that creates lasting change in students' lives and brings them joy.

To all the friends who believed in me from the get-go, encouraged me to follow my dreams, and pushed me right until the end. Thank you to Sofie, for being the friend and amazing scholar without whom none of this would have happened. To Inge, Judith, Ria, Adriana, Madelon, Teuntje, and Gabby, for being my dependable, longtime friends who are not surprised by anything I throw at them. To Bri'Ann and Peter, Suzy and Dan, Gerjanne and Jonathan, for being kind and loving friends and for making me your kids' Uncle Jules, Tante 'Anne, and Ome/Tante Bets. Being with them brings me joy and makes me realize how precious life is. To Gala, for being an inspiring fellow writing researcher and the friend who indulged me in countless Zoom writing sessions on those days when my ADHD brain simply would not get started. To Andrew, for being the kind friend who made Thanksgivings and Easters filled with warmth instead of homesickness. To Laurens, for being the amazingly funny friend that makes me laugh like no one else but also the one with the most incredible brain that solved my most mundane R problems. To Jannah, Yuting, Muhammad, Kate, Alexa, Rachel, and Gavkhar, for being my friends and fellow warriors on this hard path that is a Ph.D and for inspiring me to never give up. To Patricia, Ann, Millie, and the rest of the choir, for being the friends that brought me much-needed joy, music, and laughter week in and week out. I will forever cherish those memories.

To my parents, siblings, and other family members (tante C., you know who you are): Thank you for always believing in me, for trusting me to go all the way across the ocean and coming back, and for being there for me no matter what. I love you. Opa en oma's: Zonder jullie was me dit niet gelukt. Bedankt dat jullie er altijd voor me zijn (geweest) en dat ik weet dat er altijd knuffels op me wachten als ik terugkom. Ik houd van jullie.

Eric, my partner in crime. Who would have thought? Thank you for being my friend-till-the-end and for embarking on this crazy adventure with me all those years ago. Our journey is far from over, I just know it. You're unforgettable.

To myself. For crying but persisting, for being anxious but trying, for not knowing how but asking for help, and for not giving up.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In the present study, I explored the potential of discovery writing to deepen undergraduate students' knowledge as well as increase their metacognitive awareness of that knowledge about a topic relevant to one of the courses in which they were enrolled. In doing so, I aimed to add to the literature that is testing Galbraith and colleagues' dual-process model of writing (Galbraith & Baaijen, 2018), which suggests that, during the writing process, writers not only engage in knowledge transformation but also in knowledge *constitution*, thus shifting the focus from the writing product to what happens *during* writing. In this chapter, I will outline the rationale behind the present study, describe the conceptual model that served as its foundation, and introduce literatures relevant to the study.

In many college classrooms, reading and listening are two of the main ways for students to acquire new knowledge, illustrated by the large number of studies focused on these modalities (e.g., Lapp et al., 2008; Rost, 2011). For example, in the undergraduate courses from which I sampled participants for the present study, students were required to complete a minimum of 13 and a maximum of 26 reading assignments throughout a 16-week semester (e.g., reading textbook chapters or journal articles). Additionally, these courses contained a minimum of 13 and an approximate maximum of 26 instances of listening-focused activities (e.g., lecture sessions; UMD HDQM, 2024a, 2024b; UMD PSYC, 2023). College instructors like the ones teaching the students sampled for this study generally assign relevant course readings to be completed before or after attending class and give lectures that correspond to those readings. Subsequently, or interwoven with these activities, students discuss the content and their newly acquired understanding of the topic among themselves and with their instructor. This serves as an illustration of the important role reading and listening play in college classrooms.

Ultimately, many college courses also require one or multiple written products as part of the course assessment, including essays, reflections, reports, or research papers, that demonstrate students' knowledge and level of understanding as a result of taking the course in question. For example, in the undergraduate courses sampled in this study, instructors assigned several smaller writing assignments throughout the semester (e.g., lab reports), one big writing assignment to be finished at the end of the semester (e.g., a research proposal), or a combination of the two (UMD HDQM, 2024a, 2024b; UMD PSYC, 2023). Thus, in college classrooms, writing is often mainly used as *a tool for assessing knowledge,* and the emphasis therefore lies on the writing product.

Consequently, many said college instructors (including myself) have discovered that students struggle with these writing assignments and often lack the necessary strategies to complete them effectively (e.g., Fallahi et al., 2006). As a result, over the past twenty years, numerous interventions have been designed to target specific academic writing abilities, such as synthesizing or summarizing (e.g., Boscolo et al., 2007; Friend, 2001) or argumentation (e.g., Butler & Britt, 2011). In these interventions, writing is treated as an *outcome*, with students' writing samples most often assessed for writing quality using rubrics. If the writing quality (however defined by the researchers) of those samples is higher post-intervention than pre-intervention, the intervention is usually considered a success.

Of course, the vastness of the writing literature indicates that I am not the first instructor or researcher to realize the challenging nature of writing tasks. Over the past decades, dating at least as far back as 1981 (Flower & Hayes), many researchers have come to the same realization and have tried to map the processes at work during writing tasks to enhance our understanding of those tasks, as well as of what may be needed to succeed at them. For example, Flower and Hayes characterized writing as a goal-directed problem-solving activity (Flower & Hayes, 1981;

Hayes, 2012) in which writers develop and solve rhetorical problems. As a result, the writer's understanding of the topic at hand changes (Hayes, 2012). Bereiter and Scardamalia (1986; Scardamalia & Bereiter, 1991) suggested that writing is a knowledge-transforming process that has "epistemic benefits" (1991; p. 179). According to this model, expert writers wrestle with their existing knowledge through solving rhetorical and knowledge-related problems—writing as problem-solving, like Flower and Hayes purported, but involving an interaction between two different kinds of problems instead of solely rhetorical ones. Consequently, both the writer's level of understanding of the topic and their level of writing expertise are theorized to increase. In this model, more novice writers engage in *knowledge telling*, which involves text production but does not involve the same cognitive processes as the expert model and thus does not have the same benefits (Scardamalia et al., 1984; Scardamalia and Bereiter, 1991). Both Flower and Hayes and Bereiter and Scardamalia thus suggested that writing is an important, knowledge-related activity that has a positive effect on the writer's level of understanding about the topic at hand. In a similar vein, the writing sub field of *writing to learn* views the act and process of writing as a way to increase knowledge—to learn (e.g., Klein, 1999; Klein & Boscolo, 2016).

Using the process of writing to advance one's own knowledge of a topic is a task much less frequently encountered in college classrooms (and the writing literature at large). This process, often termed *epistemic writing*, involves writing as a means of developing knowledge and understanding (e.g., Dowst, 1980). That is, the writer develops *new* knowledge *during* writing, which presents itself in the writing product. The model of writing at the heart of the present study is Galbraith's (1999) dual-process model of discovery writing, which goes beyond Bereiter and Scardamalia's (1987) knowledge-transformation process of writing and introduces the process of knowledge *constitution*. Galbraith's extensive framework attempts to capture the multiple dynamic ways that writing transforms knowledge, drawing from the vast literature on conceptual learning in cognition, including one's perceptions of their own knowledge

In what follows, I present an overview of this framework, before introducing literatures relevant to the variables at the center of the present study, described in more detail in Chapter 2. I conclude this first chapter by presenting a brief outline of the study and the research questions it addressed.

Conceptual Model

The conceptual model that guided this investigation is largely based on David Galbraith's dual-process model of writing (Galbraith, 1999; Galbraith & Baaijen, 2018). In the development of this model, Galbraith built on existing, longstanding theories and frameworks of writing, such as Flower and Hayes' (1981) model of writing as a goal-directed problem-solving activity and Bereiter and Scardamalia's (1987) model of knowledge-telling and knowledge-transforming processes that happen during writing. He subsequently went beyond these to create a model for writing and discovery through writing. In this model, he distinguished between two main processes that occur during content generation: knowledge transformation and knowledge constitution. During knowledge *transformation* or organization of thought, the writer pulls existing ideas or schemas from episodic memory and, through maintaining and handling those ideas in working memory, generates content to satisfy rhetorical goals. The process of knowledge *constitution*, which interacts with knowledge transformation, goes beyond the use of episodic memory and involves spontaneous formulation of thought guided by the implicit structure of semantic memory (Galbraith & Baaijen, 2018). According to Galbraith, this latter process not only demonstrates existing levels of knowledge but *develops* the writer's understanding of the topic at hand. In other words, the writer may *discover* new ideas and develop new levels of understanding during the process of writing, as opposed to drawing exclusively on pre-existing knowledge. A schematic overview of this process can be seen in Figure 1.

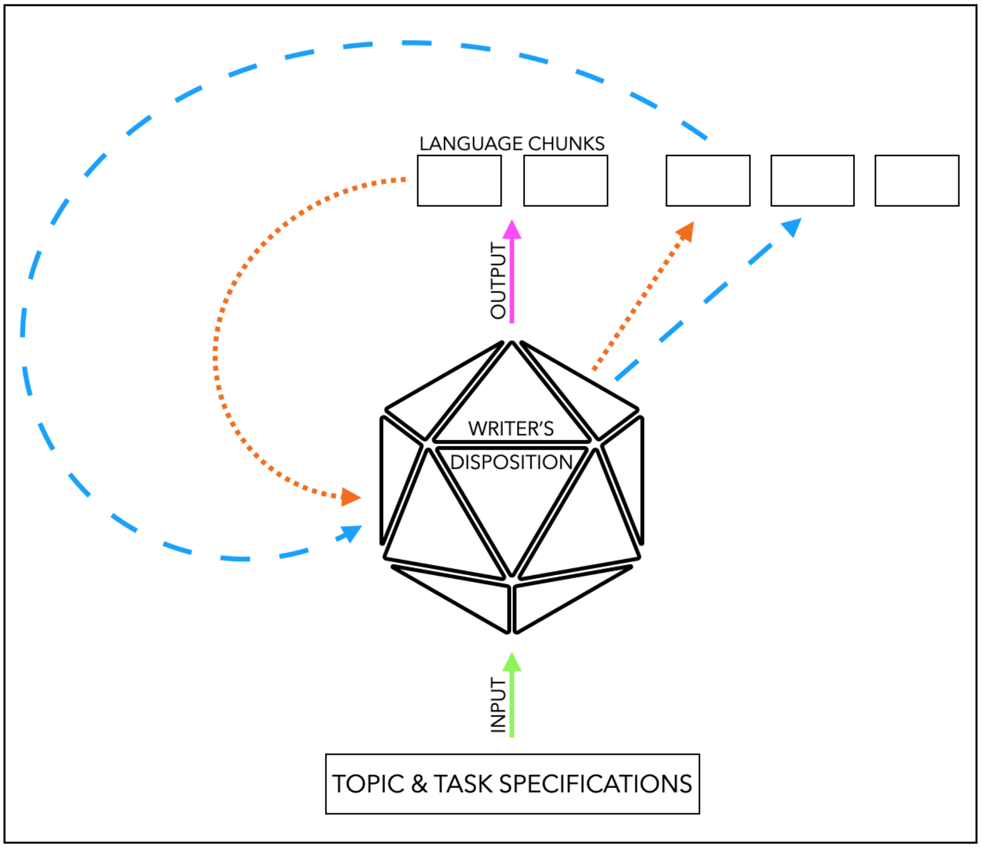
First, the topic and task specifications serve as input and are fed into the writer's disposition toward the topic, which Galbraith and Baaijen (2018) defined as follows:

> "[This disposition] is formally defined as the matrix of connection strengths between the units within a constraint satisfaction network. The strengths of these fixed connections are the product of an individual's learning history and reflect the totality of an individual's experience. They are the connection strengths consistent with the production of all the propositions that the individual has encoded and *constitute the writer's internally organized, but implicit, understanding of the topic*" (p. 242, *emphasis* added).

This description beautifully captures the complexity of any writer's knowledge base and the idiosyncratic nature of this network of understanding and associations, and, as I will argue in more detail in Chapter 2, provides a strong rationale for including students' subjective self-ratings of their knowledge on a given topic in any knowledge-focused writing activity.

Figure 1

Schematic Representation of the Knowledge-Constituting Process



Note. Figure generated based on figures published by Galbraith (1999) and Galbraith & Baaijen (2018).

Second, the topic and task specifications activate connections within this dispositional network, activation that then produces language output that is written down as language chunks. Subsequently, these chunks are fed *back into* the network and, with inhibitory feedback, lead to more new language production. This process is repeated "until the written text captures the writer's understanding" (Galbraith & Baaijen, 2018, p. 244).

Although the dual-process model paints an already rather comprehensive picture of the discovery writing process, it may not entirely do justice to an important aspect of human endeavors, their motivation and affective states. Thus, I will include a measure of interest as an explicit representative motivational variable, described in more detail in subsequent sections.

Self-Ratings of Confidence and Knowledge

One of the goals of the present study was to emphasize the importance of listening to what *learners* think and know. This involves not only focusing on objective measures of knowledge and learning, but choosing to include what learners think they know and investigating how this changes as a function of different knowledge-focused tasks, something that Galbraith and colleagues also stressed in their development of the dual-process model of writing (Galbraith et al., 2023). One way to get to the core of the learning that happens during and as a consequence of writing is to include a measure of subjective understanding. Galbraith and colleagues' contribution to this notion is their development of an 8-item subjective knowledge rating (SKR) instrument, consisting of two subscales that correspond to the two processes present in their model of writing: subjective understanding and subjective organization (described in more detail in subsequent chapters). This instrument has been tested and validated as a reliable measure of subjective understanding and organization (Galbraith et al., 2023) and serves as a crucial component of the present study's design for several reasons.

First, this instrument provides more in-depth information about learners' subjective knowledge states than typical confidence ratings or judgments of learning, which usually consist of one or two items, such as those used in a study by Persky and Dinsmore (2019), who asked participants to "rate their level of confidence (0 through 100) on the correctness of their answers." Frequently, these ratings are then compared to measures of actual performance to calculate learners' *calibration*. In the case of Persky and Dinsmore (2019), for example, students' confidence ratings were compared to their performance on knowledge assessments.

These ratings and subsequent calculations provide a general indication of where learners may be in their knowledge or performance, but often do not contain enough detail for more generalizable conclusions about or guidance specifically tailored to their subjective knowledge states.

The same holds true for the structure of knowledge measures often used in studies investigating learning and knowledge development. Many of these knowledge measures consist of multiple-choice items (e.g., Buchin & Mulligan, 2023; Dinsmore & Alexander, 2016) or, at most, short-answer questions (e.g., List, 2018), the answers to which leave much room for interpretation and often do not paint a detailed enough picture of the learner's knowledge to serve as the basis for substantial conclusions or directions for how to support the learners.

The fact that the SKR instrument distinguishes between two knowledge-related processes that occur during writing and, for each of these, goes beyond single-item ratings, creates potential for the use of this instrument in writing- and knowledge-related contexts and provides a new avenue for measuring students' calibration of their learning.

Knowledge-Focused Writing Intervention Studies

Given the present study's focus on writing and subjective knowledge, it is important to consider writing interventions that include knowledge as an outcome. An established literature that fits this description is *writing to learn*, which has, at its core, writing as a tool to increase and deepen knowledge (for a review and meta-analysis, see Graham et al., 2020). Using writing as a treatment for learning outcomes has been shown to be effective in many academic domains, such as mathematics (e.g., Jitendra et al., 2013), social studies (e.g., Morphy, 2013), and science (e.g., Kingir et al., 2012), although variability of effect sizes exists (Bangert-Drowns et al., 2004; Graham et al., 2020) suggesting that factors such as the type of task play a pertinent role.

Many of these writing-to-learn studies use Flower and Hayes' model of writing as problem-solving or Bereiter and Scardamalia's model of knowledge transformation versus knowledge telling as their main theoretical framing. After all, these models posit that the process of problem-solving as well as the process of knowledge transformation during writing have a positive effect on the writer's level of understanding and writing ability (Hayes, 2012; Bereiter & Scardamalia, 1987, respectively), thus providing a strong rationale for using the act of writing to that end. Something that is not as explicitly present in this literature and its experimental studies is attention to what happens *during* the writing process and the knowledge evident and developing in the writing product(s). This is where discovery writing can play an important complementary role.

Like many writing-to-learn studies, Galbraith's (e.g., 1992; 1999; Galbraith et al., 2005; Galbraith, 2009) empirical studies on discovery writing built on Bereiter and Scardamalia's (1987) model of knowledge transformation and, as established in previous sections, add to this the process of knowledge constitution. These studies served to test and solidify the dual-process model of writing and have demonstrated the potential of the two processes to develop and increase knowledge and learning . Ultimately, these studies came together in Galbraith and colleagues' (2023) most recent work on developing an instrument to measure subjective knowledge, capturing the two processes present in their model and providing further evidence for its validity.

Knowledge and Interest as Outcomes

Two final constructs lie at the heart of the present study and need proper introduction: knowledge and interest. Both constructs exist in long-standing (e.g., Dewey, 1913), well-established literatures, and the goal of the present study is, therefore, not to review these or provide new theoretical insights or definitions that add to them. Rather, this section outlines why these constructs are important, the definitions I have chosen for the purposes of the present study, and how these constructs were measured (although more detailed descriptions of the measures follow in Chapter 3).

Knowledge

The dominant model of learning, constructivist learning theory, articulates that making sense or meaning of new information depends on the propensity to map this information to what was *known* as the learner began the learning process (e.g., Ausubel, 1968; Alexander & Murphy, 1998; Chi, 1985; Kintsch, 1998). Thus, the breadth and depth of one's existing knowledge is inherently important to any learning-related task or study.

In order to capture the transformation of knowledge, it is critical to measure relevant knowledge prior to the introduction of new information or targeted learning and continue to measure knowledge development during and after the learning process. To that end, I measured knowledge at the start of the present study by tasking students with creating a concept map, asking them to map everything they knew about the topic at hand. This is a task that has been reliably used to measure knowledge in previous studies (e.g., Watson et al., 2016; Rittle-Johnson et al., 2001). Using Novak and Gowin's (1984) scoring method (i.e., counting the number of concepts, highest level of hierarchy, and number of cross-links; Watson et al., 2015), I assessed and scored the knowledge represented in these concept maps, the result of which served as the knowledge indicator at the start of the intervention and right after the intervention was completed (i.e., as posttest).

To measure the application of acquired knowledge in a different context, I assessed writing samples that students completed at least a week after the intervention, using the Structure of Learning Outcomes (SOLO) Taxonomy (Biggs & Collis, 1982). This taxonomy provides a framework for assessing knowledge structures, with an underlying assumption that the more structured and principled is someone's knowledge, the more they know or have learned (e.g., Biggs & Collis, 1989). The resulting scores of the writing samples served as indicators of how well acquired knowledge transferred to a new context. More details about the pretest, posttest, and transfer test as well as the corresponding scoring protocols can be found in Chapter 3.

Interest

One long-standing argument states that conceptual change does not happen outside of affective factors such as motivation or attitude toward the topic (e.g., Pintrich, 2003; Sinatra, 2005). A major motivational variable, interest (e.g., Renninger et al., 2014) impacts levels of attention to and engagement with a task (e.g., Renninger & Hidi, 2017), effort put into said task (e.g., Renninger & Hidi, 2002) as well as levels of learning (e.g., Alexander, 1997; Rotgans & Schmidt, 2011). All of these are important for and relevant to cognitively demanding tasks such as writing (e.g., Hayes, 2012). Using Hidi and Renninger's four-phase model (2006) as a basis, I identified that a specific kind of topic interest, *situational interest*, would be relevant to the present study and would need to be measured. Situational interest is described as attention and an affective reaction that is aroused in the moment by the topic or task and the way it is presented (e.g., Hidi, 1990), as opposed to topic interest at the individual interest level, which represents a more enduring predisposition in a person to come back to a particular topic. Due to the relatively brief nature of the present study, considering such sustained interest is not as relevant (Hidi & Renninger, 2006). In the vast literature on interest, the construct is often reliably measured by single- or multiple-item self-ratings of level of interest (e.g., Ainley & Patrick, 2006;

Linnenbrink-Garcia et al., 2010). Therefore, to measure levels of situational interest, I used similar interest ratings in the present study.

The Present Study

In sum, the rationale behind the present study consisted of evidence reported to date for Galbraith et al.'s (1999, 2018, 2023) dual-process model of writing—suggesting that not only do writers engage in knowledge transformation, but in the development of *new* knowledge—and a novel subjective knowledge measure, tailored to those two processes, with incredible potential to be used as a tool for calibration, knowledge activation, and learning. Additionally, Galbraith and colleagues (2023) called for research exploring the differential effect of writing genres on learning-related outcomes, which I explain in more detail in Chapter 2.

Purpose

The purpose of the present study was to (a) investigate knowledge development comparing two different writing tasks relative to a comparison task of rereading a text passage, (b) explore patterns in subjective knowledge, confidence, and situational interest ratings throughout engagement with such tasks, (c) examine the predictive power of those ratings for post-intervention knowledge, and (d) compare confidence ratings with evidence of knowledge (i.e., calculating calibration scores).

Design

The present study used a *pretest-posttest repeated measures intervention* design, in which undergraduate students were randomly assigned to experimental or comparison conditions. Students in all conditions started by reading a text on the topic at hand, after which students in the experimental conditions engaged in several writing activities. Simultaneously, students in the comparison condition reread the same text while being tasked with surface- and deep-level reading strategies. All students rated their subjective knowledge about the topic, confidence in their knowledge about the topic, and situational interest in the topic at hand multiple times throughout the study. The study occurred in real classrooms, using materials akin to existing course materials, on a topic already part of existing course curricula.

Research Questions & Predictions

The following research questions guided the present study:

1. How do students' subjective knowledge ratings (SKRs), confidence ratings, interest ratings, and knowledge scores change as a function of writing as compared to rereading?
2. How do subjective knowledge and confidence ratings compare to measures of knowledge?
    a. Specifically, how do students' SKRs and confidence ratings compare to knowledge scores of the pre-posttest concept maps and knowledge structures evident in the posttest writing sample?
3. Do SKRs, confidence ratings, and situational interest ratings predict students' post-intervention knowledge scores by experimental condition?

Because of the novelty of the present study generally as well as the subjective knowledge instrument that played a key role in the intervention, the research questions were largely exploratory in nature. I therefore formulated some general predictions rather than proposing formal hypotheses.

Based on the (limited) evidence from the literature, I predicted that students in the experimental conditions would (a) report lower subjective knowledge and confidence ratings after the first writing task than students in the comparison condition after their first reading task, but (b) would become better calibrated throughout the intervention (i.e., their subjective knowledge and confidence ratings would more accurately match their evidence of knowledge), and (c) would show evidence of transformational knowledge gains throughout the intervention relative to the comparison condition.

CHAPTER 2: REVIEW OF THE RELEVANT LITERATURE

In this chapter, I review the literature relevant to the research design and the rationale behind the research questions. I first explore the body of literature that lies at the heart of this study, *discovery writing*. Specifically, I summarize literature on writing as an *outcome* versus writing as a *process*, with the latter most relevant to discovery writing. Then, I present literature on different *outcomes of discovery writing*, focusing on those outcomes relevant to the present study: subjective knowledge, confidence and calibration, interest, and learning and accumulated knowledge. Finally, I revisit the research questions and describe the context of the current study before presenting the methods of the study in Chapter 3.

Discovery Writing

The vast theoretical and empirical literature on writing can be categorized in many different ways. For example, many scholars distinguish between cognitive (e.g., Flower & Hayes, 1981) and more social/sociocultural (e.g., Faigley, 1986; Graham, 2018) models and theories of writing (e.g., Rijlaarsdam & van den Bergh, 2006). The categorization that seems most appropriate and relevant for the present study is a distinction between a body of literature that treats writing as a *product* versus a literature that treats writing as a *process*. This distinction is intended to provide the background and rationale for the research questions and design.

Writing as an Outcome

As I mentioned in Chapter 1, many researchers and practitioners have realized that undergraduate students often struggle with (academic) writing. As a consequence, over the past 20 years, there has been a steady increase in the number of writing interventions that target specific academic writing genres and abilities, such as synthesizing or summarizing (e.g., Boscolo et al., 2007; Friend, 2001), and argumentation (e.g., Butler & Britt, 2011). These writing interventions treat writing as a product or outcome: the goal is to find and test empirically a treatment that is effective for improving undergraduate students' ability to write for the particular genre or task.

In a systematic review of classroom-based academic writing interventions in higher education (van Meerten & Alexander, 2023), my co-author and I found that researchers have used a range of different treatments to achieve this goal. The majority of studies we analyzed relied on *explicit strategy instruction*, which involved teaching students about the use of specific strategies that were designed to help them improve their writing in a specific genre. For example, Friend (2001) taught her participants argument repetition and generalization strategies to uncover what a text's authors deemed most important and thus aid in students' summary writing. Another treatment used in these interventions, often in combination with explicit strategy instruction, is *practice*, in which students receive the opportunity to engage with a strategy or specific writing task multiple times before being re-tested on their writing ability (e.g., Kwok, 2013).

Finally, a number of interventions relied on *observational learning* or *worked examples* (i.e., someone modeled the targeted writing ability or students received a good or bad example of the writing task, respectively) or on providing students with *feedback* throughout the intervention, often in combination with explicit strategy or practice. An example in which researchers employed explicit strategy instruction, practice, and observational learning together can be seen in the study by van der Loo et al. (2018), in which, depending on the conditions to which the participants were assigned, they were provided with (a) videos in which an individual demonstrated the targeted writing skill with accompanying explanation—one model that used less effective and one model that used more effective strategies; (b) sessions in which they had to complete exercises, leading up to the writing task at hand—the authors called this "learning by doing;" and (c) index cards that contained writing strategies for the task, mirroring the strategies used by the stronger model in the videos.

In addition to the body of literature that treats writing as an outcome, there are many college classrooms that operate similarly. For instance, university courses are often designed in such a way that students spend the majority of the semester reading, listening to, and studying content, and subsequently having to complete a writing task to demonstrate acquired knowledge (e.g., a final essay or report; Alexander et al., 2023).

Although treating writing as a product can be useful, the purpose of the present study is to demonstrate the potential of writing as a *process,* which may, additionally and perhaps inadvertently, contribute to the goals of the literature just described (i.e., improving writing ability or demonstrating acquired knowledge).

Writing as a Process

As briefly described in Chapter 1, two models of writing underlie the current study and its conceptual framework and therefore warrant more in-depth description and explanation: writing as goal-directed problem solving (e.g., Flower & Hayes, 1981) and writing as knowledge transformation (e.g., Scardamalia & Bereiter, 1987).

Writing as Goal-Directed Problem Solving

From an earlier history of theorizing and researching writing solely in terms of the outcome or written text, composition researchers and instructors began to focus on the process of writing. Beginning with Emig (1971) and Elbow (1973), who emphasized the importance of the *processes* in which writers engage when writing, a breakthrough occurred in 1979, when Hayes & Flower proposed a radically different way of approaching both the research and practice of writing.

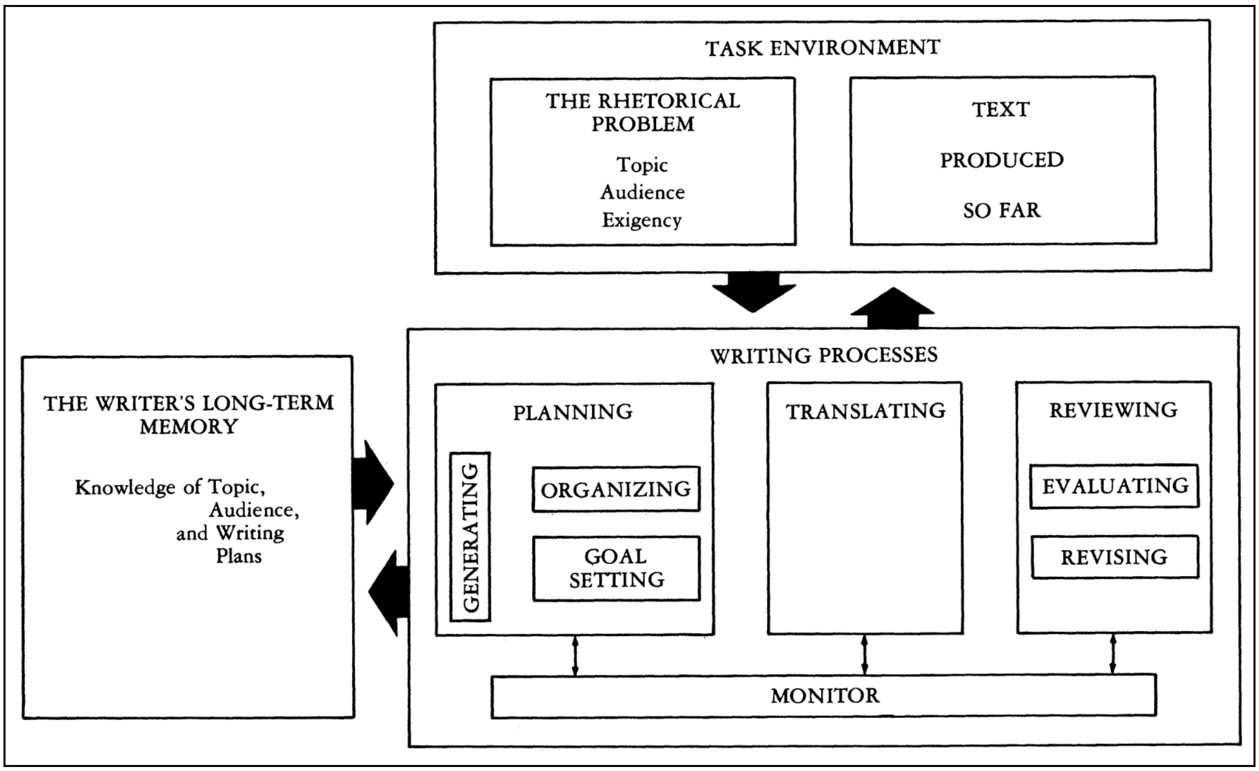
In their original model, Flower and Hayes posited that the writing process consisted of several distinct thinking processes that are being organized by the writer— planning, translating, and reviewing, under the control of a monitoring process (Flower & Hayes, 1981; Hayes & Flower, 1979, 1986; see Figure 2). *Planning* involves generating and organizing ideas and the setting of goals that can evolve throughout the writing task. While writing, writers create subgoals that support the overall purpose of the task but they can also generate new goals based on what they are learning through composing. *Translating* refers to “the process of putting ideas into visible language” (Flower & Hayes, 1981, p. 373). *Reviewing* consists of evaluating and revising on-going ideas and written text. Simultaneously (i.e., while all three processes are happening), the writer monitors progress and keeps track of when they should move on to a different process, such as moving from planning to translating; these decisions depend on the writer’s goals and their writing tendencies or strategies (Flower & Hayes, 1981).

Finally, the model depicts how any given writing task contains a *rhetorical problem* consisting of the task at hand, the target audience, and the writer’s own goals, which can be *solved* through the act of writing (Flower & Hayes, 1981; Hayes & Flower, 1986).

The characteristics of this rhetorical problem place certain constraints on what the writer can produce. For example, if the topic of the writing task is the solution to an engineering problem, addressing a professional audience, and the writer needs to receive an A to pass the class, they simply cannot produce a personal letter to, say, their best friend and meet all of these demands. A second source of constraints arises from the written text that evolves as the writer moves through the task. That is, the title and first few sentences exert influence on what the writer can produce next. The extent of this effect depends on various factors, including the writer's writing experience, their long-term memory capacity, and their topic knowledge (Flower & Hayes, 1981).

Figure 2

Cognitive Process Model of Writing by Flower & Hayes (1981)



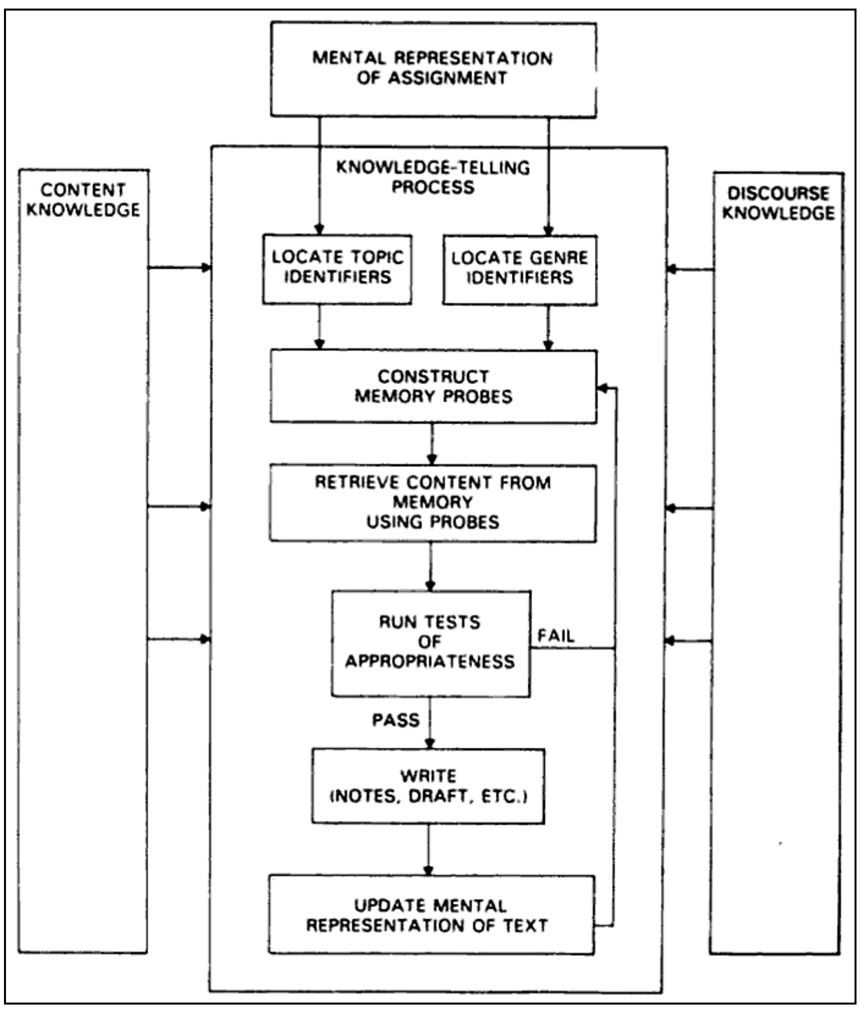
Because the purpose of this section was to lay the groundwork for understanding the present study's conceptual framing and rationale, I will not go into detail about the critiques (e.g., Faigley, 1986), rebuttals (e.g., Flower, 1989), and reformulations (e.g., Hayes, 2012) of Flower and Hayes' work. For the purposes of the present study, their work served as an initially groundbreaking way to describe writing that was distinct from what scholars and practitioners had done until then.

Writing as Knowledge Transformation

A second giant in the world of modeling writing processes, building in part on Flower and Hayes (1981), is Bereiter and Scardamalia's (1987) model. Bereiter and Scardamalia posited that, during writing, more mature writers engage in *knowledge transformation* and more novice writers engage in *knowledge telling* (e.g., Scardamalia et al., 1984; Bereiter & Scardamalia, 1987). *Knowledge telling* refers to writing down any ideas related to the topic and task that come to mind, in a manner that resembles oral discourse production (see Figure 3).

Figure 3

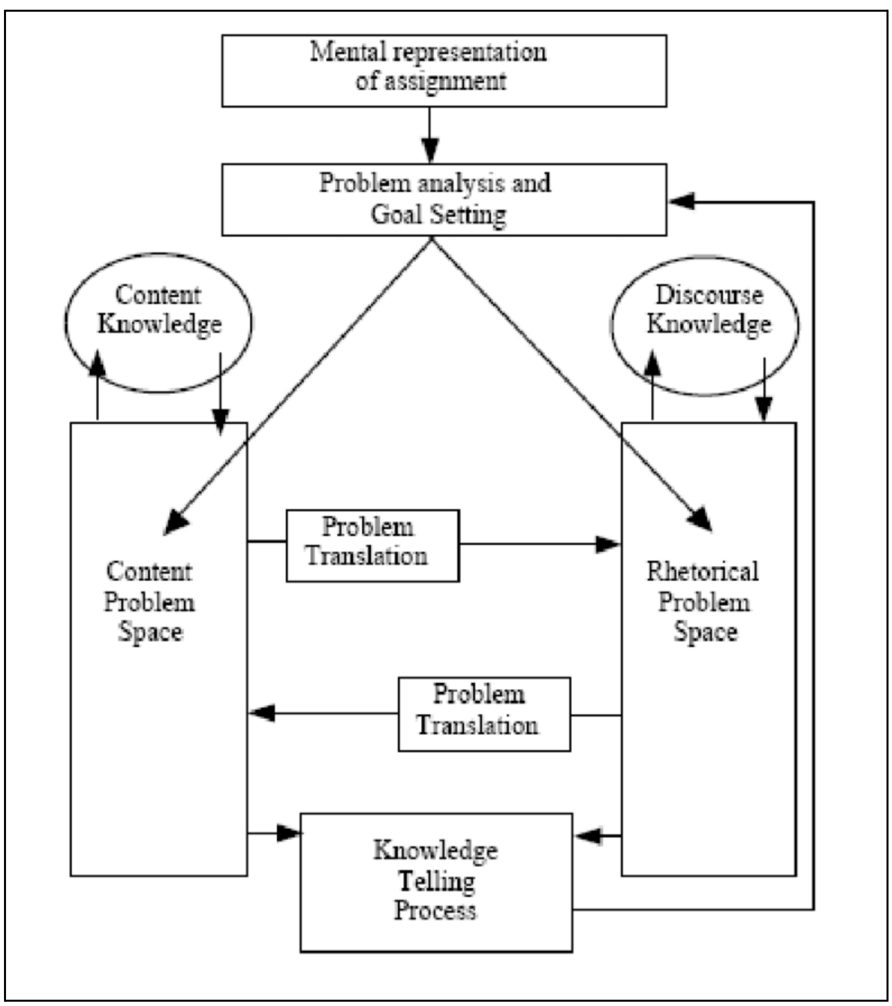
Structure of the Knowledge-Telling Model by Bereiter & Scardamalia (1987)



In contrast, *knowledge transforming* is much more complex and involves wrestling with one's existing knowledge by solving problems related to content and rhetoric, partially resembling Flower and Hayes' (1981) notion of solving rhetorical problems while writing. In this scenario, ideas are not written down as they come up, like someone engaged in knowledge telling would do, but are improved and revised first (Scardamalia et al., 1984). As a consequence of engaging in the knowledge-transforming process, both the writer's level of understanding of the topic and their level of writing expertise increase (see Figure 4; e.g., Scardamalia & Bereiter, 1991).

Figure 4

Structure of the Knowledge-Transforming Model by Bereiter & Scardamalia (1987)



Research on writing today continues to refer to the distinction between knowledge-telling and knowledge-transforming writing processes and strategies (e.g., Nückles et al., 2020), both of which play an important role in Galbraith's dual-process model of discovery writing.

Discovery Writing

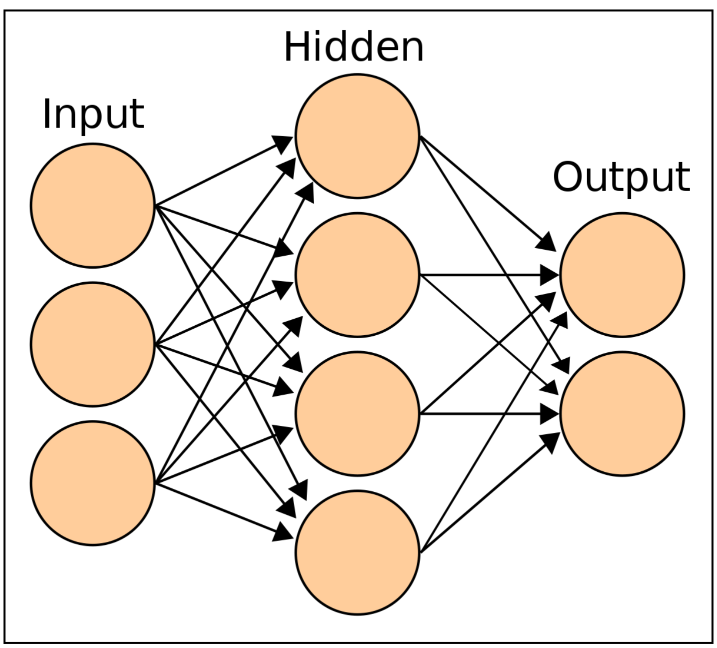
As described in Chapter 1, Galbraith's dual-process model of discovery writing is central to the intervention in this study, positing that during writing, writers engage in both knowledge transformation and knowledge constitution (i.e., developing new knowledge; Galbraith, 2009; Galbraith & Baaijen, 2018). As mentioned earlier, this model builds on the models developed by Flower and Hayes (1981) and Bereiter and Scardamalia (1987).

In what follows, I describe and summarize the theoretical and empirical literature that supports discovery writing, with a specific focus on the dual-process model. I include a brief description of work by other scholars and a more detailed summary of some of Galbraith's (e.g., 1992, 1999) own empirical studies that built toward the dual-process model in its current state.

Parallel Distributed Processing Models. At the heart of Galbraith's (2009) dual-process model lies what is known as *parallel distributed processing* (PDP) models, based in an information-processing paradigm called *connectionism*. This paradigm assumes that the brain and its functioning, such as neural networks, are an important source of information for how the mind works (Gibbons, 2018). Dating back to work by Rumelhart, Smolensky et al. (1986), researchers use this category of models when theorizing or researching tasks that require one to keep multiple pieces of information in mind at once, such as writing. PDP models are based on the assumption "that information processing takes place through the interactions of a large number of simple processing elements called units, each sending excitatory and inhibitory signals to other units" (McClelland et al., 1986, p. 10). What this means can be seen most clearly in examples of applied PDP models, such as models applied to word recognition. For instance, Figure 5 contains a PDP forward-feeding model, in this case representing a word recognition scenario. The first layer represents the word that serves as input and what is activated upon hearing the word, the second layer is hidden, an implicit representation of knowledge that guides output activation, and the third layer represents output, in this case the semantics that the listener ultimately associates with the word. Specific input activates connections that lead to certain output. These connections as well as their strength are idiosyncratic, and the same set of connections can lead to different output, depending on the nature of the input (e.g., Galbraith, 2009; Rumelhart, Hinton et al.,1986).

Figure 5

PDP Forward-Feeding Model



Note. From Burnett (2006).

A similar structure and pattern of activation can be seen in Galbraith's (2009) dual-process model when looking at the task and topic specifications as input, the activation within the writer's dispositional network as the hidden layer, and the language chunks as output. This serves to illustrate the implicit text and knowledge production process that is a key component of the dual-process model (see Figure 1 in Chapter 1). Graham et al. (2020) essentially give words to the schematics of Figure 1:

> "It is assumed that the writer's understanding of a concept is implicit and not directly retrievable. It is *only through the act of composing text that the writer discovers or has access to this knowledge*. More specifically, the writer must synthesize these implicit connections in memory until an initial fusion of understanding is obtained. [...]. The extent to which writing leads to new learning depends on the extent to which the content produced by the writer in this way differs from content already held in the writer's episodic memory. When this differs, *the writer presumably experiences a new development in understanding*" (pp. 181-182, *emphasis* added).

The following sections shed more light on the history, development, and specific elements of the dual-process model.

Work by Galbraith and Colleagues. Galbraith and colleagues have worked on developing the dual-process model for over 20 years. In one of the first relevant studies, Galbraith (1992) contrasted two different conceptions of discovery through writing. The first, the "classical position," following Aristotle and Cicero, posited that writing is fundamentally goal-directed and that discovery follows from rhetorical problem-solving (Galbraith, 1992, p. 45), a position held by Flower and Hayes (1981) and Bereiter and Scardamalia (1987). The second, the "romantic position," stems from the observation that some writers simply start writing and figure out what to say and how to say it as they go, and that this leads to discovery (Galbraith, 1992, p. 48).

To test these two positions, Galbraith (1992) conducted an experiment with undergraduate students, divided them into groups of “low and high self-monitors": students who are more or less susceptible to perceptions and expectations of others in writing and selecting what writing strategies to use (Snyder, 1979), demonstrating a parallel to Scardamalia and Bereiter’s knowledge-telling (i.e., low self-monitors) and knowledge-transforming writing approaches (i.e., high self-monitors; Galbraith, 2009). Low self-monitors would be expected to engage more in writing as follows from the romantic position, whereas high self-monitors would be expected to focus on satisfying rhetorical goals (Galbraith, 1992).

During the study, students were asked to choose a writing topic and, before receiving the actual writing task, to rate how much they felt they knew about the topic and list all topic-related ideas they could think of, including a rating of importance for each idea on the list. After completing an essay writing task, students completed the same measures and, subsequently, received their first list of ideas to find points of correspondence between the first and second list, determining whether the second list contained any new ideas (presumably discovered during writing).

After analysis of the results, Galbraith (1992) concluded that there, indeed, seemed to be *two sources of new ideas during writing*: one involving rhetorical planning, which characterizes high self-monitors, the other a more spontaneous, continuous writing down of prose and articulation of ideas (he later called this approach the *dispositional dialectic*, 1999), which characterizes low self-monitors. However, the low self-monitors’ writing was associated with clearer increases in knowledge as compared to the high self-monitors, leading Galbraith (1992) to conclude that this study supported the romantic position (i.e., discovery of new ideas and knowledge occurs more through a spontaneous writing process than one that involves satisfying rhetorical goals).

Subsequently, Galbraith (1999) used these results when building his case for a knowledge-constituting model of writing in a second important piece of the puzzle that would later become the final dual-process model. Here, he examined in more detail the mechanisms or processes responsible for the development of understanding during writing, specifically looking at whether “active rhetorical problem solving” is associated with that development, and whether that same process is active during planning and translation processes (Galbraith, 1999, p. 139). After presenting the elaborate processes and mechanisms behind the proposed knowledge-constituting model, he described an experiment that aimed to extend and expand the findings of the 1992 study.

As in the previous experiment, students were categorized as low and high self-monitors and were asked to list topic-related ideas before and after the writing task. However, this time, instead of only including a measure of subjective understanding (i.e., “how much do you feel you know about the topic?”), Galbraith (1999) also included a measure of subjective organization, to identify different kinds of transformation of thought. Additionally, he manipulated the type of planning writers used, including three conditions: (a) unplanned, asking writers to write without thinking about organization, (b) synthetically planned, giving writers five minutes to write down a sentence summarizing their response before the writing task, and (c) outline planned, giving writers five minutes to create an outline of the text to be written. In line with the previous experiment, Galbraith found that there was a clear, significant difference in the number of new ideas produced by low versus high self-monitors and a significant positive correlation between new ideas produced and increase in knowledge in the synthetic planning condition. Further, more high than low self-monitors reported a change in subjective organization of thought; for low self-monitors in the outline planned condition, the number of new ideas was associated with change in subjective organization, but this was not the case in the unplanned and synthetically planned groups (i.e., the groups with less explicit planning strategies). Galbraith concluded that these findings supported some of the knowledge-constituting model's main claims:

> That *new ideas produced by the dispositional dialectic* (assumed to be active in low self-monitors' synthetically planned text) should be associated with *increased knowledge* [...], but not with changes in organisation. It also supports the claim that n*ew ideas produced by explicit planning* (as in the low self-monitors' outline planned texts) should be associated with *changes in organisation*, but not with increases in knowledge. (Galbraith, 1999, p. 153; *emphasis* added).

Finally, in 2009, Galbraith presented the first official version of the dual-process model in his article *Writing as Discovery*, reiterated and refined in a later version, *The Work of Writing: Raiding the Inarticulate* (Galbraith & Baaijen, 2018). Moving away from models that treat discovery during writing as a "side-effect" (Galbraith, 2009, p. 5), he presented a model that moved beyond exclusively explicit thinking processes during writing (the more classical approach described earlier) and added more implicit text production processes. In formulating the model and conducting experiments to test it, Galbraith concluded that, even though they recognize the curious occurrence of novel content generation during writing, neither Flower and Hayes' (1981) model of writing as goal-directed problem-solving nor Bereiter and Scardamalia's (1987) knowledge-telling and knowledge-transforming model of writing accounted for the generation of truly novel ideas and content. Specifically, Flower and Hayes described idea generation as a subcomponent of the planning process (see Figure 2 for reference), involving the retrieval of content from memory in response to the writer's goals and the task constraints. In turn, Bereiter and Scardamalia described content generation for the knowledge-telling model specifically as constructing memory probes and retrieving content using those probes (see Figure 3 for reference). As Galbraith aptly put it: “It is difficult to see how content which is simply retrieved from a memory store can be ‘new’” (2009, p. 16).

Consequently, some other component or mechanism was needed to model this process accurately, which is where Galbraith’s (2009) dual-process model fits in. This model incorporates the planning components of Flower and Hayes and Bereiter and Scardamalia’s models, as well as some of Bereiter and Scardamalia’s knowledge-telling and knowledge-transforming components into the *knowledge-transforming process* (originally called the knowledge-retrieval process in 2009), but adds the *knowledge-constituting* process (Galbraith & Baaijen, 2018). According to Galbraith, both processes are necessary for effective writing and, during writing, they often interact and sometimes even collide with each other, requiring the writer to manage the “conflict” and generate knowledge that “satisfies rhetorical goals but at the same time fully captures the writer’s implicit understanding of the topic” (Galbraith & Baaijen, 2018, p. 246).

Self-Explanations. A concept and body of literature that supports the idea that writers can generate new knowledge during writing is research on the *self-explanation effect* (Chi et al., 1989). As the term implies, the concept of self-explanation refers to any learner generating an explanation to oneself (e.g., Chi et al., 1994), similar in some aspects as to what may occur during a writing task (e.g., Klein, 2004).

Self-explanation has been used as an effective learning strategy: students prompted to engage in more self-explaining learned more, monitored their comprehension more accurately, and became more effective problem solvers than students who engaged in less self-explaining (e.g., Chi et al., 1994; Pillow et al., 2002; De Bruin et al., 2007; for a review, see Fonseca & Chi, 2011). For example, Wong et al. (2002) tasked one group of ninth-graders with using self-explanations while studying about a new theorem in geometry, while the other group of students used their usual study habits to learn the same content, using pre- and posttest knowledge and problem-solving measures. The self-explanation group accessed more (prior) knowledge, generated more knowledge, scored higher on a problem-solving task, and performed better on a far-transfer task, requiring them to apply what they learned to a new, different set of problems (Wong et al., 2002).

Outcomes of Discovery Writing

In his commentary on a Special Issue about calibration in *Learning and Instruction* (2013), John Hattie concluded:

> "All these claims [forwarded in preceding articles regarding issues surrounding calibration] beg *more listening to learners during the teaching process*. Listening to what prior knowledge they bring; listening to their understanding of the goals of learning; listening to where they are moving, from priors to goals; and, listening to their conceptions of confidence and accuracy in answering these questions" (2013, p. 65; *emphasis* added).

This call to action, as it were, served as one of the catalysts for the present study and, throughout the writing of these chapters, served to remind me of its overarching purpose: listening to what learners think they know, observing the knowledge they present and develop in their writing, and inquiring about their motivation for the topic and tasks at hand.

In the sections that follow, I dive deeper into those outcomes of discovery writing that were measured or calculated in the present study: (a) subjective knowledge, (b) confidence and calibration, (c) situational interest, and (d) learning and accumulated knowledge.

Subjective Knowledge

As the literature on knowledge and knowledge acquisition made one's existing knowledge a central component of any occasion of learning something new, researchers such as Flavell (1971) and Brown (1978, 1988; Palincsar & Brown, 1986) proposed that one's knowledge as well as *awareness* of what one knows was an important contributor to any learning process. Finding out what learners already know, and, as an extension of that, what they *think* they know, became an important way to explain why, what, and how learners learn and introduced the concept of *metacognition*, generally defined as "awareness and control of one's learning" (Gourgey, 2002, p. 18; Flavell, 1979). As the literature on metacognition is vast and contains much conceptual and definitional ambiguity, any attempt to summarize it here would not do it justice. However, for the present study, I can say that the rating of *subjective knowledge* is a metacognitive act, and a consistent aspect of Galbraith and colleagues' work on discovery writing. From research on metacognition and metacognitive awareness, it has become apparent that students who are more metacognitively aware and are able to use more metacognitive strategies (e.g., Schraw, 2002) learn more effectively than students who are less metacognitively sophisticated, which is one of the reasons for including subjective knowledge as a variable in knowledge- and learning-related studies.

Recently, Galbraith and colleagues (2023) developed a novel measure of subjective knowledge and used this to measure outcomes of thinking as a result of discovery writing. The measure, which plays an important role in the present study, reflects the dual-process model and includes subscales of subjective *understanding* and subjective *organization of thought*. As described in earlier sections, Galbraith previously used single-item ratings to measure these constructs, which provided some insight but not enough to warrant more solid, generalizable conclusions (Galbraith et al., 2023). In the construction of the most recent measure, Galbraith et al. first generated a list of questions that potentially captured the two constructs, including the "old" items used before, as well as new items, such as "How coherent are your thoughts about the topic?" and, after piloting for comprehension, settled on a set of 12 items (see Table 1) and distributed the measure to undergraduate students, 165 of whom returned a completed version.

Subsequent analyses resulted in a two-factor solution: items that loaded on the first factor all related to understanding and knowing, and items that loaded on the second factor referred to organization, hence the naming of the two subscales *understanding* and *organization*. The final version of the measure includes eight of the original 12 items (see Table 1 and Appendix I). Preliminary results from two studies referenced in Galbraith et al.'s chapter (2023) indicated that the measure was reliable and replicated results from earlier studies.

Table 1

Initial Subjective Knowledge Measure (Galbraith et al., 2023)

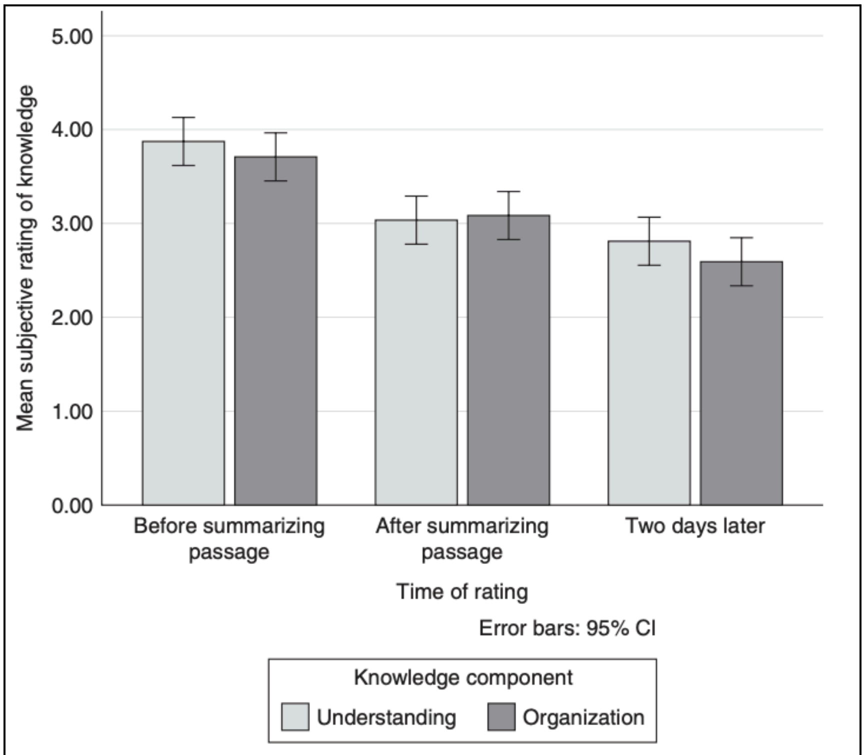
Items
How much you feel you know about the topic*
How well you understand the topic*
How organized your thoughts on the topic are*
How well you could explain the topic*
How clear your thoughts on the topic are
How clear your interpretation of the topic is
How coherent your thoughts about the topic are*
How structured your thoughts about the topic are*
How much you can make sense of the topic's issues*
How clear the relationships between your ideas about the topic are
How well you comprehend the topic's issues
How well-ordered your thoughts about the topic are*

Note. *Items included in the final version of the measure

During the first study (Peters, n.d., as cited in Galbraith et al., 2023), participants read a passage and were given ten minutes to summarize it. Two days later, they completed comprehension measures about the passage. After reading, after writing, and after the comprehension measures, students were asked to fill out the subjective knowledge measure. Results showed that writing the summary led to a significantly greater decrease in understanding than in organization, and the measure completed two days after the writing task showed a significantly greater decline in organization than in understanding (see Figure 6). Additionally, Peters found a significant positive correlation between the subjective knowledge ratings and the comprehension scores.

Figure 6

Results from Peters' study in Galbraith et al. (2023)

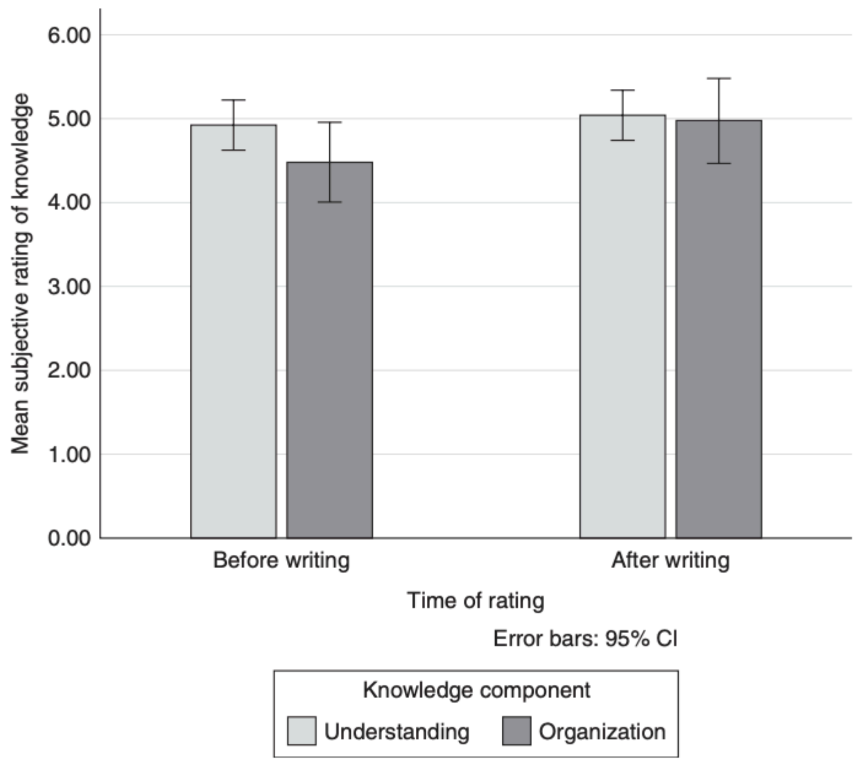


Ratings of understanding and organization as a function of time of testing (N = 78).

The second study (Hall, 2023) tasked participants with writing an essay. This experiment was similar in structure to the earlier studies conducted by Galbraith (e.g, 1992, 1999): students were asked to list ideas about the topic, rate their understanding of the topic, make a plan for the writing task in two different conditions, write the essay, and rate their understanding of the topic again. The results of this study indicated that ratings of understanding increased after writing, regardless of the type of planning carried out before writing. Specifically, subjective organization increased significantly, and subjective understanding increased but not significantly so (see Figure 7).

Figure 7

Results from Hall's study in Galbraith et al. (2023)



Subjective ratings of understanding and organization before and after writing (N = 43).

Interestingly, summarizing tasks such as those used by Peters (n.d.) and others (e.g., Arnold, 2017) seem to have a different effect on subjective organization than argumentative tasks such as the one used by Hall (2023), but there is not enough evidence yet to warrant conclusions about the specifics of these differences and why they occur. In their chapter (2023), Galbraith and colleagues therefore suggested that future research endeavors be focused on examining the nature of the differential effects of writing tasks on learning and subjective knowledge, which was one of the purposes of the present study.

As described above, so far, the subjective knowledge instrument has been used exclusively to gauge differences in subjective knowledge before and after a writing task. In the present study, I went beyond this and evaluated the use of this measure as a tool for gaining more insight into the effect of reading versus writing on subjective knowledge and explored its use as a tool for confidence and calibration assessments.

Confidence and Calibration

An important part of assessing what learners know and think they know is the evaluation of learners' confidence and calibration—the latter "a measure of the relationship between confidence in performance and accuracy in performance, indicating how aware individuals are of what they do and do not know" (Stone, 2000, p. 437). Another more abstract and generalized definition describes calibration as "the accuracy or alignment between a judgment and a meaningful standard" (Hadwin & Webster, 2013, p. 39; Winne & Muis, 2011).

Being well-calibrated (i.e., a good match between confidence levels and accuracy) when it comes to knowledge, learning, and academic outcomes is important because it helps students determine their goals (e.g., Hadwin & Webster, 2013), the steps they need to take to meet those goals (e.g., how much attention they pay in class, how much effort they put into classwork, etc; e.g., van Loon et al., 2013), and their help-seeking behaviors (e.g., Boekaerts & Roozendaal, 2010; Winne & Jamieson-Noel, 2002; for a review, see Hacker & Bol, 2019). However, even though confidence ratings and subsequent calibration scores are frequently used as important indicators of knowledge and performance (e.g., Raedts et al., 2007; Singer & Alexander, 2017), these measures often consist of single-item ratings and do not provide detailed information beyond a numeric score that represents confidence in, for instance, the correctness of an answer to a multiple-choice question. Galbraith and colleagues' subjective knowledge instrument could constitute an easy-to-incorporate tool that would yield much more information about learners' subjective ratings regarding their knowledge and organization of that knowledge than those often used in the calibration literature.

Interest

A third important variable in any learning-related task is motivation (see Urhahne & Wijnia, 2023 for a review). As mentioned in Chapter 1, for the purposes of the present study, I chose *interest* as a representational motivational variable, which I will define and describe in more detail in this section.

As evidenced by the large number of studies on interest and its impact on learning-related outcomes (e.g., Renninger & Hidi, 2017), starting as early as 1913 (i.e., Dewey's classic work *Interest and Effort in Education),* interest has been said to play a crucial role in any educational endeavor focused on learning and developing new knowledge, especially when the task at hand is cognitively demanding (e.g., a writing task; Hayes, 2012). For example, interest has been found to impact students' levels of attention to a task (e.g., Schiefele, 1991; Ainley et al., 2002), engagement with said task (e.g., Renninger & Hidi, 2017), effort put into the task, (e.g., Renninger & Hidi, 2002), as well as learning gains (e.g., Alexander, 1997; Rotgans & Schmidt, 2011), the use of learning strategies (e.g., Schiefele, 1991), and goal setting (e.g., Harackiewicz et al., 2000; Pintrich & Zusho, 2002). Therefore, given the focus of the current study on learning and the development of new knowledge, it seemed important to include a measure of interest.

In choosing how to frame interest, Hidi and Renninger's influential four-phase model of interest development (2006) has served as an important foundation. These authors, as well as many other interest theorists, distinguish between *situational* and *individual* or personal interest. On the one hand, situational interest refers to attention and an affective reaction activated by the topic or task parameters and the way they are presented (e.g., Hidi, 1990; see Schraw & Lehman, 2001 for a review). This kind of interest occurs *in the moment*, in response to environmental stimuli (e.g., Hidi & Renninger, 2006; Linnenbrink-Garcia et al., 2010). Individual interest, on the other hand, constitutes a more "dispositional quality, residing in the person across situations" (Linnenbrink-Garcia et al., 2010, p. 648).

In their model, Hidi and Renninger (2006) referred to four phases of interest development, reflecting and refining this distinction: (a) triggered situational interest, (b) maintained situational interest, (c) emerging individual interest, and (d) well-developed individual interest. In this model, one phase, if sustained, may develop into the next phase. For example, if sustained, *triggered* situational interest (i.e., interest "sparked" by contextual factors; Hidi & Renninger, 2006 p. 114) may develop into *maintained* situational interest (i.e., a psychological state of interest that occurs more persistently; Hidi & Renninger, 2006).

Because of the relatively brief nature of the current study, triggered and maintained situational interest are more relevant than emerging and well-developed individual interest (i.e., a more enduring predisposition in a person to come back to a particular topic; e.g., Schraw & Lehman, 2001). This is why I measured both triggered and maintained situational interest throughout the current intervention (see Chapter 3 for more details).

Learning and Accumulated Knowledge

Last, I want to consider learning and accumulated knowledge as an outcome of discovery writing and, really, of any educational endeavor. Individuals engaged in teaching of any kind ultimately seek to help students learn and acquire knowledge. To assess knowledge and learning, educators typically use tasks that, at least in theory, reflect what students have learned until that point: multiple-choice tests, essays with corresponding rubrics, more hands-on projects, etc. For example, in the undergraduate courses sampled for this study, instructors used exams, lab reports, and collaborative research projects to assess their student's knowledge levels (UMD HDQM, 2024a, 2024b; UMD PSYC, 2023). Compared to measures used in research, measures used by instructors are often less systematic and usually do not involve a pretest-posttest procedure, simply because of contextual constraints such as time and resources.

Like educators, many researchers across domains consider learning or some knowledge-related variable as the outcome of their intervention studies. The assessment of learning and knowledge gains in these cases is often more systematic. This holds true for the writing-to-learn body of literature referred to earlier, in which researchers used writing as a treatment to advance students' knowledge in different academic domains, and quite effectively so (for a review and meta-analysis, see Graham et al., 2020). Within this literature, researchers have used various forms of assessment to evaluate learning gains, including multiple-choice items, short-answer questions, and open-ended response tests with corresponding rubrics (Graham et al., 2020).

In parallel with Galbraith and colleagues' work (e.g., 2009, 2018, 2023) and in line with the claims of the discovery writing paradigm and the dual-process model, knowledge was one of the outcomes of the present study, as well. As mentioned in Chapter 1, I used two different types of knowledge assessment: (a) concept maps with corresponding scoring schemes (Novak & Gowin, 1984) and (b) a writing task that was scored using the SOLO Taxonomy (Biggs & Collis, 1982). I selected both of these evaluation tools because of their potential to provide more in-depth and insightful information about the state of students' knowledge before and after the intervention than many short-answer or multiple-choice knowledge measures.

Concept maps are visual representations of students' knowledge, including how concepts are linked and hierarchically ordered, providing insight into the structure and depth of students' knowledge (Novak & Gowan, 1984). Similarly, the SOLO Taxonomy aided me in providing a holistic score for students' writing samples, based on the structure and interconnectedness of students' knowledge as written down. Both of these types of assessment have been reliably used as a method of evaluating change in knowledge structures and meaningful learning (e.g., Hung & Lin, 2015; Watson et al., 2016; Dinsmore & Alexander, 2016). Detailed descriptions of these measures as well as how they were used for analytic purposes can be found in Chapter 3.

Learning From and With Text

The primary focus of my research has been on *writing* as a tool for developing knowledge. However, in the present study, students in all conditions started their learning activities with a *reading* task, and students in the comparison condition subsequently engaged in rereading strategies. This warrants a brief section on the effect of reading on learning, providing a summary of the relevant literature as well as strengthening the rationale for why the comparison rereading condition served as a "worthy opponent" of the experimental writing conditions.

Learning from (re)reading. As with the literature on writing, knowledge, and interest, the literature on reading and learning is too vast and complex to attempt to define and summarize it in a short section such as this one. However, I have selected relevant highlights of the theoretical and empirical literature on reading, which should suffice in providing the background for reading as a subcomponent of the present study.

Learning from and with text through reading is an age-old practice (e.g., van Dijk et al., 1983) that remains an important vehicle for activating existing and acquiring new knowledge in current educational practice (e.g., Greenleaf et al., 2023; Lapp et al., 2016). Consequently, there is a large body of literature on reading and text comprehension strategies, ranging from theories and frameworks (e.g., Rayner et al., 2011; Tracey & Morrow, 2024) to empirical interventions (for reviews see e.g., McNamara, 2007; Melby-Lervåg & Lervåg, 2014), all to try and understand what strategies are more or less effective for learning from and with text and, importantly, how one can get students to employ the more effective strategies. Summarizing, drawing inferences, generating questions, and rereading are examples of effective comprehension strategies put forth by the literature (e.g., Dole et al., 1991; McNamara, 2007), although specific task parameters or conditions may impact the level of effectiveness. For example, Rawson et al. (2000) found that rereading a text is more effective for comprehension and metacomprehension accuracy than reading a text once. Furthermore, in the context of having to engage with *multiple* texts, McNamara and colleagues (2023) recently concluded that rereading these texts was more beneficial for comprehension than summarizing them.

Although the reading and writing researchers largely operated separately for a long time (e.g., Consalvo & Schallert, 2002), a seminal piece by Tierney and Pearson (1983) posited that both reading and writing consist of *actively constructing meaning* and that processes involved in reading are remarkably similar to those involved in writing (e.g., *planning* to retrieve relevant knowledge; Consalvo & Schallert, 2002). These processes echo writing processes proposed by researchers like Flower and Hayes (1981) as described earlier in this chapter. This emphasizes that reading and writing can be considered worthy counterparts in a learning-focused intervention such as the one I tested in the present study. However, following the dual-process model of writing that frames the current study (i.e., Galbraith & Baaijen, 2018), students engaging in reading *and* writing tasks, as opposed to reading and re-reading, are expected to create explicit output of the meaning-making process and then feed that output back into their dispositional network. In other words, students engaging in learning tasks beyond (re-)reading engage more deeply in constructing ideas from their own knowledge base as opposed to mainly in response to cues from the author of a text. Whereas familiarizing oneself with new material can be achieved through (re)reading, deeper learning, beyond acquiring novel ideas, may require something more (e.g., Dunlosky, 2013).

Revisiting the Research Questions

In sum, the writing process literature and Galbraith et al.'s (2009, 2018) dual-process model of writing, specifically, formed the foundation for the present study. The dual-process model and its claims about knowledge development during writing, combined with the potential of the novel subjective knowledge measure (Galbraith et al., 2023) served as a catalyst for the present study's research design. Before presenting the details of my design, I want to re-introduce the research questions that guided the present study:

1. How do students' SKRs, confidence ratings, situational interest ratings, and knowledge scores change as a function of writing as compared to rereading?
2. How do subjective knowledge and confidence ratings compare to measures of knowledge?
    a. Specifically, how do students' SKRs and confidence ratings compare to knowledge scores of the pre-posttest concept maps and knowledge structures evident in the posttest writing sample?
3. Do SKRs, confidence ratings, and situational interest ratings predict students' post-intervention knowledge scores by experimental condition?

Context of the Current Study

Reading and Writing

In the current study, students were tasked with reading and writing tasks of varying nature, the procedural details of which I explain in Chapter 3. The reason for choosing these two modalities as learning tasks was that reading and writing already serve as two of the primary course components and ways in which undergraduate students typically acquire knowledge in the classes they are taking (see earlier sections in Chapter 1 and the current chapter). Any intervention study that uses a combination of reading and writing tasks would, therefore, come quite close to naturally-occurring circumstances, which would strengthen conclusions drawn about learning processes and gains as a result of the study.

Writing Genres

The writing tasks assigned to students in the experimental conditions in this study consisted of a free-write, an explanation, and a persuasion task. Students in all conditions completed a posttest argument task (see Chapter 3, hereafter referred to as the *transfer test*). I chose these tasks not only to test the differential effect of these writing genres on potential knowledge development throughout the intervention (i.e., in response to Galbraith et al.'s call, 2023), but also to avoid a so-called floor effect—in other words, assigning only a free-writing task would likely result in low variability in knowledge scores awarded to those writing samples, whereas more complex tasks such as an explanation, persuasion, or argument task (in order of increasing complexity) would create more variability and thus more insight into those specific writing genres and their effect on learning. This is because these four genres (i.e., free writing, explanation, persuasion, and argumentation) involve increasingly complex coordination of writing processes. Free writing requires students to write rather quickly and continuously for ten to fifteen minutes, while inhibiting any impulse to edit or modify the text (e.g., Elbow, 1973), whereas explanatory writing requires students to consider their understanding of the topic at hand and portray this to the reader by constructing plausible cause-and-effect relationships and inferences (e.g., Klein & Rose, 2010). Further, persuasive writing requires students to provide claims and evidence and even include or rebut counterarguments in order to convince their reader of their standpoint (e.g., Wolfe & Britt, 2008).

Additionally, and perhaps most importantly, the argument task (i.e., transfer task) that all students completed, represents learning outcomes central to many, if not all, undergraduate courses. Instructors, myself included, want students to be able to do more than simply regurgitate information or explain a phenomenon. Instead, we would like them to (a) be able to evaluate evidence, (b) draw nuanced conclusions, and (c) apply learned information to contexts beyond the original (classroom) context, all of which are present in an argumentative writing task (e.g., Feretti & Lewis, 2018). For example, instructors for the courses sampled for this study included learning outcomes such as “the ability to critically evaluate the strengths and weaknesses of theories in human development,” “use scientific reasoning in the evaluation of research,” and “understanding of how course concepts are translated into applied tasks in a variety of career contexts” (UMD HDQM, 2024a, 2024b; UMD PSYC, 2023), all of which are reflective of those hopes and expectations. Because this study took place in real college courses and I aimed to draw conclusions that are relevant for educators as well as researchers, it became all the more important to create a transfer test that represents an authentic learning outcome, such as the argumentative writing task.

CHAPTER 3: METHOD

Design

The present study was a pretest-posttest, within- and between-subjects intervention that took place in real college classrooms. During the study, I compared the outcomes of two groups of students who engaged with reading materials related to their course content and subsequently took part in one of two writing interventions (writing for explanation versus persuasion) to a third group, serving as a comparison group that engaged in rereading the text materials about the topic. Students participated in all tasks during out-of-class activities designed to be completed as course homework.

Participants and Course Contexts

Participants in this study were undergraduate students enrolled in several courses at the University of Maryland, College Park. These courses included two human development courses, one called *Paradigms and Perspectives in Human Development* and the other *Research Methods in Human Development*, as well as sections of a psychology course called *Research Methods in Psychology Laboratory*. All three of these courses can be taken as a requirement by students majoring in human development or psychology, but also as an elective to satisfy General Education requirements.

Naturally, learning outcomes for these courses were varied, but there was some important overlap regarding the content of these courses in terms of presenting scientific reasoning and research principles in behavioral and social sciences. For example, the human development course about paradigms included demonstrating "knowledge of scientific reasoning" as a learning outcome; the human development research methods course had demonstrating "knowledge of all phases of the research process in the study of human development, including development of research questions and hypotheses, study design, ethical considerations, data collection, measurement, data analysis, and presentation of results" as one of its outcomes; and for the psychology research methods course, learning outcomes included being able to "formulate scientific questions, and translate these into ethical research studies, including an appropriate design and hypotheses" (UMD HDQM, 2024a, 2024b; UMD PSYC, 2023). This overlap in learning outcomes guided me in deciding on the topic of the intervention—*what makes a good research design*—which is relevant to all three of the courses in which participants were enrolled and fit seamlessly into the respective course curricula.

For one of the human development courses, the intervention was an integrated homework activity. For the other courses, it served as an extra-credit assignment presented as part of or instead of one of the assigned course readings or written discussion assignments. The intervention tasks were presented through two separate Qualtrics surveys (see Measures and Procedures for more details).

The initial sample for this study consisted of 241 students. However, as I inspected the data more closely, a total of 94 cases had to be removed from the sample for one or more of the following reasons:

- Students submitted a survey more than once. If this was the case, I kept the first submission for the sake of validity, unless the student only filled out demographics in the first submission and did not get exposed to any other content;
- Students did not complete the first survey *and* did not complete the second survey;
- More than one of the three knowledge measures were missing (i.e., pretest/posttest/transfer test);

- Students did not sign the consent form.

Following these criteria and after removal of one outlier case (as explained in more detail in Data Analysis), the final sample that was analyzed consisted of 146 students.

The fact that the courses sampled could be taken as both electives and General Education requirements resulted in rather diverse student backgrounds and class standings. Academic majors ranged from public health (6.8%) to natural sciences (6.8%) and education (10.3%), with a majority in behavioral and social sciences (57.5%; see Table 2). Fifteen students (10.3%) reported multidisciplinary majors. The majority of the students identified as White (57.5%) and female (78.1%), and considered themselves native English speakers (86.3%). The second most common racial/ethnic identity reported was Asian or Asian American (17.1%), followed by Black or African American (8.2%), Hispanic or Latinx (7.5%), or multiple identities (7.5%). Additionally, most of the students were juniors (40.4%) or sophomores (36.3%), with an average age of 20.18. More detailed demographic information can be found in Table 2.

Of the 146 students in the final sample, 54 were assigned to the comparison condition, 48 to the explanation condition, and 44 to the persuasion condition.

Table 2

Demographic Information About Sample

	(N = 146)	
	N	**%**
Gender Identity		
Male	25	17.1%
Female	114	78.1%
Non-Binary/Third Gender	6	4.1%
Other/Self-Reported	0	0.0%
Prefer not to say	1	0.7%

Native English Speaker		
Yes	126	86.3%
No	20	13.7%
Racial/Ethnic Identity		
American Indian or Alaska Native	0	0.0%
Asian or Asian American	25	17.1%
Black or African American	12	8.2%
Hispanic or Latinx	11	7.5%
Middle Eastern or Northern African	3	2.1%
Native Hawaiian or Other Pacific Islander	0	0.0%
White	84	57.5%
Other/Self-Reported	0	0.0%
Multiple	11	7.5%
Prefer not to say	0	0.0%
Class Standing		
Freshman	10	6.8%
Sophomore	53	36.3%
Junior	59	40.4%
Senior	23	15.8%
Other/Self-Reported	0	0.0%
Prefer not to say	1	0.7%
Course Enrollment		
EDHD200-0101-0102	34	23.3%
EDHD200-0201	17	11.6%
EDHD306-0101	13	8.9%
EDHD306-0201	16	11.0%
PSYC300	66	45.2%
Academic Major		
Agriculture	2	1.4%
Architecture	0	0.0%
Arts & Humanities	2	1.4%
Behavioral & Social Sciences	84	57.5%

Business	5	3.4%
Natural Sciences	10	6.8%
Education	15	10.3%
Engineering	0	0.0%
Information Studies	1	0.7%
Journalism	0	0.0%
Letters & Sciences	0	0.0%
Public Health	10	6.8%
Public Policy	0	0.0%
Multidisciplinary	15	10.3%
Prefer not to say	2	1.4%
Learning Disorders/Disabilities		
Oral/Written Language Disorder	0	0.0%
Non-Verbal Learning Disability	1	0.7%
Dysgraphia	0	0.0%
Dyscalculia	0	0.0%
Dyslexia	2	1.4%
AD(H)D	15	10.3%
Other/Self-Reported	4	2.7%
None	121	82.9%
Prefer not to say	4	2.1%

	M *(SD)*
Experience	
Writing for Coursework	71.71 *(21.35)*
Topic Familiarity	71.06 *(22.99)*
Age	20.18 *(1.09)*

Measures

Demographics

Before starting the intervention tasks, students were asked to fill out a demographics survey. This survey contained questions about students' age, gender identification, ethnicity, class standing, academic major, language proficiencies, learning disorders or disabilities, academic writing experience, and experience with the intervention topic—for example “What is your class standing?”, with options “freshman”, “sophomore”, “junior”, “senior”, and “other, please specify” to choose from. See Appendix I for the full survey.

Knowledge and Learning

To measure knowledge at the starting point of the intervention as well as knowledge and learning throughout and after the intervention, I analyzed and evaluated two sets of tasks: concept maps and writing samples.

Concept Maps

Before the intervention as well as right after the intervention took place, students were asked to create a concept map in answer to the question: “What makes a good research design?”. Concept maps consist of visual representations of students’ knowledge, including how concepts are linked and hierarchically ordered (Novak & Gowan, 1984). In providing a focus question rather than a topic as the prompt for this task, I followed Novak and Cañas’ (2007) recommendation for constructing concept maps.

Concept maps go beyond more straightforward multiple-choice knowledge questions and provide more insight into the structure of students’ knowledge (e.g., Hung & Lin, 2015; Watson et al., 2016), hence constituting a highly suitable pre- and posttest measure of knowledge for the present study.

The pre- and posttest concept maps were scored for evidence of knowledge using Novak and Gowin’s (1984) coding schemes (see Data Analysis for details).

Writing Samples

A week after the intervention took place and the posttest concept map was created, students in all conditions were assigned a final argument writing task (i.e., the transfer test), which was analyzed and assessed for knowledge structures, using the SOLO Taxonomy (Biggs & Collis, 1982, 1989; see Analyses for details).

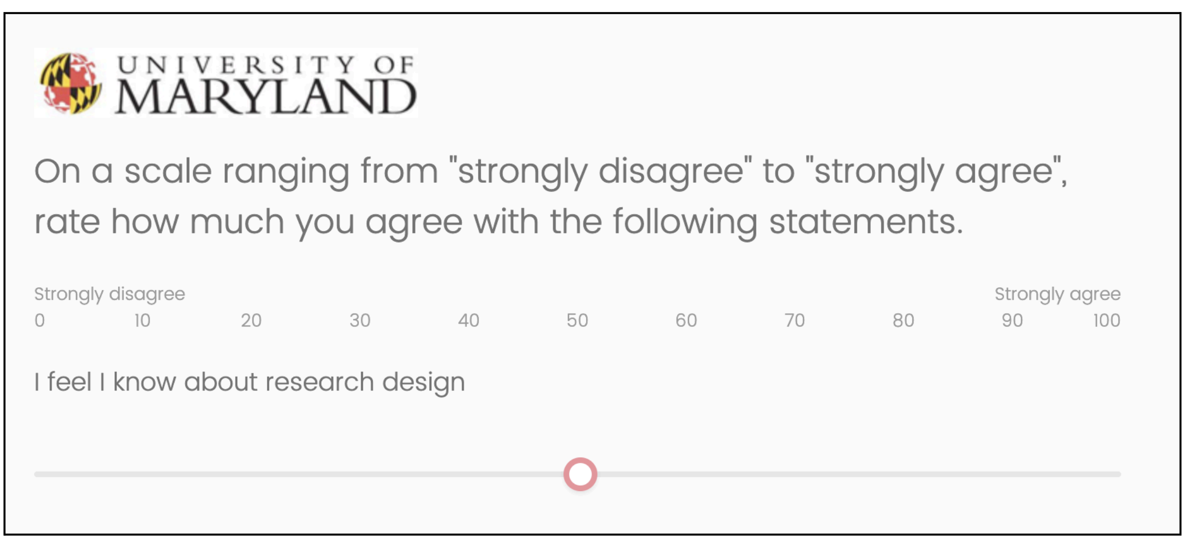
The comparison of the scores given to the pretest concept map, the posttest concept map, and the final writing sample served as a measure of learning.

Subjective Knowledge, Confidence, and Situational Interest

Throughout the intervention, students filled out a rating instrument a total of five times. This instrument contained subjective knowledge, confidence about knowledge, and situational interest items (17 items total). All items were phrased as statements (as opposed to questions) and were rated on a sliding scale that ranged from "strongly disagree" to "strongly agree" (see Figure 8). Example items will be given throughout this section, but the full rating instrument as it was used in this study can be found in Appendix I.

Figure 8

Example of Rating Instrument Item as Presented to Students



Subjective Knowledge

To measure students' subjective knowledge state, I used the eight items presented by Galbraith and colleagues in their Subjective Knowledge Rating Instrument (Galbraith et al., 2023) and adapted them to fit the intervention topic. This instrument measures two processes that together constitute subjective knowledge: subjective understanding and subjective organization. For example, as presented by Galbraith and colleagues (2023), "I understand research design well" is an item that measures subjective understanding; "My thoughts about research design are coherent" is an item that measures subjective organization.

Situational Interest

As part of the same instrument, students rated their levels of situational interest (SI) in the topic at hand by filling out eight items adapted from Linnenbrink-Garcia et al.'s Situational Interest Survey (2010). This survey consists of triggered and maintained situational interest items, distinguishing between a positive affective reaction to the way material is presented (i.e., triggered) and positive reactions to the material itself (i.e., maintained; Linnenbrink-Garcia et al., 2010). The items measuring *maintained situational interest* were most suitable for the purposes of the present study, which is why I chose to include only those. I included both subscales of maintained situational interest as presented by Linnenbrink-Garcia et al. (2010), *maintained-SI-feeling* and *maintained-SI-value*, both represented by four items for eight items total. These items focus on how enjoyable and meaningful the course material presented is, respectively.

Example items include "What I am learning about research design is fascinating to me" for maintained-SI-feeling and "What I am learning about research design can be applied to real life" for maintained-SI-value.

Confidence and Calibration

Finally, as part of the rating instrument, students filled out a confidence item, rating their level of confidence about their knowledge of the topic at hand. This item was constructed in a manner often used in the calibration literature and empirical research: "How confident are you about your knowledge of research design?" (e.g., Persky & Dinsmore, 2019).

Subsequently, I compared students' confidence ratings as completed throughout the intervention to the knowledge scores extracted from the pre- and posttest concept maps and in the transfer-test writing sample. By comparing the knowledge scores awarded to these tasks with the confidence items filled out after the respective tasks and calculating relative accuracy, I measured students' calibration (e.g., Schraw et al., 2013). See Data Analysis for more details about the calibration calculation and analysis.

Materials

Intervention Text

At the start of the intervention, all students were asked to read a text called *What Makes a Good Research Design?* The text was approximately 1800 words (i.e., 3 pages, single-spaced) and had two main sections: *Introduction and Key Concepts* and *RCTs vs. Observational Research: Examples from the Medical Field.*

The first section introduced important key terms related to research design, such as correlational and experimental research and internal and external validity. This section was based on excerpts from two textbook chapters on research design (i.e., Brown et al., 1999; Siegler et al., 2003). The second section of the text delved into strengths and limitations of randomized controlled trials (RCTs) and population-based observational research, using the medical field as a base for examples throughout. This section was based on excerpts from a journal article titled

Randomised controlled trials and population-based observational research: Partners in the evolution of medical evidence (Booth & Tannock, 2014).

The text was structured in such a way that it related to and matched existing course readings and content (see Appendix II for the full text).

Intervention Tasks

What follows is a general description of the intervention tasks. The detailed task prompts as they were used in the intervention can be found in Appendix II. In the Intervention and Procedures section, I describe how these tasks were implemented for each of the conditions.

Concept Mapping Tasks

Before and after the intervention, students were assigned pre- and posttest concept mapping tasks, asking them to create a concept map in response to the question: *What makes a good research design?* By clicking on a button that read "click here to start creating your concept map", they were directed to a simple, online tool (draw.io) where they could create their concept map. Before starting the actual concept map, they were provided with (a) an example concept map (see Figure 9), (b) a 1-minute instructional video about how to use the online tool for creating the concept map (see Figure 10 for a screenshot and Appendix II for the URL), and (c) a list of step-by-step instructions once they were present in the online tool (see Figure 11 and Appendix II).

Figure 9

Example Concept Map Provided to Students

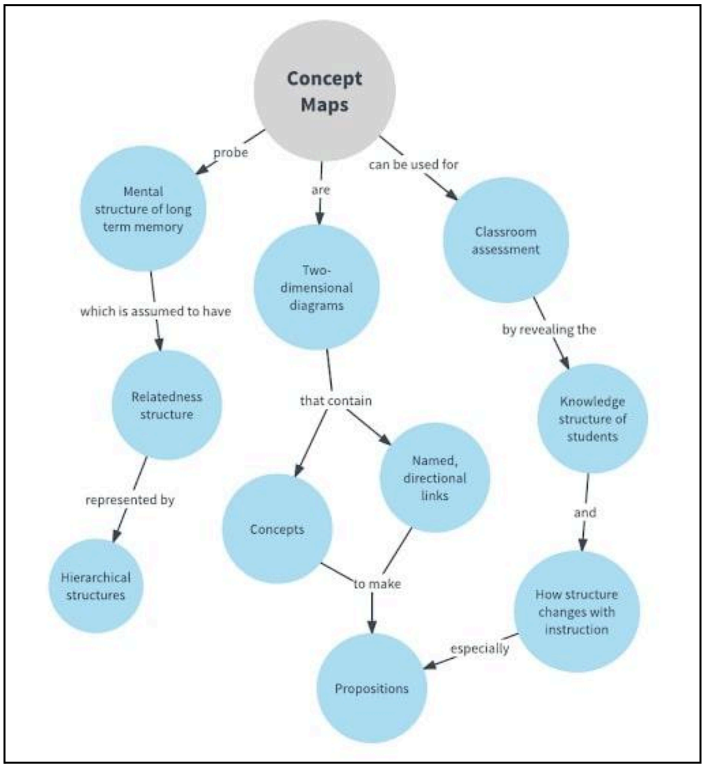


Figure 10

Screenshot of Instructional Video About How to Use Online Tool

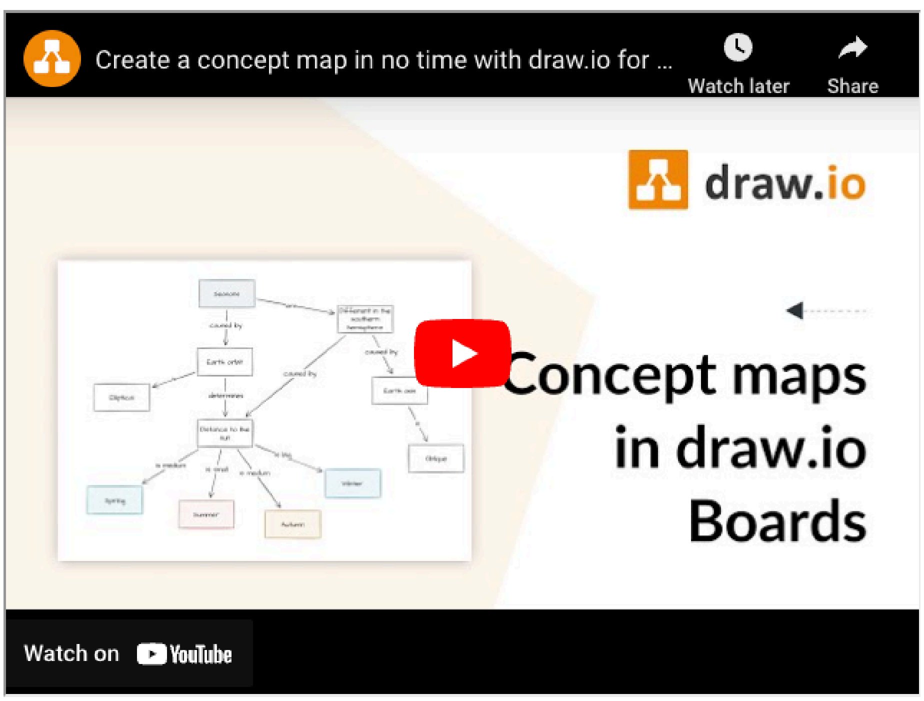
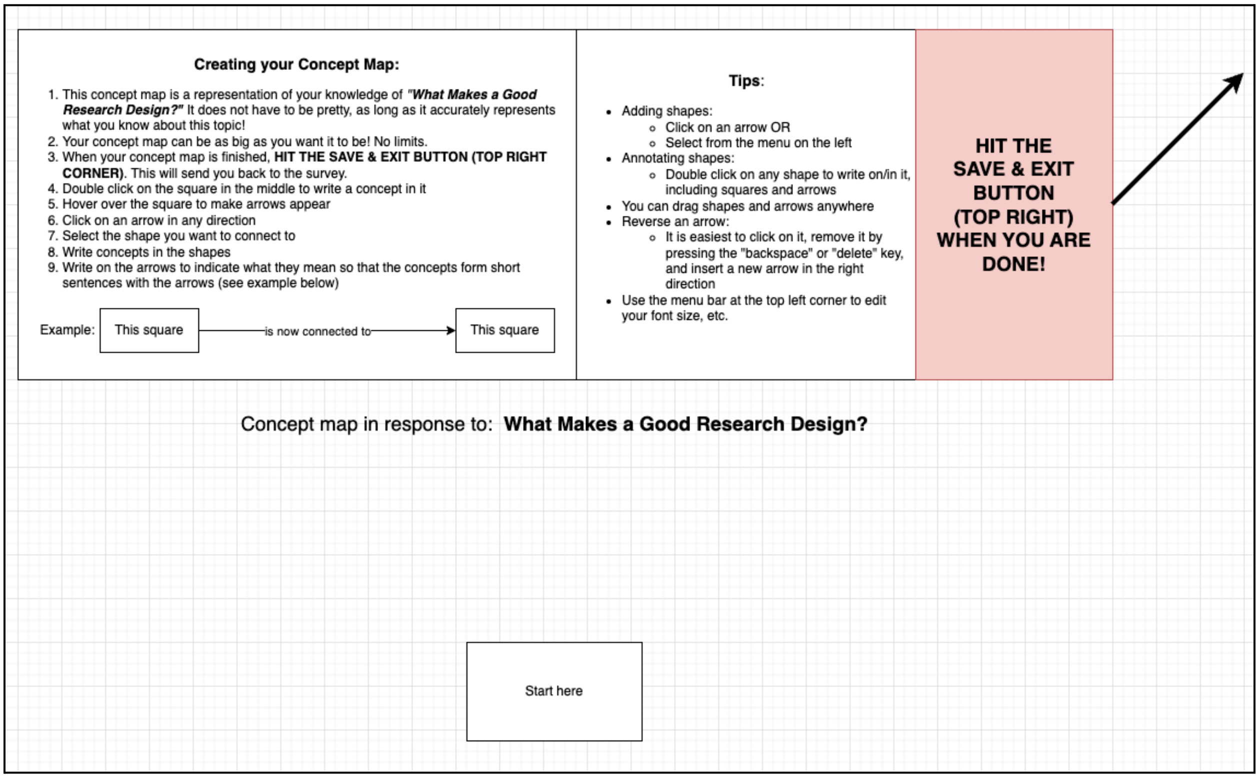


Figure 11

Online Tool Presented to Students for Creating Pre- and Posttest Concept Maps



Reading Tasks

At the start of the intervention all students were assigned a reading task, asking them to read the intervention text. Additionally, the comparison group received two *rereading tasks*, asking them to engage in a surface-level reading strategy (i.e., highlighting key terms) and a more deep-level reading strategy (i.e., highlighting any sentence or passage that contributed to their understanding of the topic; e.g., Dinsmore & Alexander, 2016), respectively.

Writing Tasks

After the initial reading task, the experimental groups were assigned two out of three writing tasks: (a) a *free-writing* task, asking them to write down everything they knew about what makes a good research design; (b) an *explanation* task, asking them to explain research design in psychology to someone else; and, (c) a *persuasion* task, asking them them to convince a friend that RCTs should be the gold standard for research design in psychology.

The experimental writing tasks were presented in order of difficulty (i.e., from a free write to an explanation task or from a free write to a persuasion task), leading up to the final writing task: as a transfer test, students in all groups (i.e., experimental and comparison) were assigned an *argument* writing task, in which they took a stance on what type of research design would work best for studying the effects of masking on COVID-19 transmission. The subtopic of this transfer test (i.e., masking and COVID-19) was slightly different from the experimental writing ones (i.e., research design in psychology), in order not to give students in the experimental conditions an "unfair" advantage over students in the comparison condition.

Qualtrics

All intervention-related tasks and measures were presented through Qualtrics surveys. Participants received the relevant links through their course instructors and online course pages (see Appendix II).

Intervention and Procedures

As mentioned, the intervention centered on a topic that played a role in all of the courses in which participants were enrolled: What makes a good research design? The intervention took place across two sessions (see Table 3): during the first session, students completed the majority of the intervention tasks; during the second session, students completed a transfer test.

Pretest

Before the intervention began, all students were asked to (a) fill out a demographics questionnaire, (b) create a pretest concept map about the intervention topic, and (c) complete a

9-item rating instrument about their level of subjective knowledge about the topic and their confidence about their topic knowledge.

Subsequently, students were randomly assigned to one of the experimental groups or to the comparison group (see Table 3). Then, students in all conditions were tasked with reading the text about research design described above.

After at least ten minutes of reading, all students were asked to fill out the same rating instrument they had filled out initially (i.e., subjective knowledge and confidence), this time including situational interest (resulting in 17 items total). Then, the intervention tasks were different for each condition—Experimental Condition 1, Experimental Condition 2, and the Comparison Condition.

Experimental Conditions

After the initial reading and rating tasks, students in both of the experimental conditions were assigned a writing task in which they were asked to write down everything they know about research design. Subsequently, they filled out the rating instrument again (i.e., rating their subjective knowledge, confidence in their knowledge, and situational interest). Following this, they were assigned a second writing task.

For students in Experimental Condition 1, the second writing task consisted of answering the question "How would you explain research design in psychology to a friend who is trying to learn about this topic?", constituting an *explanation task* with a slight increase in complexity compared to the first writing task.

Students in Experimental Condition 2 were tasked with writing a persuasive text, convincing their "friend" that RCTs should be the gold standard for research design in the field of psychology, similarly constituting a more complex task compared to the first writing task.

After the second writing task, students in both experimental groups filled out the rating instrument again. During all of the writing tasks, students in both experimental groups had access to the intervention text and could, therefore, reread it if they wished. Access was provided as part of the writing task instructions: “If you want, you can access the text you just read here,” with a link to the text attached to the word “here”.

Comparison Condition

During the two experimental writing tasks, students in the comparison condition engaged in two tasks that involved re-reading the initially assigned text on the topic of research design. During the first re-reading task, students were asked to highlight key terms in the text; during the second task, they were asked to highlight key sentences that contributed to their understanding.

Like students in the experimental conditions, students in the comparison condition filled out the rating instrument again after every task (see Table 3).

Table 3

Schematic Overview of the Intervention

Timing	Duration in mins.	**Comparison Condition**	**Experimental Condition 1**	**Experimental Condition 2**
Session 1	10–15	Demographics Survey Pretest Concept Map Rating Instrument Time 1		
	10–20	**Reading Task**		
	2–5	Rating Instrument Time 2		
	10–20	**Rereading Task** ***Key Terms***	**Writing Task** ***Free Write***	**Writing Task** ***Free Write***
	2–5	Rating Instrument Time 3		
	10–20	**Rereading Task** ***Key Inferences***	**Writing Task** ***Explanation***	**Writing Task** ***Persuasion***
	10–15	Rating Instrument Time 4 Posttest Concept Map		
Session 2	15–20	**Transfer-Test Writing Task** ***Argument***		
	2–5	Rating Instrument Time 5		

Posttests

Subsequently, students in all conditions were asked to create a posttest concept map about the topic at hand, visually outlining everything they know about what makes a good research design.

To prevent or reduce cognitive (over)load (e.g., Kirschner, 2002) and information saturation (e.g., Alexander, 2018), all students were assigned a final writing task (a transfer test) *one week after finishing the posttest*, asking them to write a response to the question "What type(s) of research design would work best for studying the effects of masking on COVID-19 transmission?”. In doing so, students had to synthesize and apply learned information and form a nuanced position, constituting the most complex of the four writing tasks (i.e., ranging from *free write* to *explanation*, *persuasion,* and *argument*).

Finally, students in all conditions filled out the rating instrument one more time, resulting in five sets of ratings in total.

Data Analysis

Scoring Protocols

Concept Maps

To score and analyze the pretest and posttest concept maps, I used an adaptation of Novak and Gowin’s (1984) scoring method, one that has been used as a reliable method of extracting knowledge scores from concept maps and of assessing change in individuals’ knowledge structures and meaningful learning (e.g., Hung & Lin, 2015; Watson et al., 2016). Although there are several scoring methods presented in the literature (e.g., holistic scoring: Besterfield-Sacre et al., 2004; categorical scoring: Segalàs et al., 2010), the knowledge *breadth*, *depth*, and *connectedness* components of Novak and Gowin’s scoring method seemed to fit best with the present study’s purpose and the role of the concept maps in the study—namely as a pretest and posttest measure of (the structure of) knowledge on a particular topic. Novak and Gowin’s (1984) method consists of several steps that can be summarized in the following computation: (NC – NCL) + (HH) * 5 + (NCL) * 10.

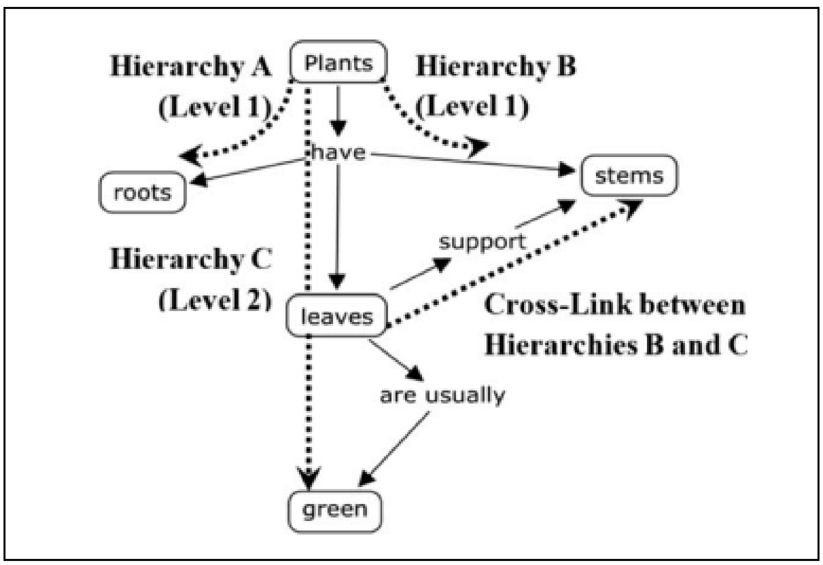
First, NC stands for *number of concepts* and refers to all propositions included in the concept maps (i.e., a concept with corresponding, annotated arrow). These are counted to indicate breadth of knowledge about the topic.

Second, HH stands for *highest level of hierarchy* and refers to the number of concepts in the longest path of propositions that include the concept map topic (Novak & Gowin, 1984; Watson et al., 2016).

Finally, NCL stands for *number of cross-links*, and refers to links between concepts from different hierarchies that create propositions. These are counted to indicate knowledge connectedness. The computation (NC – NCL) is included so that cross-linked concepts are not counted twice. For an example of a concept map illustrating these terms, see Figure 12.

Figure 12

Example of a Concept Map with Explanation of Terms



Note. From Watson et al. (2016)

After reviewing the pretest and posttest data, I observed that students varied widely in their approach to creating the concept maps. Specifically, numerous students failed to annotate the arrows leading to and from the concepts they included. This was true for more than half of the pretest concept maps, thereby violating the first category to be counted in the scoring scheme: propositions. To be able to conduct a meaningful analysis of the concept maps, I made the decision to adapt Novak and Gowin's (1984) formula slightly, resulting in (NC + (NAA* .5) – NCL) + (NCL) * 10 + (HH) * 5, in which NAA stood for Number of Annotated Arrows. In using the formula this way, I counted the number of concepts, with an additional 0.5 awarded for any annotated arrows leading to or from those concepts.

Additionally, to control for the quality and accuracy of the concepts included in the concept maps and not unduly inflate scores, as when a student may receive a high score simply for including multiple unrelated concepts, I generated lists of key concepts from the intervention text using ChatGPt and Gemini (see Appendix III). I reviewed the lists for quality and kept the most accurate concepts from both. Subsequently, as I went through the training for scoring the concept maps, I generated synonyms for many of the concepts. Ultimately, I only awarded points for those concepts that were on the list of key concepts or could conceivably be a synonym of one of those concepts. In this way, I could more accurately assess change in knowledge structure from pretest to posttest and control for any extraneous concepts included by the students.

To establish interrater reliability, I first went through training sessions with a fellow graduate student who had a background in science learning and linguistics. Throughout these early sessions, we scored a total of 19 concept maps independently, and came back together to talk through and resolve our disagreements. After these first 19 concept maps, I calculated intraclass correlations for each subscore as well as the final score and determined that these were sufficient enough (they ranged from .81 to .92) to start scoring a randomly drawn percentage of maps independently. I then randomly listed all response IDs from the sample and selected the first 20% from the pretest and the posttest separately, ascertaining that no response ID was present in both lists so as not to overrepresent any one participant's responses. After my colleague and I both scored these two sets of concept maps and I started calculating reliability scores, I had to remove one special case in which a student had created a directional flowchart that simply did not fit any of the rules and criteria discussed in our training sessions. I also excluded a few cases where a student technically submitted a concept map but did not actually provide any content. This resulted in a total of 26 pretests (17.7%) and 25 posttests (17.0%) scored for purposes of interrater reliability. Intraclass correlations are reported in Table 4 and ranged from 0.86 to 0.98. We resolved any discrepancies between our scores through discussion. The rest of the concept maps were scored by me, consulting about more difficult cases with the student who had gone through the training, as well as with my advisor.

Table 4

Intraclass Correlations for Concept Map Scoring (n = 51)

	Absolute Agreement	**Consistency**
Final Score	.944	.944
Number of Concepts (NC)	.976	.975
Number of Annotated Concepts (NAA)	.956	.956
Number of Crosslinks (NCL)	.907	.907
Highest Level of Hierarchy (HH)	.857	.857

Note. ICC analyses were two-way, mixed effects, based on single rater.

Transfer-Test Writing Samples

I scored the transfer-test writing samples for *evidence and structure of knowledge* using the Structure of Observed Learning Outcome (SOLO) Taxonomy (Biggs & Collis, 1982, 1989). This taxonomy represents a “learning cycle” (p. 152, 1989) and measures structure of knowledge at five different levels, based on “capacity (working memory and attention span), relating operations (how question and response interrelate), consistency and closure (amount of data used and openness of conclusion), and overall structure of the response” (Dinsmore & Alexander, 2016, pp. 223-224). Consequently, students’ writing samples were classified as one of these five levels and assigned scores ranging from 0 to 4, respectively: *prestructural*, *unidimensional*, *multistructural*, *relational*, or *extended abstract*. Following Dinsmore and Alexander (2016), I chose to include a separate score for transitional responses, awarded to those writing samples that fell in between two of the five levels (e.g., 2.5 if a writing sample fell between *multistructural* and *relational*). A detailed explanation of each level (and score) can be found in Table 5.

Table 5

SOLO Taxonomy

Taxonomy level	Score	Response characteristics
Prestructural	0	Cue and response undifferentiated No logical interrelation for cue and response High closure or low consistency Cue linked with irrelevant feature(s)
Unidimensional	1	Relate question with one piece of relevant data with a logical operation Drawing a conclusion from a particular instance Responses equally correct but inconsistent with each other One relevant feature to link question and response
Multistructural	2	Two or more relevant concepts or data Uses several features but does not link them Closure but lack of consistency Several relevant features link question and purpose
Relational	3	Response which interrelates multiple concepts Overall concept or principle accounting for data present Waits for all aspects before interrelating to make coherent whole Definite overgeneralized answer tied to concrete experience Uses relevant data in a conceptual scheme
Extended abstract	4	Give information comprehended in relevance to an overriding abstract principle True logical deduction Heavily qualifies set out principle to application in given situations Question left relatively open Relevant data with interrelations under hypothetical abstract structure with alternative outcomes and no definite closure
Transitional responses	.5; 1.5; 2.5; or 3.5	At a level of the taxonomy but marked by confusion or inconsistency Handles more information than able to cope with Loses track of the argument Forced to give up before reaching next SOLO level

Note. Based on Bigg & Collins (1982) and Dinsmore & Alexander (2016).

To establish interrater reliability, I worked with a scholar who had experience in using the SOLO Taxonomy. We undertook several training sessions in which we scored a total of 24 writing samples together and refined the scoring rubric with examples. Subsequently, we scored multiple subsets independently (for a total of 21), after which we reconvened to discuss and resolve our disagreements. At this point, I calculated median and average absolute deviation scores to determine our interrater reliability before scoring the remaining writing samples. For example, if three writing samples received a 2, 2.5, and 3 from one of us and 1, 3, and 2 from the other, this would result in absolute deviations of 1, 0.5, and 1, a median absolute deviation score of 1, and an average absolute deviation score of 0.83. Our median and average absolute deviation scores were 0.5 and 0.48, respectively, for this sample of 21. We determined that this was sufficient for both of us to score the remaining writing samples independently. After the scoring was finished, we re-calculated median and average absolute deviation scores (0 and 0.07, respectively) and resolved the few discrepancies between our scores through discussion.

Preliminary Analyses

Before running the primary analyses to answer the research questions, I ran several preliminary analyses, including a missing data analysis, general data inspection, checking of assumptions, and testing between-group equivalence, as well as an exploratory factor analysis, which I will describe first. All analyses were conducted in RStudio (Posit Team, 2024) and Numbers (Apple Inc., 2024).

Missing Data

After removing several cases from the dataset following criteria described in Chapter 3 (e.g., neither of the intervention surveys was fully completed and submitted or more than one of three knowledge measures were missing), a sample of n = 147 remained. Within this sample, there were several students who were missing *one* of the three knowledge measures (i.e., pretest, posttest, or transfer test), for a total of 44 cases. For two of these cases, I treated the transfer-test scores as missing because it became clear from the writing sample that the students had used OpenAI to generate their responses (see Appendix III).

Using a regression-based test (RB-test; Rouzinov & Berchtold, 2022), I established that these data were missing completely at random (MCAR) and subsequently ran a Multiple Imputation by Chained Equations (MICE; e.g., van Buuren & Groothuis-Oudshoon, 2011) analysis to impute the missing data values. Following the imputation procedure, I checked the correlations between the three knowledge measures for the original dataset and compared these to the same correlations for the imputed dataset and concluded that these were similar enough to warrant using the imputed dataset as the basis for subsequent analyses (see Appendix III for correlation tables and density plots).

Data Inspection and Assumptions

After the missing data analysis, I inspected the data more closely by looking at overall means and standard deviations, examining variable distributions, and checking between-group equivalence at the time of pretest. As I explain in more detail in Chapter 4, this analysis indicated that random assignment of participants to conditions failed to result in equivalent groups before the intervention. Therefore, in the analyses pertaining to Research Questions 1 and 3, I only compared the *explanation* and *persuasion* groups, both of which were considered experimental conditions.

For any ANOVA and multiple linear regression reported in subsequent sections, I checked the relevant assumptions, including normality, linearity, homogeneity of variance, and sphericity where necessary (i.e., in the case of mixed ANOVAs) and used corrections where needed. I also inspected the data for outliers and removed one case in which knowledge scores for pretest, posttest, and transfer test were abnormally high, thus skewing the data and corresponding data distributions.

Between-Group Equivalence

To check between-group equivalence at pretest, I examined the effect of condition and course section on pretest knowledge scores using three variations of the dataset: (a) the original dataset from which missing cases were omitted; (b) a dataset in which missing observations were imputed; and (c) the imputed dataset with the one observed outlier removed, which would be used in subsequent analyses.

The between-group equivalence tests yielded significant effects of condition on pretest for datasets (a), $F(2, 133) = 3.22$, $p = .043$ and (b), $F(2, 144) = 3.32$, $p = .039$, indicating a *failure of randomization*, despite the fact that students had been randomly assigned to conditions within each course section in which they were enrolled. Post-hoc independent t-tests of the analyses conducted using dataset (b) with Tukey-adjusted p-values confirmed the significant effect, with a mean difference of -4.98, 95% CI [0.36, 9.59], between the Comparison and Explanation conditions, $t(144) = -2.56$, $p = .03$. None of the other contrasts yielded significant results. Further, a similar test on (c) returned a non-significant effect of condition on pretest $F(2, 143) = 2.68$, $p = .072$ (see Table 6). However, because the observed p-value approached .05, I erred on the side of caution and maintained the failure of randomization conclusion.

A follow-up consideration of the relation between course section and pretest suggested that this was a wise course of action. Using the same datasets (a–c), I found that course section significantly predicted pretest for dataset (a), $F(4, 131) = 2.47$, $p = .048$, and (c), $F(4, 141) = 3.23$, $p = .014$. Post-hoc independent t-tests of the analyses conducted using dataset (c) with

Tukey-adjusted p-values did not confirm the significant effect for any of the contrasts between the different course sections, with p-values ranging from .067 to 1.00. Further, I found a non-significant effect for dataset (b), $F(4, 142) = 2.20$, $p = .072$ (see Table 7). However, the observed *p*-value again approached .05, which corroborated the earlier found relation between condition and pretest, suggesting that I could not assume group equivalence at pretest.

Table 6

Results of One-Way ANOVA: Pretest Knowledge Scores (Effect of Condition)

Source	**df**	**SS**	**MS**	***F***	***p***	**η^2**
Dataset (a)						
Condition	2	651	325.7	3.22	.043*	.05
Residuals	133	13453	101.2			
Dataset (b)						
Condition	2	647	323.5	3.315	.039*	.04
Residuals	144	14053	97.6			
Dataset (c)						
Condition	2	426	213.10	2.679	.072	.04
Residuals	143	11374	79.54			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2 = effect size (eta squared).

Table 7

Results of One-Way ANOVA: Pretest Knowledge Scores (Effect of Course Section)

Source	df	SS	MS	*F*	*p*	η^2
Dataset (a)						
Course Section	4	989	247.2	2.469	.048*	.07
Residuals	131	13116	100.1			
Dataset (b)						
Course Section	4	857	214.15	2.197	.072	.06
Residuals	142	13843	97.49			
Dataset (c)						
Course Section	4	991	247.69	3.231	.014*	.08
Residuals	141	10810	76.67			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2 = effect size (eta squared).

Given this initial non-equivalence, I explored plausible explanatory factors for the observed differences by examining student characteristics. As can be seen in Table 6, students from all course sections (i.e., EDHD200-0101-0102, EDHD200-0201, EDHD306-0101, EDHD306-0201, and PSYC300) were present in each of the conditions. The distribution of key demographic variables were also analyzed across conditions. However, results showed similar distributions of gender, age, racial/ethnic identity, class standing, learning disorders/disabilities, English as native language, and academic experience across the conditions.

Table 8

Course Enrollment Across Conditions

	***N* = 54**		***N* = 48**		***N* = 44**	
	Comparison Condition (Rereading)		Experimental Condition 1 (Explanation)		Experimental Condition 2 (Persuasion)	
Course Section	***N***	**%**	***N***	**%**	***N***	**%**
EDHD200-0101-0102	12	22.2%	7	14.6%	15	34.1%
EDHD200-0201	3	5.6%	8	16.7%	6	13.6%
EDHD306-0101	7	13.0%	3	6.3%	3	6.8%
EDHD306-0201	5	9.3%	7	14.6%	4	9.1%
PSYC300	27	50.0%	23	47.9%	16	36.4%

Additionally, using dataset (c) (i.e., the final dataset), I examined the effect of condition and course section on SKC ratings at pretest (i.e., Time 1) and on situational interest ratings at Time 2, given that situational interest was not measured at pretest. (Note that the SKC rating scores used in these analyses were derived from a factor analysis I report in the next section.) Results indicated that there was no significant effect of condition on SKC ratings at pretest, $F(2, 143) = 1.84$, $p = .162$ (see Table 9), nor a significant effect of condition on situational interest at Time 2, $F(2, 143) = 1.09$, $p = .34$ (see Table 10).

Finally, there was a significant effect of course section on SKC ratings at pretest, $F(4, 141) = 6.80$, $p <.001$ (see Table 11) and a significant effect of course section on situational interest at Time 2, $F(4, 141) = 5.11$, $p <.001$ (see Table 12). Post-hoc independent t-tests with Tukey-adjusted p-values confirmed significant differences in SKC ratings at pretest between the PSYC300 and EDHD200-0101-0102 course sections, $t(141) = 3.82$, $p = .002$, and between the PSYC300 and EDHD306-0201 course sections, $t(141) = 4.32$, $p = <.001$. They also confirmed significant differences in situational interest ratings at Time 2 between the PSYC300 and EDHD306-0201 course sections, $t(141) = 3.62$, $p = .004$ and between the PSYC300 and EDHD200-0201 course sections, $t(141) = 3.24$, $p = .012$. However, the importance of these differences at the course section level are assumed to be slight, given that every course section had a nearly equal distribution of participants across conditions (see Table 8).

Table 9

Results of One-Way ANOVA: SKC Factor at Pretest (Effect of Condition)

Source	**df**	**SS**	**MS**	***F***	***p***	**η^2**
Condition	2	3.61	1.806	1.841	.162	.03
Residuals	143	140.27	0.981			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2 = effect size (eta squared).

Table 10

Results of One-Way ANOVA: Situational Interest at Time 2 (Effect of Condition)

Source	**df**	**SS**	**MS**	***F***	***p***	**η^2**
Condition	2	2.18	1.088	1.088	.34	.01
Residuals	143	143.05	1.000			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2 = effect size (eta squared).

Table 11

Results of One-Way ANOVA: SKC Factor at Pretest (Effect of Course Section)

Source	df	SS	MS	*F*	*p*	η^2
Course Section	4	23.28	5.819	6.804	<.001*	.16
Residuals	141	120.60	0.855			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2 = effect size (eta squared).

Table 12

Results of One-Way ANOVA: Situational Interest at Time 2 (Effect of Course Section)

Source	df	SS	MS	*F*	*p*	η^2
Course Section	4	18.4	4.599	5.113	<.001*	.13
Residuals	141	126.8	0.900			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2 = effect size (eta squared).

Thus, because of the failure of randomization as indicated by the significant effects of condition and course enrollment on pretest knowledge scores (i.e., one of the variables at the heart of the intervention), I could no longer draw a meaningful comparison between the experimental and comparison conditions, examining the effect of writing beyond rereading. Therefore, for all subsequent analyses relating to Research Question 1 and 3, I only drew comparisons between the two experimental conditions (i.e., explanation and persuasion). Because Research Question 2 is of a slightly different nature (calculating change in calibration scores over time per condition), these analyses did include all conditions (i.e., comparison, explanation, and persuasion).

Scale Structure and Reduction

Finally, I conducted an exploratory factor analysis (EFA) using a principal component analysis (PCA) of the rating instrument items. This analysis was conducted for each of the five time points for subjective knowledge and confidence and for each of the four time points for which situational interest was measured. I report the sample size, number of items, variance explained, omega (total) score, and the range of factor loadings for each PCA in Table 13. Corresponding scree plots can be found in Appendix III.

At Time 1, measures only included *subjective knowledge* and *confidence* items for conceptual reasons (i.e., situational interest did not apply before the intervention began). As described in Chapter 3, these were nine items in total: (a) four items measuring subjective understanding, (b) four items measuring subjective organization, and (c) one item measuring confidence. The PCA on this first set of items indicated a clear one-factor solution, a pattern that remained throughout the analyses of the other four time points (see Table 13). In accordance with these results, I report *subjective knowledge* and *confidence* as one factor score (referred to hereafter as the *SKC* factor) in subsequent analyses and results.

Because situational interest is a conceptually distinct variable from subjective knowledge and confidence, I ran separate PCAs on those items for each of the four time points. The situational items consisted of eight items in total, four measuring maintained-SI-feeling, and four measuring maintained-SI-value. Starting at Time 2, the situational interest items also indicated a one-factor solution (see Table 13), which remained consistent for the remainder of the analyses.

Table 13

Results of PCAs: One-Factor Solutions

	N	# of items	Variance Explained	Omega (Total)	Factor Loading Range
Time 1					
Subjective Knowledge + Confidence	147	9	85%	.98	.88–.95
Time 2					
Subjective Knowledge + Confidence	147	9	85%	.98	.86–.94
Situational Interest	147	8	70%	.94	.65–.90
Time 3					
Subjective Knowledge + Confidence	147	9	91%	.99	.94–.96
Situational Interest	147	8	78%	.96	.73–.94
Time 4					
Subjective Knowledge + Confidence	147	9	90%	.99	.91–.96
Situational Interest	147	8	77%	.96	.73–.93
Time 5					
Subjective Knowledge + Confidence	132	9	91%	.99	.95–.97
Situational Interest	132	8	80%	.96	.76–.96

Primary Analyses

Mixed ANOVAs

To answer the first research question, I assessed changes in (a) SKC, the factor that combined subjective knowledge ratings and confidence ratings, (b) interest ratings, and (cc) pretest, posttest, and transfer-test knowledge scores as a function of writing at different time points in the intervention using *four two-way mixed ANOVAs*. In these analyses, the independent variables were *condition* (i.e., explanation and persuasion) and *time*. The time variable in the first two ANOVAs referred to the five rating instances for SKC and situational interest scores and referred to two and three time points for the ANOVAs dealing with knowledge scores (i.e., pretest/posttest and pretest/posttest/transfer test, respectively). The dependent variables were *SKC*, *situational interest*, and *knowledge* (twice; raw scores for the first and z-transformed scores for the second ANOVA), respectively.

After running these analyses, I assessed whether the residuals were approximately normally distributed and whether there were any outliers or influential cases that could be having an effect on the results. Given that there were three conditions and five time points in the study's design, I also checked the sphericity assumption—assessing the variances of the differences between conditions at different time points—using Mauchly's test (Mauchly, 1940). Whenever the sphericity assumption appeared to be violated, I used Greenhouse-Geisser (Greenhouse & Geisser, 1959) corrections to account for this and subsequently re-examined the results.

After running the ANOVAs, I used pairwise comparisons to shed light on the specific time points or groups for which there were statistically significant differences indicated.

Relative Accuracy (Calibration)

To answer Research Question 2, I assessed calibration by calculating relative accuracy between knowledge and subjective knowledge/confidence.

The first step in this analysis was to consider standardizing all relevant scores (i.e., subjective knowledge, confidence, and knowledge scores) to make them comparable. However, because the results from the EFA indicated a one-factor solution that combined subjective knowledge and confidence, I used the SKC factor scores here instead of a z-transformation to standardize each score. I did use z-transformations for the pretest, posttest, and transfer-test knowledge scores, to make sure all scales were comparable.

This complete set of SKC and knowledge scores then allowed me to create a meaningful *calculation of calibration* (i.e., comparing judgments of learning with performance as in Winne & Muis, 2011). For the purposes of the present study, I calculated *relative accuracy* (as opposed to absolute accuracy), that is, the "proportion of concordant or discordant judgments relative to total judgments" (Schraw et al., 2013, p. 50), by subtracting the SKC factor scores from the standardized knowledge scores. In this way, a resulting positive score indicated *overconfidence*, a negative score *underconfidence*, and a score close to zero pointed to good *calibration*.

I compared the standardized knowledge scores from the pretest and posttest concept maps as well as from the transfer-test writing samples with the SKC factor scores corresponding to each of these tasks, resulting in three separate comparisons, one for each timepoint.

Finally, after inspecting the data and checking relevant assumptions, I ran a *two-way mixed ANOVA* using the *calibration scores* resulting from the calculations described above as the dependent variable. This ANOVA provided insight into differences in calibration scores for any of the two experimental groups (i.e., explanation writing and persuasive writing) at any of the three time points (i.e., pretest, posttest, and transfer test). I again used pairwise comparisons to shed light on the specific time points or groups to which any significant differences pertained.

Multiple Linear Regressions

To answer the final research question, I ran two multiple linear regression models with post-intervention knowledge at posttest and transfer test as the analytic outcomes, respectively. Rather than using two "kitchen sink" models with all measured variables predicting the outcome variables, I opted to use conceptually more delimited models, using variables that were relevant predictors for posttest and transfer-test knowledge scores.

First, in the model predicting posttest knowledge scores, I used condition (i.e., explanation or persuasion), SKC and situational interest factor scores from Time 4 (i.e., at posttest), and pretest knowledge scores as predictor variables.

For the second model, predicting transfer-test knowledge scores, I used similar predictor variables: condition (i.e., explanation and persuasion), SKC and situational interest factor scores from Time 5 (i.e., at transfer test), and posttest knowledge scores.

CHAPTER 4: RESULTS

Descriptive Statistics

Before reporting the results of the primary analyses of this study, I report several descriptive statistics that contextualize the forthcoming results.

Means and Standard Deviations

The means and standard deviations of the main variable of interest, knowledge at pretest and posttest are displayed in Table 14. As displayed in that table and even more clearly in Figure 13, knowledge scores increased from pretest to posttest for all groups. Because the pretest and posttest were scored on a different scale than the transfer test, the latter is not displayed in the figure.

Table 14

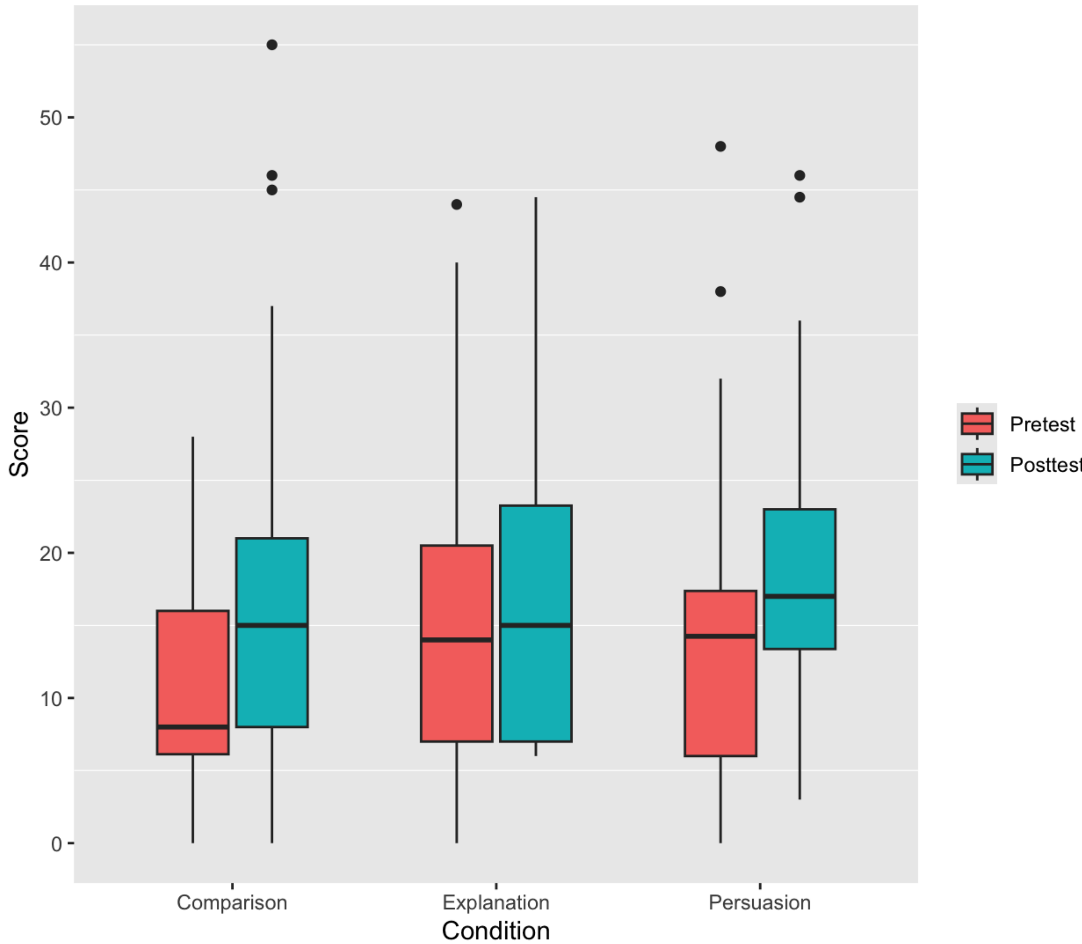
Means and Standard Deviations of Knowledge Scores

		Pretest	**Posttest**	**Transfer Test**
	N	**M *(SD)***	**M *(SD)***	**M *(SD)***
All Conditions	146	12.97 *(9.02)*	17.55 *(10.69)*	2.23 *(0.88)*
Comparison Condition (Rereading Group)	54	10.80 *(7.12)*	17.06 *(11.27)*	2.24 *(0.89)*
Experimental Condition 1 (Explanation Group)	48	14.17 *(9.58)*	17.29 *(11)*	2.23 *(0.95)*
Experimental Condition 2 (Persuasion Group)	44	13.74 *(10.08)*	18.43 *(9.79)*	2.41 *(0.82)*

Apart from demonstrating increase in knowledge, Figure 13 provides additional illustration of the failure of randomization described earlier, as shown by the starting point and trajectory of the comparison condition.

Figure 13

Knowledge Scores Per Condition at Pretest and Posttest



Correlations

In addition to overall means and standard deviations, Table 15 displays correlations between the knowledge measures across time points. Correlations between the main variables (i.e., knowledge, SKC and situational interest) across time points can be found in Table 16. Note that where variables were not measured, as, for example, situational interest at Time 1, no correlation is reported.

For calculations that involved the rating instrument (i.e., measuring SKC and situational interest) at Time 5, 15 cases were excluded because of missing data.

Table 15

Correlations Between Knowledge Measures

Time Point	Pretest	Posttest	Transfer Test
Pretest	1.00		
Posttest	0.47	1.00	
Transfer Test	0.27	0.24	1.00

Table 16

Correlations between Variables by Time Point

Variables	Time 1 (Pretest)	Time 2	Time 3	Time 4 (Posttest)	Time 5 (Transfer Test)
Knowledge x Subjective Knowledge	0.35	-	-	0.19	0.19
Knowledge x Situational Interest	-	-	-	0.04	0.13
Subjective Knowledge x Situational Interest	-	0.57	0.64	0.61	0.73

Main Findings

Research Question 1: Relating Ratings of Subjective Knowledge, Confidence, and Situational Interest to Knowledge Scores

As described in the Data Analysis section of Chapter 3, my first decision was to combine subjective knowledge and confidence ratings into a single factor (i.e., the SKC factor). Additionally, results of the between-group equivalence test led me to eliminate the comparison condition from analyses testing the first research question. Thus, my research question became: How do students' SKC ratings, situational interest ratings, and knowledge scores change as a function of explanation and persuasion writing interventions?

I ran four separate two-way, mixed ANOVAs to answer this question, with *condition* (i.e., the two experimental conditions) as the between-subjects variable and *time* as the within-subjects variable for each analysis.

The dependent variable I examined in the first ANOVA was the SKC factor across five time points. The results of the ANOVA indicated no significant main effects for condition or time, nor an interaction effect between the two (*p*s > 0.05; see Table 17).

In the second ANOVA, I examined situational interest as the dependent variable across four time points. Again, results indicated no significant main effects for condition or time, nor an interaction between the two (*p*s > 0.05; see Table 18).

Table 17

Results of Two-Way, Mixed ANOVA: SKC Ratings

Source	**SS**	**df**	**MS**	***F***	***p***	**η^2_p**
(Intercept)	2.006	1	2.006	0.460	0.499	
Condition	0.344	1	0.344	0.079	0.779	<.001
Error (Condition)	361.30	83	4.35			
Time	0.319	4	0.080	0.454	0.695[1]	<.001
Condition * Time	0.716	4	0.179	1.018	0.380[1]	.01
Error (Time)	58.39	332	0.176			

[1]For these p-values, Greenhouse-Geisser corrections were used to mitigate violating the sphericity assumption. *Note.* SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2_p = effect size (partial eta squared).

Table 18

Results of Two-Way, Mixed ANOVA: Situational Interest Ratings

Source	**SS**	**df**	**MS**	***F***	***p***	**η^2_p**
(Intercept)	1.949	1	1.949	0.479	.491	
Condition	0.274	1	.274	0.068	.796	<.001
Error (Condition)	337.92	83	4.071			
Time	0.028	3	0.009	0.099	.931[1]	<.001
Condition * Time	0.247	3	0.082	0.843	.447[1]	.01
Error (Time)	24.36	249	.098			

[1]For these p-values, Greenhouse-Geisser corrections were used to mitigate violating the sphericity assumption. *Note.* SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2_p = effect size (partial eta squared).

I examined knowledge scores as the dependent variable across two time points (i.e., pretest and posttest) in the third ANOVA. Results indicated a main effect for time (see Table 19). A post hoc pairwise t-test confirmed the effect, with a mean difference of 3.59, 95% CI [1.39, 5.79], between the posttest and the pretest, $t(91) = 3.24$, $p = .002$, $\eta^2_p = .11$. Results indicated no other significant effects.

Because the transfer test writing sample was scored on a different scale than the pretest and posttest concept maps, I ran a final ANOVA using the standardized knowledge scores (i.e., using z-transformation as described in Data Analysis for Research Question 2) to examine potential effects of the intervention on the three conditions across the three time points (i.e., pretest, posttest, and transfer test). The results of this model indicated no significant main or interaction effects (ps > .05, see Table 20). However, spaghetti plots across the three time points for each condition show that there is movement between the time points (see Figure 14), and this movement seems more dynamic for the experimental conditions (i.e., the movement within the comparison condition seems to be more homogeneous).

Table 19

Results of Two-Way, Mixed ANOVA: (Pretest-Posttest) Knowledge Scores

Source	**SS**	**df**	**MS**	***F***	***p***	**η^2_p**
(Intercept)	47266	1	47266	317.285	<.001	
Condition	0	1	0	0.002	.962	<.001
Error (Condition)	13407.2	90	148.969			
Time	608	1	608	10.744	**.001***	**.11**
Condition * Time	51	1	51	0.903	.344	<.001
Error (Time)	5091.1	90	56.568			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2_p = effect size (partial eta squared).

Table 20

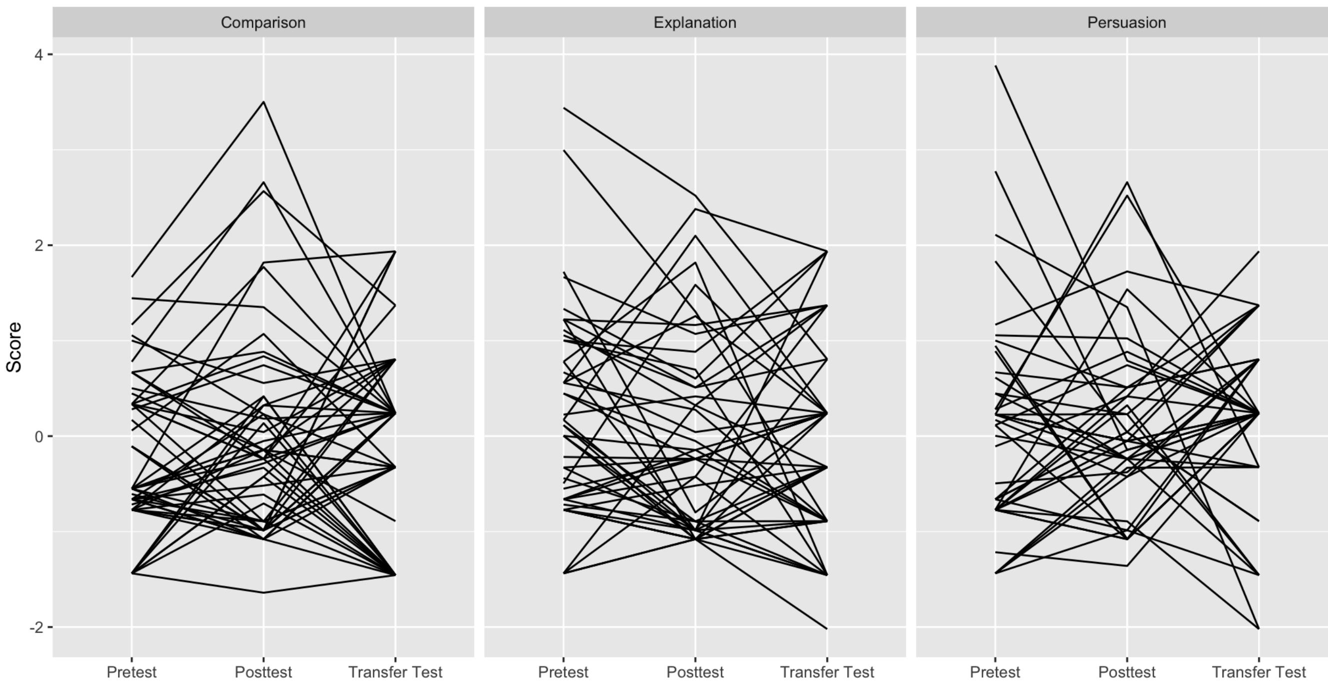
Results of Two-Way, Mixed ANOVA: Standardized Knowledge Scores

Source	**SS**	**df**	**MS**	***F***	***p***	**η^2_p**
(Intercept)	1.273	1	1.273	0.745	.390	
Condition	0.314	1	0.314	0.184	.669	<.001
Error (Condition)	153.71	90	1.708			
Time	0.697	2	0.349	0.486	.616	<.001
Condition * Time	1.162	2	0.581	0.811	.446	<.001
Error (Time)	129.05	180	0.717			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2_p = effect size (partial eta squared).

Figure 14

Trajectories from Pretest to Posttest to Transfer Test for Each Condition (Standardized Knowledge Scores)

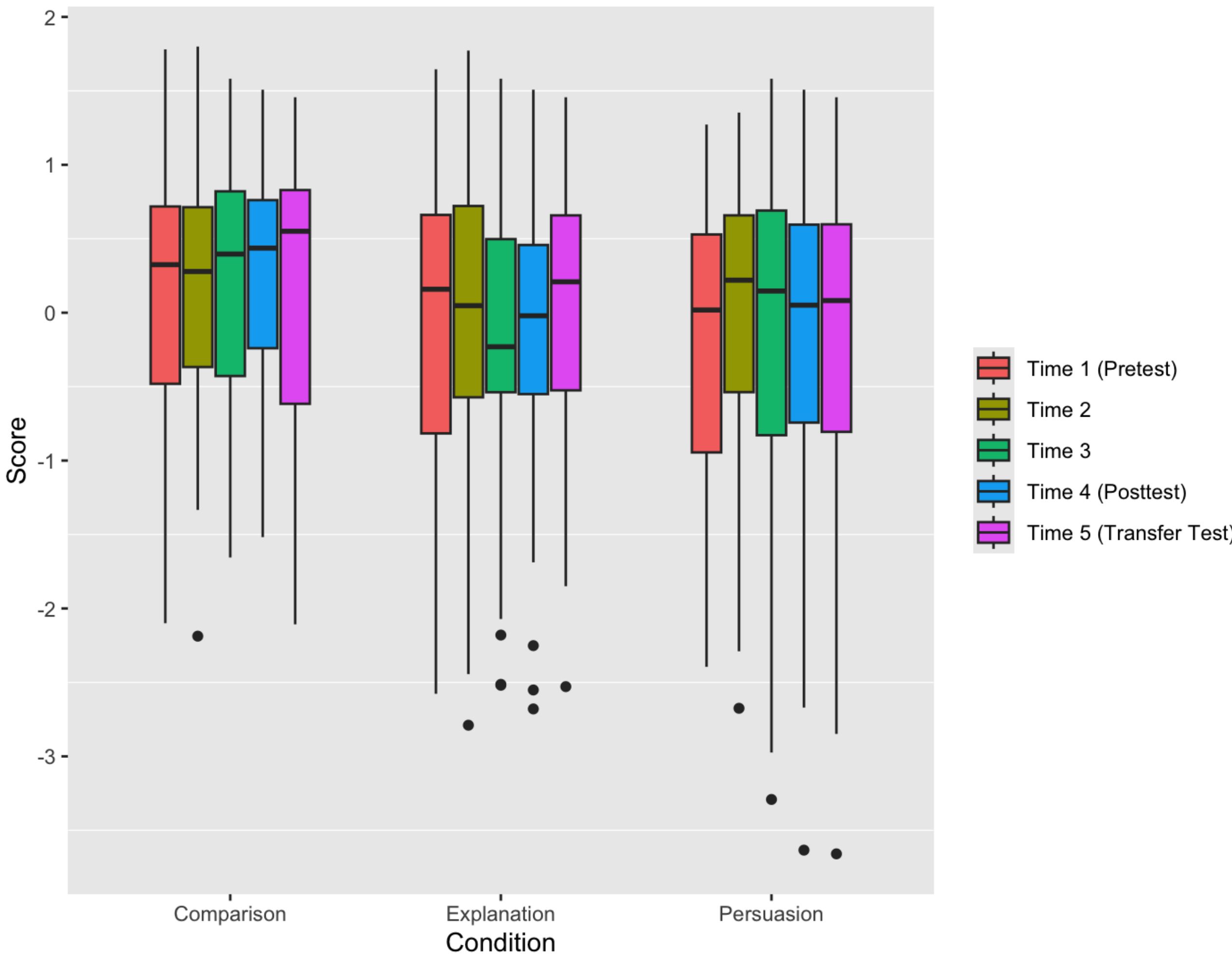


Research Question 2: Relating Subjective Knowledge and Confidence Ratings to Measures of Knowledge

The initial analysis leading to combining the subjective knowledge and confidence ratings into one SKC factor necessitated the re-writing of the second research question into: How do students' SKC ratings compare to their actual knowledge scores? As can be seen in Figure 15, SKC ratings followed a U-shaped trajectory for students in all conditions, decreasing initially and increasing in later tests of the intervention, with a more visible dip for the explanation group.

Figure 15

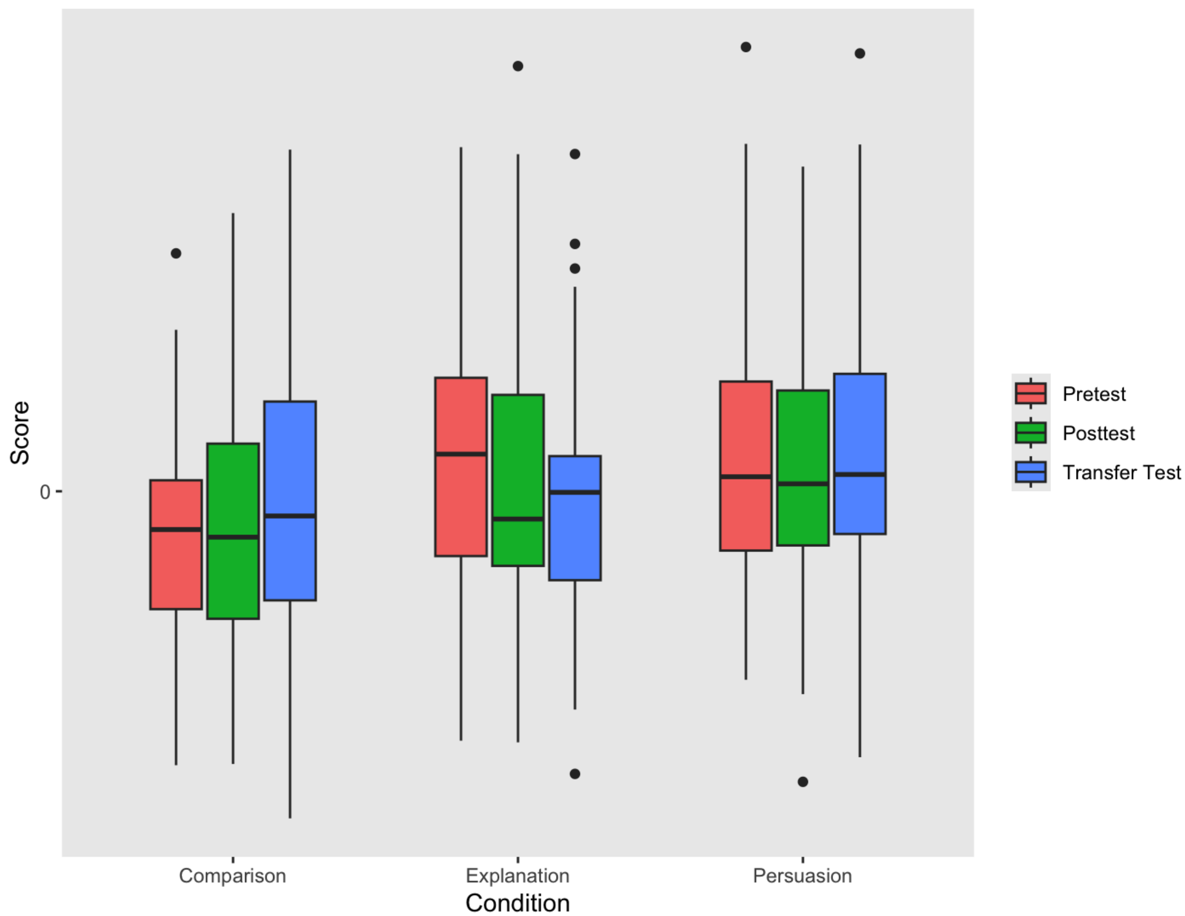
SKC Factor Scores Per Condition Over Time



The overall trajectory of students' calibration scores per condition and over time can be seen in Figure 16. A score above zero indicates overconfidence (i.e., confidence scores that are higher than actual knowledge scores), a score below zero indicates underconfidence (i.e., confidence scores that are lower than actual knowledge scores), and a score close to zero indicates good calibration (i.e., confidence and actual knowledge scores that match). Similar to the trajectory of the SKC factor overall as described in the previous section, students in all conditions seemed to become less confident after the pretest, and in some cases even *under*confident, whereas confidence seemed to increase and calibration seemed to get better toward and after the posttest.

Figure 16

Calibration Scores Per Condition Over Time



The two-way, mixed ANOVA I conducted to examine differences between conditions over time indicated a significant effect of *condition* (see Table 21). Post hoc independent t-tests (with a Bonferroni adjustment) indicated a significant difference between the rereading and persuasion groups (p = .002, η^2_p = .05). Figure 15 confirms that students in the persuasion group became significantly better calibrated compared to students in the rereading group.

Table 21

Results of Two-Way, Mixed ANOVA: Calibration Scores

Source	**SS**	**df**	**MS**	***F***	***p***	**η^2_p**
(Intercept)	0.016	1	0.016	0.006	.938	
Condition	16.910	2	8.455	3.134	**.047***	**.05**
Error (Condition)	345.31	128	2.698			
Time	0.014	2	0.007	0.009	.992	<.001
Condition * Time	2.824	4	0.706	0.877	.478	.01
Error (Time)	207.34	256	0.810			

Note. SS = sum of squares; df = degrees of freedom; MS = mean squares; *F* = F-statistic; *p* = p-value, η^2_p = effect size (partial eta squared).

Research Question 3: Predicting Students' Post-Intervention Knowledge Scores

Again, combining subjective knowledge and confidence ratings into one SKC factor led me to reword the third research question to the following: Do SKC and situational interest ratings as well as condition predict students' post-intervention knowledge scores? To answer this question, I ran the two multiple regression models described in Chapter 3, predicting posttest and transfer-test knowledge scores, respectively.

For the first regression model, I tested *pretest* (measured at Time 1), *condition* (with the explanation condition as the referent group), *SKC* at Time 4, and *situational interest* at Time 4 as predictors of the posttest knowledge scores (also measured at Time 4). Note that I used the Time 4 SKC and situational interest ratings in this regression because these would likely be most influential in accounting for variability in Time 4 knowledge scores. Additionally, adding the Time 2 and Time 3 SKC and situational interest scores would have added redundant predictors, leading to possible multicollinearity problems, as the time point means of both measures were highly intercorrelated (ranging from 0.71 to 0.95 for SKC and 0.86 to 0.94 for situational interest, see Table 22 and Table 23, respectively).

Table 22

Correlations Between SKC Ratings at All Time Points

	Time 1	Time 2	Time 3	Time 4	Time 5
Time 1	1.00				
Time 2	0.88	1.00			
Time 3	0.85	0.95	1.00		
Time 4	0.79	0.90	0.94	1.00	
Time 5	0.71	0.76	0.82	0.83	1.00

Table 23

Correlations Between Situational Interest Ratings at All Time Points

	Time 2	Time 3	Time 4	Time 5
Time 2	1.00			
Time 3	0.94	1.00		
Time 4	0.93	0.93	1.00	
Time 5	0.86	0.89	0.90	1.00

Results of this first model indicated an overall R^2 of .23, meaning the model accounted for 23% of the variance in posttest knowledge scores. In addition, the model was statistically significant, $F(4, 87) = 6.56$, $p < .01$, suggesting that the predictors collectively explained a significant proportion of the variance in posttest knowledge scores. In this model, pretest knowledge scores were a significant predictor that accounted for 18.7% of the variance in posttest knowledge scores ($p < 0.05$).

Similar to the rationale for how the first regression model was constructed, the second regression model tested *posttest* (a Time 4 knowledge measure), *condition*, *SKC* at Time 5, and *situational interest* at Time 5 as predictors of the transfer-test knowledge scores (measured at Time 5). Results indicated an overall R^2 of .16, meaning the model accounted for 16% of the variance in transfer-test knowledge scores. Again, the model was statistically significant, $F(5, 79) = 2.99$, $p < .05$, suggesting that the predictors collectively explained a significant proportion of the variance in transfer-test knowledge scores. In this model, posttest knowledge scores and the Time 5 SKC ratings were significant predictors that accounted for 6.08% and 6.02% of the variance in transfer-test knowledge scores, respectively ($p < 0.05$).

Summary of Findings

In sum, the results indicated that the two subscales of subjective knowledge ratings (i.e., subjective understanding and subjective organization; four items per subscale) and the 1-item confidence rating needed to be combined into one factor (SKC) and treated as such in all analyses due to collinearity. Second, the two subscales of situational interest (i.e., maintained-SI-feeling and maintained-SI-value; again, four items per subscale) were combined into one factor and treated as such in all analyses. Third, due to the failure of randomization, the comparison condition was removed from the analyses concerning Research Questions 1 and 3. Instead, I drew comparisons between the two experimental writing conditions to determine whether particular writing tasks may have a differential effect on the outcomes of interest. Fourth, neither SKC nor situational interest ratings changed significantly over time across the two experimental conditions, but knowledge scores did increase significantly for both conditions from pretest to posttest. Fifth, there was a significant difference in calibration scores between the persuasion and comparison group regardless of time, meaning students in the persuasion group were better calibrated than students in the comparison group. Finally, for the experimental groups, pretest knowledge scores significantly predicted posttest knowledge scores, and posttest knowledge scores and Time 5 SKC scores (i.e., the measure of confidence in one's own knowledge) significantly predicted transfer-test knowledge scores.

CHAPTER 5: DISCUSSION

This study consisted of an intervention primarily based on work by Galbraith and colleagues (e.g., 2018, 2023). It was set in the context of coursework and topics student participants actually needed to learn as part of existing curricula, meaning the study had an unusual level of ecological validity, bearing important consequences for the implications of the findings. The main findings of the study included (a) a positive, significant effect of the intervention on learning about the topic of research design for students in both experimental conditions (i.e., explanation and persuasion), (b) significantly better calibration on the part of students in the persuasion group as compared to the rereading group, and (c) a predictive relationship between SKC ratings at posttest and transfer-test knowledge scores for both the explanation and the persuasion groups, indicating an improved relationship between confidence and actual knowledge levels.

In this chapter, I first describe the limitations of the present study and how these may have impacted the results and any subsequent conclusions. Then, I discuss each of the main findings of this study and how they should be positioned within the existing literature on writing and learning through writing. Finally, I discuss the implications of the findings for both research and practice.

Limitations

The first limitation is in regard to the sample selected. The students from the sample were from one relatively selective research-intensive public university in the mid-Atlantic U.S. Admitted students have an average high school GPA above 4.0 (4.45), indicating that students have a strong history of Advanced Placement (AP) and honors courses as these courses contribute additional value to GPAs (Office of Institutional Research, Planning & Assessment, 2023). It is possible that a sample representing a wider variety of preparation for academic writing, as one would find in for instance a community college setting, would benefit more from the intervention, both in terms of knowledge and in terms of the confidence and situational interest factors. Alternatively, the effectiveness of this intervention could hinge on students being prepared to benefit from it, as, for example, by possessing a certain level of knowledge or confidence at the outset. Exploring this avenue of research would contribute to a deeper understanding of individual difference factors that may constrain the conditions under which this intervention would be most beneficial.

Further, these students were selected from methodology courses in their respective programs; however, the topics selected for this study are frequently introduced in foundational courses that often precede these courses, suggesting that participants may have already had a strong knowledge base of the topic. Even though a strength of the present study lies in the fact that the sample included students from five different course sections related to research methodology, ideally I would have recruited students with even more varying backgrounds, including those with potentially more limited knowledge of the topic in question.

Second, the intervention was relatively brief (approximately one hour) and was taken by most students as an extra-credit assignment. Ideally, an intervention such as the present one would take place over several weeks and would be incorporated even more than it was into required coursework. This would likely result in more significant and long-lasting changes and may negate any effects of recruiting a specific group of students that is interested in extra-credit points (e.g., students who want to get the highest grade possible or students who need the extra-credit points to get a passing grade).

Third, as described in preceding chapters, there was a failure of randomization despite taking all precautionary measures and implementing randomization procedures correctly. This made it difficult to draw conclusions about the effects of writing above and beyond rereading. To correct for the fact that the conditions were not equivalent at pretest, some researchers have proposed calculating gain scores (e.g., subtracting posttest from pretest scores). However, doing so potentially conflates error of measurement and would have caused problems for correctly interpreting the results (e.g., Forstmeier et al., 2017). Therefore, I erred on the side of caution and chose to exclude the comparison group from some of the main analyses, instead.

Fourth, there are some limitations that pertain to the pretest and posttest concept maps. There appeared to be a bug in the software students were using to create their pretest and posttest concept maps, which caused them to be able to return to and submit the intervention survey without uploading their concept map(s). This resulted in incomplete or missing data for quite a few students and led me to having to exclude a number of cases from the final sample.

A final limitation pertains to the intervention tasks. These tasks, however brief in terms of interventions generally, amounted to approximately one hour. This may have meant that students' attention span and cognitive energy may have waned toward the end of said hour. After about 50 minutes of the intervention tasks, possibly at the point where they were quite fatigued, students were asked to create their posttest concept map, a measure of knowledge that was vital to the study. To mitigate this possible fatigue at posttest, I could have presented students with their pretest concept maps and asked them to edit and adjust the map based on what they had learned, as opposed to starting anew.

These limitations should be kept in mind when considering the interpretations presented in the following sections, as well as the applicability of the findings.

Interpreting Results by Research Question

In this section, I will discuss the findings of the preliminary and primary analyses by research question in light of the existing literature.

Results of the Preliminary Analyses

The preliminary exploratory factor analysis I conducted on the subjective knowledge and confidence items resulted in a one-factor solution. One interpretation of these data is that subjective knowledge and confidence are tapping into the same underlying dimension or construct, as opposed to the two constructs proposed by Galbraith et al. (e.g., 2018; i.e., knowledge transformation and knowledge organization) that together formed his construct of subjective knowledge. This consequently means that the subjective knowledge instrument (Galbraith et al., 2023) may not be adding new dimensions to the existing literature on confidence and students' subjective knowledge state. In other words, those two dimensions do not appear to comport with sources of confidence found in previous literature (e.g., Dinsmore & Parkinson, 2013). In other words, the sources of confidence that students draw on did not appear to discriminate about the type or nature of knowledge as Galbraith had hypothesized, at least in this instance. However, multiple items can be used as evidence for reliability of confidence in future studies, rather than relying on a single item.

Additionally, the exploratory factor analysis on the situational interest items also resulted in a one-factor solution, as opposed to the two subscales distinguishing between two types of maintained situational interest focused on *maintained-feeling* and *maintained-value situational interest* as posited by Linnenbrink-Garcia et al. (2010). Similarly to how subjective knowledge and confidence behave, the two subscales seemed to tap into a singular *maintained situational interest* construct. As supported by interest researchers like Hidi and Renninger (2006), this finding is relatively unsurprising. In their work on the development of interest, they describe how maintained situational interest is a combination of feeling and value, but they do not make a sharp distinction between the two. Rather, they propose a much more intertwined and dynamic relationship in which, as triggered situational interest begins to develop and change into a more stable maintained form, valuing of the domain or topic of interest becomes more prominent and is added to the positive feelings that are strongly associated with the earliest phase of situational interest (e.g., Hidi & Renninger, 2006; Renninger, 2009). Because the intervention took place in research methods classes at a relatively introductory level of the topic, it is safe to say that most students were at the beginning stages of their maintained situational interest development (even if their triggered situational interest had already started to form). Following Hidi and Renninger's (2006) reasoning, at this stage, a distinction between feeling and value would be almost impossible to make, whereas at a later stage in their interest development, this distinction may have been a bit clearer.

Research Question 1: How do students' SKC ratings, situational interest ratings, and knowledge scores change as a function of explanation and persuasion writing interventions?

Subjective Knowledge/Confidence (SKC)

As described in Chapter 4, SKC ratings for students in all conditions followed a U-shaped trend, most visible in the explanation group. Even though the starting levels of confidence for the conditions were not the same, students in all conditions experienced a mean drop in confidence levels after the initial tasks, and gained some confidence toward the transfer test, which aligns with what I predicted in earlier chapters as well as with the literature on subjective knowledge, confidence, and calibration (e.g., Galbraith et al., 2023). Because there were no significant differences in SKC ratings between the conditions across time, this may mean that the rereading and writing tasks equivalently made students reconsider their confidence levels initially. However, there were some descriptive trends in SKC ratings within the conditions that are worth noting.

First, SKC ratings for students in the comparison condition began to rise immediately after Time 2, meaning they reported higher SKC ratings after they had read the intervention text twice. SKC ratings for these students continued to rise consistently after that. In contrast, the switch from decreasing to increasing confidence levels occurred later for students in the experimental conditions. In the explanation group, students' reported confidence levels kept decreasing until after they had completed their second writing task (i.e., the explanation task). Only then did their confidence levels start to increase and kept increasing until after the transfer test. Students in the persuasion group did not report an increase in confidence levels until after they had completed the posttest concept map. Follow-up analyses could test whether quadratic models would be more appropriate for examining these trends.

Based on the existing literature, a plausible explanation of these trends is that the nature of writing tasks is such that they make writers more actively wrestle with what they know and do not know and that, through writing about a topic, writers become more acutely aware of gaps in their knowledge base (Galbraith et al., 2023). Rereading, in contrast, seems to have a positive effect on subjective outcomes, such as students' confidence levels about their knowledge, although this confidence does not necessarily translate into substantive improvements in more objective outcomes, such as text comprehension (e.g., Phillips et al., 2016). This mismatch between confidence and actual knowledge is confirmed by the findings for Research Question 2, which demonstrate that the persuasion group was significantly better calibrated than the rereading group.

It seems that the persuasion task, in particular, had students wrestle with their knowledge base and their potential lack of knowledge about the topic. A possible explanation for this is the nature of persuasion tasks, which require students to provide claims and evidence and even include or rebut counterarguments in order to convince their reader of their standpoint, a task that students typically struggle more with than more straightforward narrative or expository writing tasks (e.g., Alexander et al., 2023; Wolfe & Britt, 2008). As supported by the literature on self-explanations (e.g., Fonseca & Chi, 2011), engaging in an explanation writing task may have made students in that condition feel more confident about their knowledge after (successfully) completing the explanation.

Situational Interest

Situational interest levels for students in the comparison group decreased over time, while it remained stable in the experimental groups, although a descriptive, non-significant upward trend appeared in those groups. This may indicate that writing tasks offer students something that rereading tasks do not, such as a more active engagement with the content, which contributes to maintained situational interest. On the other hand, students who engage in multiple rereadings of a text, thus being exposed to the same material multiple times, may hit a ceiling of engagement that results in lower interest ratings (e.g., Philips et al., 2016).

In sum, even though there were some descriptive trends, differences in levels of confidence or situational interest between conditions were not statistically significant over time. This may simply be attributed to the fact that the writing interventions were too brief to allow for a meaningful advantage when compared to the rereading condition. An alternate conclusion may be that writing genre (i.e., explanation or persuasion in this case) does not seem to have a differential effect on self-reported confidence and situational interest in writing interventions such as the one at hand. As noted by Galbraith and colleagues (2023), despite two meta-analyses on the effects of writing on learning (Graham et al., 2020; Klein, 1999), there have been few direct tests of different writing genres or approaches on learning. Thus, it is unclear whether effects shown in the literature are a reflection of generalized processes of expression (i.e. explanation; Fonseca & Chi, 2011) or stem from the demands of specific types of writing tasks. As described in more detail in Chapter 2, for the two studies they reported, they found that summary tasks seemed to have a different effect on subjective knowledge than argumentative tasks (Galbraith et al., 2023). The lack of a significant difference between the explanation and persuasion genres in the present study may be explained by the fact that these genres are different from the genres Galbraith et al. used, implying that certain writing genres do and other writing genres do not have a significant effect on confidence levels.

Knowledge

Notwithstanding the fact that there was no significant effect of condition on knowledge gain, there was a significant effect of *time*, meaning students in both writing conditions improved significantly from pretest to posttest. Because there was no difference between the two writing interventions nor an interaction effect of condition by time, it appears that writing genre (in particular, explanation or persuasion) did not have a significant effect on learning. However, the variance in knowledge scores for the persuasion group is much smaller than the same variance for the explanation group. This may mean that the persuasion writing tasks worked equivalently well for most students in the persuasion group. In contrast, the variance of the explanation group was larger, indicating that the explanation tasks worked better for some students compared to others in the same group. Thus, an educator may want to choose an intervention that benefits most students in a group, such as a persuasion writing task, rather than an intervention that seems to exacerbate differences among students, such as an explanation writing task, reminiscent of the criticism that certain interventions may aggravate achievement gaps among student subgroups (e.g., Dittman & Stephens, 2017).

Even though I could not draw a meaningful comparison between the writing groups and the rereading group regarding their knowledge scores, descriptive statistics indicate that the rereading group experienced knowledge gains, as well. This begs the question of what writing may offer students beyond rereading, a question about which I can only provide educated speculations, based on the existing literature.

Students in all three groups read the intervention text at Time 2. After this initial reading activity, students in the comparison group went on to reread the text at Times 3 and 4 (i.e., they read the same text three times in total). Meanwhile, students in both experimental groups, having read the text at Time 2, first completed a free-writing task at Time 3, after which the explanation group went on to write an explanatory text, whereas the persuasion group wrote a persuasive text. It seems noteworthy that the two experimental groups were showing improvement even though they had read the intervention text only once, suggesting the two writing tasks were at least as good, if not better, at getting students to learn more about the topic than the students who read the text three times (i.e., students in the comparison condition).

As demonstrated by the pervasiveness of reading tasks in college-level curricula as well as the literature on the benefits of reading and rereading for learning (e.g., Greenleaf et al., 2023; Lapp et al., 2016) and confirmed by the findings of the present study, reading and rereading offer advantages for learning about a topic. However, there are downsides to using reading as a main tool for learning, such as its impact on interest levels, as illustrated by the downward trend in situational interest present in the rereading group. Simply put, rereading may be beneficial for learning but may not be the most effective approach for increasing knowledge. The findings of the present study suggest, however, that the same thing may be said about writing. Like reading and rereading, writing tasks promote learning and learning-related outcomes (see Graham et al., 2020 for a review), some of which were measured in the present study. Additionally, however, writing may provide students with advantages that were not measured in the present study. These may include improved writing ability (e.g., Johnstone et al., 2002) or the targeting of higher-level, valuable learning outcomes that rereading tasks typically do not achieve, such as the ability to articulate one's thoughts and viewpoints and support claims with evidence (e.g., Boscolo et al., 2007). Further, the trends in the descriptive data of the present study show learning gains for all groups but higher means for the writing groups on the posttest and transfer test. Had the intervention continued over a longer period of time, the writing tasks in the experimental groups may have encouraged a way of approaching new content that rereading would not.

In sum, in college-level curricula that are often permeated by reading-intensive learning tasks, educators may have to consider how reading and rereading may set the stage for writing tasks, or, how writing tasks can help students approach content in a different way, after which they could return to the text to reread it. The short answer may be that, to meet both lower- and higher-level learning outcomes as well as consider students' interest levels, (re)reading and writing should take turns in a way that corresponds to the nature of the topic and type of new content.

Research Question 2: How do students' SKC ratings compare to their actual knowledge scores?

As described in previous chapters, calibration scores were calculated using confidence levels and knowledge scores at pretest, posttest, and the transfer test. As with the U-shaped trend of levels of confidence described in the previous section, calibration scores for each condition follow a similar pattern, although at different levels of calibration (i.e., relative accuracy between confidence levels and actual knowledge scores). The rereading group seemed to be generally underconfident, whereas the persuasion group seemed generally to be well-calibrated. The explanation group seemed to experience the most visible decrease, from being overconfident to being underconfident, and back to being overconfident again. Nonetheless, confidence levels went down and students in all conditions became less overconfident or more underconfident after the pretest. Subsequently, confidence levels went up and students in all conditions became better calibrated or more overconfident following the posttest. This pattern indicates that both rereading and writing tasks seemed to have an effect on confidence levels and overall calibration scores. In other words, students seemed to become less confident, and in some cases even underconfident, after being confronted with new content (through reading or writing), and more confident, and in some cases even overconfident, after completing the final writing task (i.e., the transfer test). In all conditions, students seemed to be better calibrated at that final time point.

There was a significant difference between the comparison group and the persuasion group, with the persuasion group being significantly better calibrated than the comparison group. Thus, it seems that engaging in writing tasks helps students more accurately assess their topic knowledge as compared to engaging in rereading tasks.

Research Question 3: Do SKC and situational interest ratings as well as condition predict students' post-intervention knowledge scores?

I ran two multiple linear regression models on the experimental conditions: one predicting posttest knowledge scores by pretest, condition, and SKC and situational interest ratings at posttest, and one predicting transfer-test knowledge scores by posttest, condition, and SKC and situational interest ratings at transfer test.

First, pretest knowledge scores were a significant predictor of posttest knowledge scores. This rather unsurprising finding is strongly supported by the literature on knowledge and learning. As described in the first chapters of this dissertation, previous learning is strongly related to new learning; knowledge acquired before engaging with new content is strongly related to new knowledge acquisition (e.g., Alexander & Murphy, 1998; Chi, 1985; Kintsch, 1998).

Second, posttest knowledge scores were a significant predictor of transfer-test knowledge scores. This means that, even though at least a week passed between the two measurement points and the nature of the transfer-test writing task was different from the posttest concept mapping task, those that had acquired more knowledge during the intervention phase held onto those knowledge gains at the transfer test. Additionally, this demonstrates a strong connection between the knowledge scores I extracted from the concept maps and the knowledge scores that resulted from the transfer-test writing tasks, regardless of the fact that these tasks were different in nature.

Finally, SKC ratings at transfer test were a significant predictor of transfer-test knowledge scores. In other words, confidence levels at this time point predicted students' actual knowledge scores. This means that, for both of the experimental groups, confidence levels at transfer test became more in line with their actual knowledge—a hint at becoming better calibrated. As mentioned in previous sections, the variance of the knowledge scores for the persuasion group was smaller than said variance for the explanation group, indicating that the persuasion intervention was effective for most students in that group. Combined with the fact that confidence levels predict knowledge scores at transfer test for both of the writing conditions and that the persuasion group was significantly better calibrated than the rereading group, this confirms that the nature of (especially persuasion) writing tasks seems to help students more accurately evaluate gaps in their knowledge base and match their confidence levels appropriately.

Implications of the Present Study

Implications for Theory and Research

First, the sample for the present study consisted of undergraduate students recruited from five different course sections related to research methodology. Even though some of the conclusions of the present study were limited because of the failure of randomization, the conclusions that were presented are more generalizable in nature than those of typical writing and learning interventions, many of which are based on a single class or a single case (e.g., Boscolo et al., 2007; Mateos et al., 2018; Raedts et al., 2007). Therefore, future research should continue the practice of trying to include students from multiple classrooms and contexts, as to increase the generalizability and applicability of findings.

Second, preliminary analyses for the present study included exploratory factor analyses on the rating instrument, which changed the way the main variables of the study were treated in subsequent analyses. To solidify the findings that subjective knowledge and confidence as well as maintained-feeling and maintained-value situational interest are tapping into a single construct instead of into multiple separate constructs, future research should include confirmatory factor analyses using rating instrument data from a different sample.

Third, in the present study, there was a significant effect of course enrollment on pretest scores, one of the indicators that randomization had failed. To be able to draw more substantial conclusions about the effect of writing on (new) knowledge development, follow-up studies could include multilevel modeling to test and explain the potential differential effects of the intervention on each course section across conditions, as well as possibly mitigate those effects in order to formulate more generalizable conclusions. In addition, content analysis of the experimental writing samples (i.e., the results of the free-write, explanation, and persuasion tasks) could substantiate any forthcoming conclusions by shedding more light on the specifics of the writers' knowledge development.

Finally, the results of the present study seem to be touching on multidimensional, dynamic processes and constructs. To be able to truly understand what is happening during writing tasks and who writing is beneficial for exactly, future studies would have to tap into even more learner characteristics than the present study did. These could include reading comprehension and writing ability, more in-depth information about motivational constructs such as self-efficacy, or students' performance in other writing-intensive undergraduate courses.

Implications for Practice

One of the big takeaways from the present study is that rereading and writing both help students to build knowledge. Rereading presents limitations for students' motivation, but may be enough for acclimating learners, whereas more advanced learners may need higher-level strategies, such as writing tasks, to advance their knowledge about a topic (e.g., Alexander, 1997; Dinsmore & Hattan, 2020). Using both rereading and writing to help students acquire knowledge provides them with an array of learning strategies. It is up to them and up to their educators to intuit which strategy is most effective when.

Concluding Remarks

In conclusion, even though the comparison condition (i.e., rereading) had to be removed from several important analyses because of a failure of randomization, the writing intervention central to the present study had a positive, significant effect on learning about the topic of research design for students in both experimental conditions (i.e., explanation and persuasion). Additionally, students in the persuasion group were significantly better calibrated than students in the rereading group, and SKC ratings at posttest were a significant predictor of transfer-test knowledge scores for both the explanation and the persuasion groups, indicating an improved relationship between confidence and actual knowledge levels. This test of effectiveness was conducted in settings where the tasks and topics were relevant to the existing coursework. Constraints from embracing an ecologically valid context notwithstanding, it would seem promising that findings indicated that a writing-based intervention in a classroom context contributed to learning and calibration outcomes.

Follow-up studies should include confirmatory factor analyses to further examine the structure of the SKC construct, multilevel modeling to investigate potential differential effects of the intervention on students from different course sections, and quadratic models to analyze patterns in SKC data.

The study's results underscore the importance of providing students with a range of learning strategies, including rereading *and* writing, to help them acquire knowledge. Educators can use this study's findings to inform their instructional decisions, recognizing that students' individual needs will vary and that a combination of strategies may be most effective in promoting knowledge development.

APPENDICES

Appendix I: Surveys and Instruments

Table 24

Demographics Survey

Variable	Item
Age	● How old are you?
Gender	● I would describe my gender identity as: ○ Man ○ Woman ○ Non-binary ○ Something else, feel free to specify: ○ Prefer not to say
Race/Ethnicity	● I would describe my race/ethnicity as: ☐ American Indian or Alaska Native ☐ Asian or Asian American ☐ Black or African American ☐ Hispanic or Latinx ☐ Middle Eastern or Northern African ☐ Native Hawaiian or Other Pacific Islander ☐ White ☐ Something else, please specify: ☐ Prefer not to say
Class Standing	● What is your class standing? ○ Freshman ○ Sophomore ○ Junior ○ Senior ○ Other, please specify:
Academic Major	● What is your academic major?
Language Proficiencies	● Is English your native language? ● If not, what is your native language?
Learning Disorders or Disabilities	● Do you have any of the following learning disorders or disabilities? ☐ Oral/Written Language Disorder ☐ Non-Verbal Learning Disability ☐ Dysgraphia

	☐ Dyscalculia ☐ Dyslexia ☐ AD(H)D ☐ Something else, please specify: ☐ None
Academic Writing Experience	• On a scale of 0 to 10, how much does your coursework generally require you to write?
Experience with the Intervention Topic	• On a scale of 0 to 10, how familiar are you with the topic of research design?

Table 25

Self-Rating Instrument (Subjective Knowledge, Confidence, Situational Interest)

Variable	Item
Introduction	*On a scale ranging from "strongly disagree" to "strongly agree", rate how much you agree with the following statements.*
Subjective Knowledge* *Subjective Understanding*	● I feel I know about research design. ● I understand research design well. ● I can explain research design well. ● I can make sense of research design issues.
Subjective Knowledge *Subjective Organization*	● My thoughts about research design are organized. ● My thoughts about research design are coherent. ● My thoughts about research design are structured. ● My thoughts about research design are well-ordered.
Confidence	● I am confident about my knowledge of research design.
Situational Interest** *Maintained-Feeling*	● What I am learning about research design is fascinating to me. ● I am excited about what I am learning about research design. ● I like what I am learning about research design. ● I find the activities we are doing about research design interesting.
Situational Interest *Maintained-Value*	● What I am studying about research design is useful for me to know. ● What I am studying about research design is important to me. ● What I am learning about research design can be applied to real life. ● I am learning valuable things about research design.

**Adapted from Galbraith et al.'s (2023) Subjective Knowledge Rating instrument*
***Adapted from Linnenbrink-Garcia et al.'s (2010) Situational Interest Survey*

Appendix II: Intervention Tasks and Materials

Table 26

Intervention Tasks

<table>
<tr><th>Task</th><th>Conditions</th><th>Prompt</th></tr>
<tr><td>Pretest Concept Map</td><td>All Conditions</td><td>“For at least the next 10 minutes, create a concept map in response to the following question: what makes a good research design?

A concept map is a visual representation of your knowledge about a topic, in this case "what makes a good research design?". For an example of a concept map, see below.”

----------------------Page Break-----------------------

“Create a concept map in response to the following question: What makes a good research design? In creating your concept map, try to be as thorough as possible and include as much information as you can.
Click the button below to start making your concept map
Make sure to hit the "Save & Exit" button at the top right of the screen when you have finished your concept map; this will send you back to the survey
If you want to go back to your concept map after returning to the survey, simply hit the "Continue editing your concept map" button
You are free to click "Continue" whenever your concept map is finished.”</td></tr>
<tr><td>Posttest Concept Map</td><td>All Conditions</td><td>“On the next page, you will be asked to create a SECOND concept map in response to the following question: what makes a good research design?

As a reminder, a concept map is a visual representation of your knowledge about a topic, in this case "what makes a good research design?". For an example of a concept map, see below.”</td></tr>
</table>

		----------------------Page Break----------------------- "For at least the next 10 minutes, **create a second concept map** in response to the following question: ***What makes a good research design?*** In creating your concept map, try to be as thorough as possible and include as much information as you can. ● Click the button below to start making your concept map ● Make sure to hit the "**Save & Exit**" button at the top right of the screen when you have finished your concept map; this will send you back to the survey ● If you want to go back to your concept map after returning to the survey, simply hit the "Continue editing your concept map" button ● You are free to click "Continue" whenever your concept map is finished."
Reading and Rereading Tasks		
Reading Core Text	All Conditions	"For at least the next 10 minutes, **read the text below** on research design to develop your understanding of this topic. After 10 minutes (see timer above), the "Continue" button will appear, but you are free to click it whenever you are finished after that."
Highlighting Key Terms	Comparison Condition	"For at least the next 10 minutes, re-read the text on research design. This time, **highlight key terms** while you are reading. ● Simply hover over a word and click on it ***once*** to highlight it. ● After 10 minutes (see timer above), the "Continue" button will appear, but you are free to click it whenever you are finished after that."
Highlighting Key Inferences	Comparison Condition	"For at least the next 10 minutes, re-read the text on research design. This time, while you are reading, **highlight any sentences or passages that**

		contribute(d) to your understanding of this topic. ● Simply hover over a word and click on it ***once*** to highlight it. ● After 10 minutes (see timer above), the "Continue" button will appear, but you are free to click it whenever you are finished after that.”
Writing Tasks		
Free Write	Experimental Conditions	“For at least the next 10 minutes, **write down everything you know** about what makes a good research design. ● After 10 minutes (see timer above), the "Continue" button will appear, but you are free to click it whenever you are finished after that. ● If you want, you can access the text you just read here.”
Explanation	Experimental Condition 1	“For at least the next 10 minutes, write a response to the following question: How would you **explain research design** in psychology to a friend who is trying to learn about this topic? ● After 10 minutes (see timer above), the "Continue" button will appear, but you are free to click it whenever you are finished after that. ● If you want, you can access the text you just read here.”
Persuasion	Experimental Condition 2	“For at least the next 10 minutes, write a response to the following prompt: **Convince** your friend that RCTs should be the gold standard for research design in psychology. ● After 10 minutes (see timer above), the "Continue" button will appear, but you are free to click it whenever you are finished after that. ● If you want, you can access the text you just read here.”

Argument (Transfer Test)	All Conditions	“For at least the next 15 minutes, write a response to the following question: **What type(s) of research design would work best** for studying the effects of masking on COVID-19 transmission? ● After 10 minutes (see timer above), the "Continue" button will appear, but you are free to click it whenever you are finished after that. ● If you want, you can access the text you just read here.”

Figure 17

Intervention Core Text

What Makes a Good Research Design?

In this text, you will first find an introduction to basic research design principles in the fields of psychology and human development. Then, you will read examples from the medical field to help you think about research design in different settings.

Introduction and Key Concepts

Correlations and Experiments
There are two general ways of studying relationships between variables. Correlational research is descriptive and non-manipulated, involving observation or measurement of variables as they naturally occur. Two problems with the correlational method are (1) determining the direction of cause and effect and (2) the possibility that extraneous variables influenced the results.

Experimental research involves manipulation of variables, control of the experimental situation, and random assignment of participants. Using the experimental approach, it is easier to infer that one variable caused the other. A major drawback of experimental studies is their artificiality, but this could be addressed by conducting experiments in field settings. The experimental approach cannot be used to answer many important questions because it would be unethical or impractical to manipulate variables. Experimental approaches are also not appropriate for description and prediction of behavior. Sometimes an experiment is the best approach; however, often a descriptive approach is preferable. Thus, there is no single best method for studying behavior.

Validity
The validity of a test or experiment refers to the degree to which it measures what it is intended to measure. Researchers strive for two types of validity: internal and external.

Internal validity refers to whether effects observed within experiments can be attributed with confidence to the factor that the researcher is testing. For instance, suppose that a researcher tests the effectiveness of a type of psychotherapy for depression by administering it to a number of depressed adolescents. If, three months later, many of the adolescents are no longer depressed, can it be concluded that this type of psychotherapy caused the improvement? No, because the students' recovery may have been due to the mere passage of time. Moods fluctuate, and many people who are depressed at any given time will be happier at a later date even without psychotherapy. In this example, the passage of time is a source of internal invalidity because the factor believed to cause the improvement (the psychotherapy) may have had no effect.
External validity, in contrast, refers to the ability to generalize research findings beyond the particulars of the research in question. Studies of child development are almost never intended to apply only to the particular children and research methods involved in a given study. Rather, the goal is to draw conclusions that apply to children more generally. Thus, the findings of a single experiment are only the first step in determining the external validity of the results.

Additional studies with participants from different backgrounds and with research methods that vary in their particulars are needed to establish the external validity of the findings.

RCTs vs. Observational Research: Examples from the Medical Field

Strengths and Limitations of RCTs

The strength of the randomized controlled trial (RCT) rests on its excellent internal validity, which is based largely on the power of randomization to ensure that the only difference between two conditions is their exposure to the treatment of interest. Although randomization minimizes the risk of bias by confounding, there are other biases inherent to RCTs that limit their applicability to the care of patients in routine practice. In particular, patients, providers, and concurrent care in the general population are different from those in clinical trials, and the generalizability (or external validity) of RCTs may be limited. Although population-based observational research does not enjoy the same level of internal validity as RCTs, well-designed observational studies can offer superior external validity and provide a unique opportunity to evaluate the uptake of new treatments and their outcomes in routine practice.

Most of the substantial improvements in treatment and outcome of patients with cancer over the past four decades have been identified in RCTs. However, patients are highly selected to participate in RCTs, and the greatest limiting factor in interpreting them is that patients seen in routine practice are very different from patients included in RCTs. This is not surprising given that <10% of patients with cancer are entered into a clinical trial. Patients with advanced age and greater comorbidity (coexisting or co-occurring conditions), and those from lower socioeconomic backgrounds are under-represented in RCTs. There can also be important differences in the provision of care for patients on RCTs (i.e., highly regulated trial protocols at specialized centers of excellence) compared with patients in routine practice. Given the greater toxicity that is expected when a treatment is applied to a non-selected population with greater comorbidity than subjects included in trials, a small increase in overall (or progression-free) survival observed in a large RCT is likely to disappear when some treatments are applied in routine practice. There is a major difference in generalizability between RCTs that report substantial gain and limited toxicity and RCTs evaluating drugs with marginal effects and substantial toxicity. RCTs reporting results of marginal clinical significance to the selected patients recruited to them can be very misleading.

Most oncologists and patients would define a treatment as having benefit if it allows patients to live longer, live better, or both. Unfortunately, a minority of new treatments evaluated in RCTs have achieved these goals. The marked increase in sample size of RCTs provides statistical power to detect treatment differences between conditions that are statistically significant but of marginal clinical relevance. RCTs evaluating treatments for cancer are reporting smaller incremental benefits than previously. There is increasing use of surrogate endpoints that have not been validated as predictive of improvement in duration or quality of survival and growing recognition that RCTs underestimate and under-report harms from new cancer therapies.

We and others have described suboptimal reporting of trial findings and various forms of bias associated with disseminating RCT results to practitioners that can adversely influence patient care. Failure to publish studies with negative results can influence results of meta-analyses and treatment guidelines through an imbalanced perspective of the benefits (or lack thereof) of new medical therapies; hopefully, mandatory trial registration will reduce this bias. Presentation of non-final analyses of RCTs at oncology meetings is commonplace, and in up to 10% of cases the study conclusions will alter substantially between conference presentation and publication of final results. Industry-funded RCTs are more likely to be reported as positive than those not sponsored by industry.

Strengths and Limitations of Population-Based Observational Research
Population-based observational studies differ from traditional institutional retrospective studies in that the former include all patients within a given jurisdiction and are therefore less prone to selection and referral biases that plague more traditional forms of observational research.

Large-scale studies have been enabled by advances in computer technology that allow interactions between databases, such as cancer registries or Surveillance, Epidemiology and End Results data, and hospital records. Population-based observational studies can provide information about rare diseases for which there are no RCTs. Furthermore, changes in the biology and epidemiology of cancer can be best described using observational research. Observational studies also provide insights into the care and outcomes of patients under-represented in RCTs, including the elderly and those with comorbidity, and patients from under-represented ethnic and socioeconomic backgrounds. Potential risk factors related to developing cancer and the prognostic significance of disease- related characteristics can also be described using observational data.

Given the differences between patients recruited to trials and those seen in routine practice, increased toxicity might be expected when the results of RCTs are applied to routine practice. Population-based studies can provide information about toxicity associated with treatment. Examples include the finding of increased cardiovascular disease and diabetes among men treated with androgen deprivation for localized prostate cancer, the risks of cardiac disease after radiotherapy for breast cancer, and long-term toxicities associated with treatment of testicular cancer. Despite compelling evidence from RCTs and published treatment guidelines, physicians may not adopt new medical therapies. Population-based studies can identify gaps in care following publication of pivotal RCTs. They have described under-utilization of adjuvant therapy for breast, colon, and non-small cell lung cancers despite publication of compelling RCT data and strong recommendations from treatment guidelines. Physicians may also overtreat patients and therefore expose them to risks and harms without meaningful chance of benefit. Observational data have quantified overtreatment of patients with early-stage prostate cancer and breast cancer. Observational data can also allow investigators and policy makers to evaluate knowledge translation in efforts to improve care and outcomes in the general population; for example, implementation of an audit and feedback tool led to improved nodal harvest in people with colorectal cancer. Moreover, population-based studies of health-system performance can inform policy and be used to improve access to care.

Many important clinical questions have not, cannot, and will not be ever addressed in the context of an RCT. In these situations, clinicians rely on information provided by observational research. Oncology practice and policy have been influenced by population-based studies showing that patient outcome is influenced by the interval between surgery and adjuvant chemotherapy for colorectal and breast cancer, hospital and surgeon volume of cancer surgery, and the extent of lymph node harvest in colorectal cancer.

Observational studies do have important limitations that must be carefully considered when evaluating treatment benefit. The most important limitation is in differentiating between outcomes that are due to adoption of a new treatment and those due to other unrecognized changes in the population under study. Factors that may not be identified or measurable using observational data include stage migration, changes in disease biology, changes in other aspects of management, and confounding by indication. Although statistical modeling techniques such as time series and instrumental variable analyses can mitigate these potential sources of bias, they remain inherent limitations of the study design. However, these limitations do not render this form of research less valuable than insights provided by RCTs, which have their own limitations.

Figure 18

Canvas Module as Presented to Students

Figure 19

Introduction on Canvas as Presented to Students

Home

Syllabus

Modules

Grades

People

Namecoach Roster

Zoom

Panopto Recordings

Adobe Creative Cloud

Introduction

In this module, you will find a link to tasks that are part of a dissertation study conducted by Julianne van Meerten in the Department of Human Development and Quantitative Methodology. The topic of these tasks is ***What Makes a Good Research Design?*** The tasks have been designed for students enrolled in courses that cover research methods or research principles, such as the one you are enrolled in.

Engaging in these tasks should help you become more familiar with the topic of research design. The procedures involve two (2) parts. Part 1 will take approximately 60 to 75 minutes, Part 2 will take 15 to 20 minutes. You will receive ***full credit*** as indicated by your course instructor if you complete tasks (i.e., get to the end of the survey and submit it) for ***both parts.*** Participating in this study will result in the number of points specified by your instructor. You must be 18 years or older to participate in this research.

Important reminders BEFORE you get started:

- Make sure the device you are using for this task has a keyboard (i.e., not a phone)
- Try to complete the tasks in a quiet, distraction-free environment

Contact information:
Julianne van Meerten | jmeerten@umd.edu
Dr. Donald J. Bolger | djbolger@umd.edu

If you do not wish to participate in the study, an alternative assignment that is no more challenging or time consuming than research participation will be provided by your instructor.

Figure 20

Part 1 of the Intervention on Canvas as Presented to Students

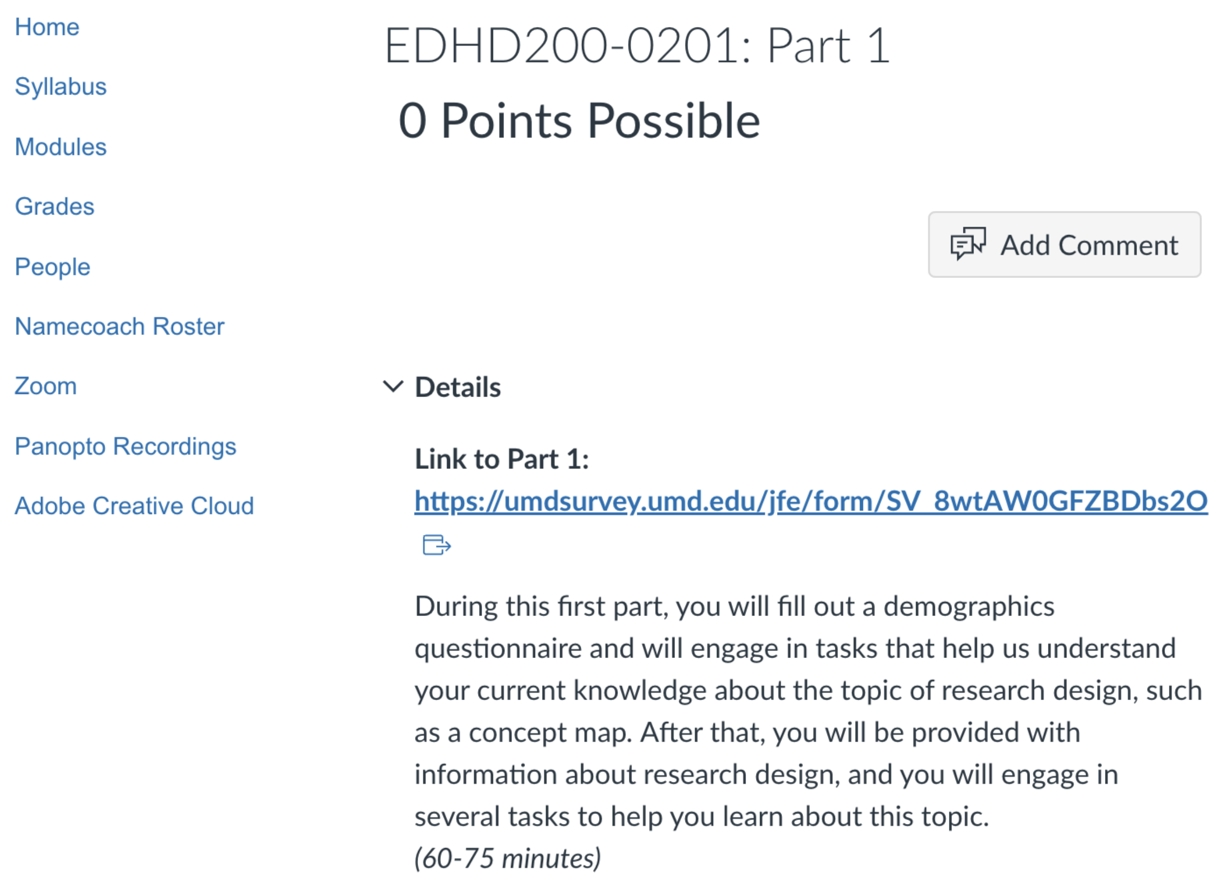


Figure 21

Part 2 of the Intervention on Canvas as Presented to Students

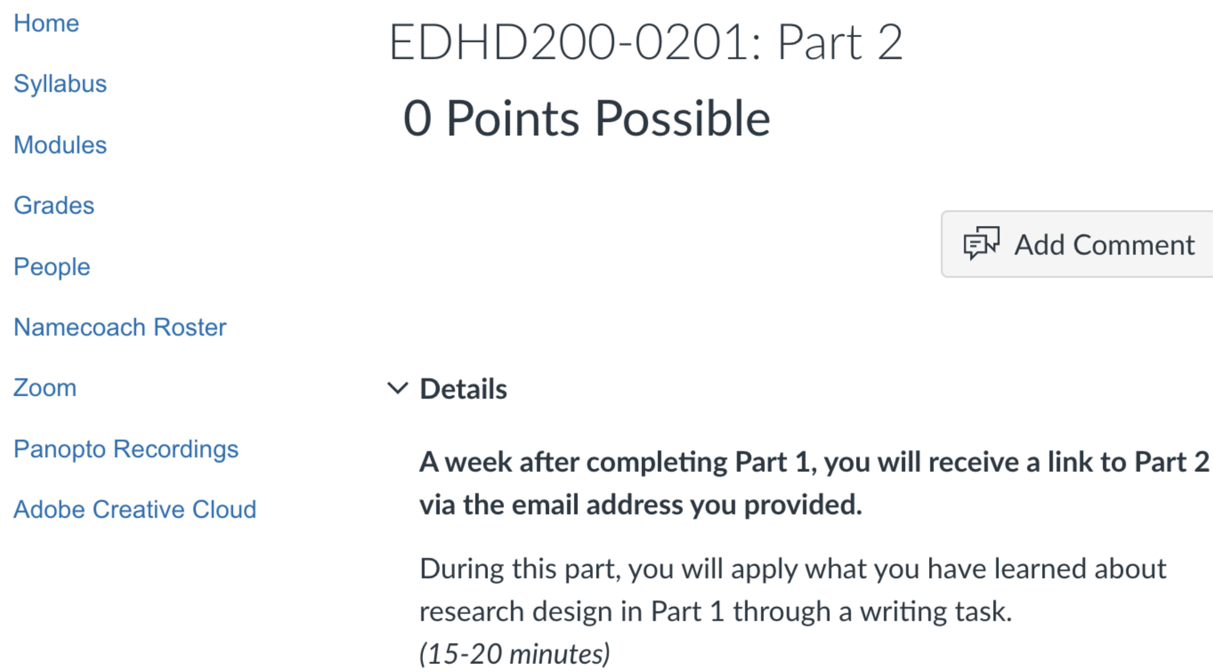


Figure 22

Concept Map Instructions as Presented to Students on Draw.io

Creating your Concept Map:

1. This concept map is a representation of your knowledge of ***"What Makes a Good Research Design?"*** It does not have to be pretty, as long as it accurately represents what you know about this topic!
2. Your concept map can be as big as you want it to be! No limits.
3. When your concept map is finished, **HIT THE SAVE & EXIT BUTTON (TOP RIGHT CORNER)**. This will send you back to the survey.
4. Double click on the square in the middle to write a concept in it
5. Hover over the square to make arrows appear
6. Click on an arrow in any direction
7. Select the shape you want to connect to
8. Write concepts in the shapes
9. Write on the arrows to indicate what they mean so that the concepts form short sentences with the arrows (see example below)

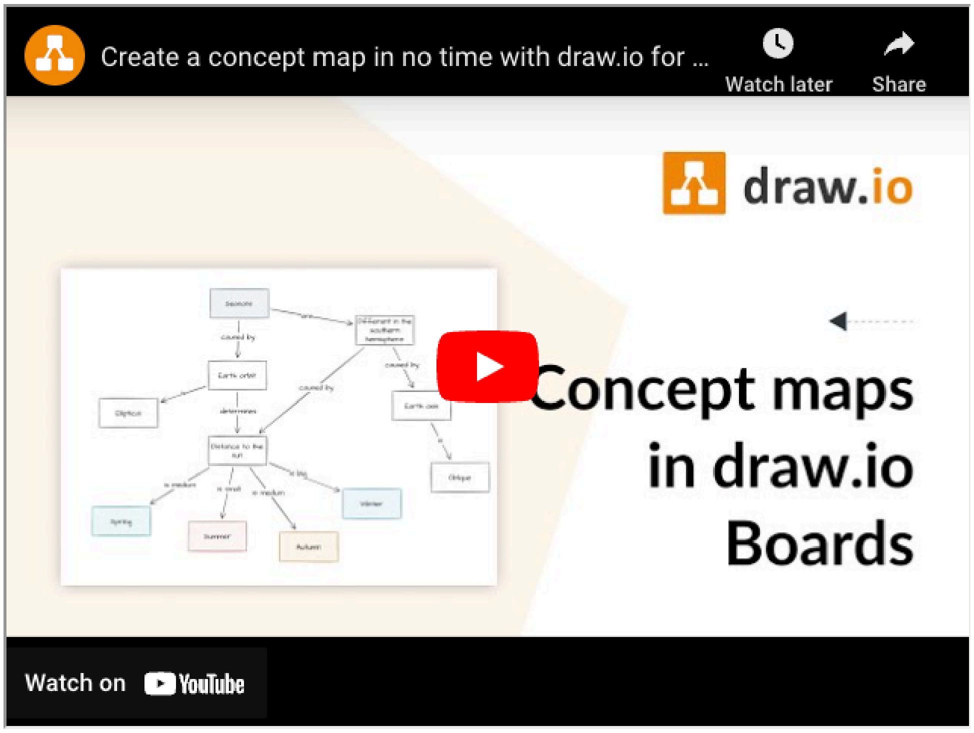
Example: This square —is now connected to→ This square

Tips:

- Adding shapes:
  - Click on an arrow OR
  - Select from the menu on the left
- Annotating shapes:
  - Double click on any shape to write on/in it, including squares and arrows
- You can drag shapes and arrows anywhere
- Reverse an arrow:
  - It is easiest to click on it, remove it by pressing the "backspace" or "delete" key, and insert a new arrow in the right direction
- Use the menu bar at the top left corner to edit your font size, etc.

Figure 23

Screenshot of Instructional Video About Concept Maps & Corresponding URL



URL: https://youtu.be/22nc9l-6Ikw?t=31

Appendix III: Additional Tables and Figures

Table 27

List of Relevant Concepts Used During Concept Map Scoring

	Relevant Concepts	Additional Related Concepts/Synonyms
Considerations in Research Design:		
	Trade-offs between internal and external validity	
	Importance of mitigating biases and limitations inherent in different study designs	
Randomization:	Assigning participants to groups by chance to minimize bias	participants/sample
Bias:	Systematic errors that can influence research results	lack of influence by external factors; objectivity; biased/unbiased
Generalizability:	How well research findings apply to a broader population	interpretable, (sample) population, population of interest, target population
Validity:		co(n)founds; reliability
Internal Validity:	Confidence in attributing observed effects to tested factors	
External Validity:	Generalizability of research findings beyond specific study parameters	representative sample; applicability
Correlational Research:		
	Descriptive and non-manipulated	
	Observes or measures variables as they naturally occur	observation(al)
	Challenges: determining cause and effect direction, influence of extraneous variables	
Experimental Research:		scientific method; nonexperimental
	Involves manipulation of variables	(control/dependent/independent/predictor/outcome/measurable) variables
	Controls experimental situations	

	Employs random assignment of participants	
	Facilitates inference of causation	
	Drawbacks: artificiality, ethical constraints, impracticality	feasibility; informed consent; (do no) harm; ethics/ethical guidelines; confidentiality; deception
Randomized Controlled Trials (RCTs):		
Strengths:	Excellent internal validity due to randomization	
Limitations:	Limited external validity, potential biases inherent to RCTs	
	Considered the "gold standard" due to strong internal validity, but may have limited generalizability	
Observational Research:		
Strengths:	Higher external validity, opportunity to evaluate real-world outcomes	
Limitations:	Lower internal validity, challenges in differentiating treatment effects from other factors	
	Studies populations without manipulating variables, can provide insights into real-world settings but may have limitations in establishing cause-and-effect	
	Difficulty in differentiating treatment effects from other changes	
	Challenges in identifying and measuring certain factors	
	Statistical limitations in mitigating bias	
Population-Based Observational Studies:		
	Include all patients within a given jurisdiction	
	Provide insights into underrepresented groups and rare diseases	
	Offer information on treatment toxicity and gaps in care	
Not most relevant, specific to medicine:		

Complementary Therapy Utilization:	Discrepancy between RCT evidence and actual clinical practice
	Identification of gaps in care through population-based studies
Impact on Clinical Practice and Policy:	Influence of observational research on oncology practice and policy decisions
	Usefulness in addressing questions not suitable for RCTs

Figure 24

Student 1 Transfer-Test Sample Seemingly Generated by OpenAI

"I am not familiar with research design. but I did some research and found this:
Studying the effects of masking on COVID-19 transmission requires a robust research design that can provide reliable evidence. Here are a few research designs that could work well:

1. **Randomized Controlled Trials (RCTs):** Considered the gold standard in research, RCTs involve randomly assigning participants to either a masked or unmasked group and then observing the incidence of COVID-19 transmission over time. This design helps establish causality between masking and transmission reduction.

2. **Quasi-experimental Designs:** In situations where RCTs are impractical, quasi-experimental designs, such as interrupted time series or controlled before-and-after studies, can be employed. These designs involve comparing COVID-19 transmission rates before and after the implementation of masking policies in different communities or regions.

3. **Observational Studies:** These studies involve observing naturally occurring instances of masking (or lack thereof) and their association with COVID-19 transmission rates. Cohort studies, case-control studies, and cross-sectional studies are examples of observational designs that can be used to investigate this relationship.

4. **Simulation Studies:** Using mathematical models and computer simulations, researchers can estimate the potential impact of masking on COVID-19 transmission under various scenarios. These studies can help predict outcomes and inform public health policies.

5. **Meta-analyses and Systematic Reviews:** These approaches involve synthesizing data from multiple studies to provide a comprehensive overview of the relationship between masking and COVID-19 transmission. They help consolidate evidence from various sources and can offer valuable insights into the effectiveness of masking.

Ultimately, the choice of research design will depend on various factors, including the research question, available resources, ethical considerations, and practical constraints. Combining multiple approaches or using a mixed-methods approach may provide a more comprehensive understanding of the effects of masking on COVID-19 transmission."

Figure 25

Student 2 Transfer-Test Sample Seemingly Generated by OpenAI

"To thoroughly investigate the effects of masking on COVID-19 transmission, a combination of research designs would likely be most effective. Here are some:

Experimental Design:
An experimental design involves manipulating variables to observe their effect on an outcome. In this case, researchers could conduct controlled experiments where they randomly assign participants to either wear masks or not wear masks in various settings (e.g., indoors, outdoors, crowded places). By comparing infection rates between the two groups, researchers can directly assess the impact of masking on COVID-19 transmission. How ever this would not be very ethical to conduct.

Quasi-Experimental Design:
Quasi-experimental designs are similar to experimental designs but lack random assignment. In the context of studying masking effects, researchers may not have the ability to randomly assign individuals to mask-wearing conditions due to ethical or practical constraints. Instead, they could leverage naturally occurring situations (e.g., mask mandates in certain regions) to examine the association between mask usage and transmission rates. this would be the best design to answer the question.

Observational Design:
Observational studies involve observing and analyzing ongoing behavior without intervening or manipulating variables. Researchers could employ observational designs to assess real-world mask usage patterns and their relationship to COVID-19 transmission. This could involve surveys, interviews, or direct observations of individuals in various settings to understand compliance with mask mandates and its impact on transmission. this is dangerous for the researcher so it is not advisable.

Correlational Design:
Correlational studies examine the relationship between variables without manipulating them. Researchers could use correlational designs to analyze existing data sets, such as regional or national COVID-19 infection rates and corresponding mask-wearing policies or behaviors. By identifying correlations between these variables, researchers can gain insights into the potential effectiveness of masking in reducing transmission. This is the second best design because it can give researchers an idea on the relation ship between their variables

Descriptive Design:
Descriptive studies aim to describe characteristics or behaviors of a population. In the context of masking and COVID-19 transmission, descriptive research could involve documenting trends in mask usage over time, examining cultural or socioeconomic factors influencing mask-wearing behavior, and detailing the prevalence of COVID-19 cases in different mask-wearing contexts. How it won't allow you to test any hypothesis or theory so it is poorly suited for this question.

Each design offers unique insights, contributing to a more robust and nuanced understanding of this critical public health issue. That being said a Quasi Experimental design is the best."

Table 28

Correlations Between Knowledge Measures Original Dataset

Time Point	Pretest	Posttest	Transfer Test
Pretest	1.00		
Posttest	0.46	1.00	
Transfer Test	0.26	0.39	1.00

Table 29

Correlations Between Knowledge Measures Imputed Dataset

Time Point	Pretest	Posttest	Transfer Test
Pretest	1.00		
Posttest	0.47	1.00	
Transfer Test	0.27	0.24	1.00

Figure 26

Density Plots: Missing Data Analysis

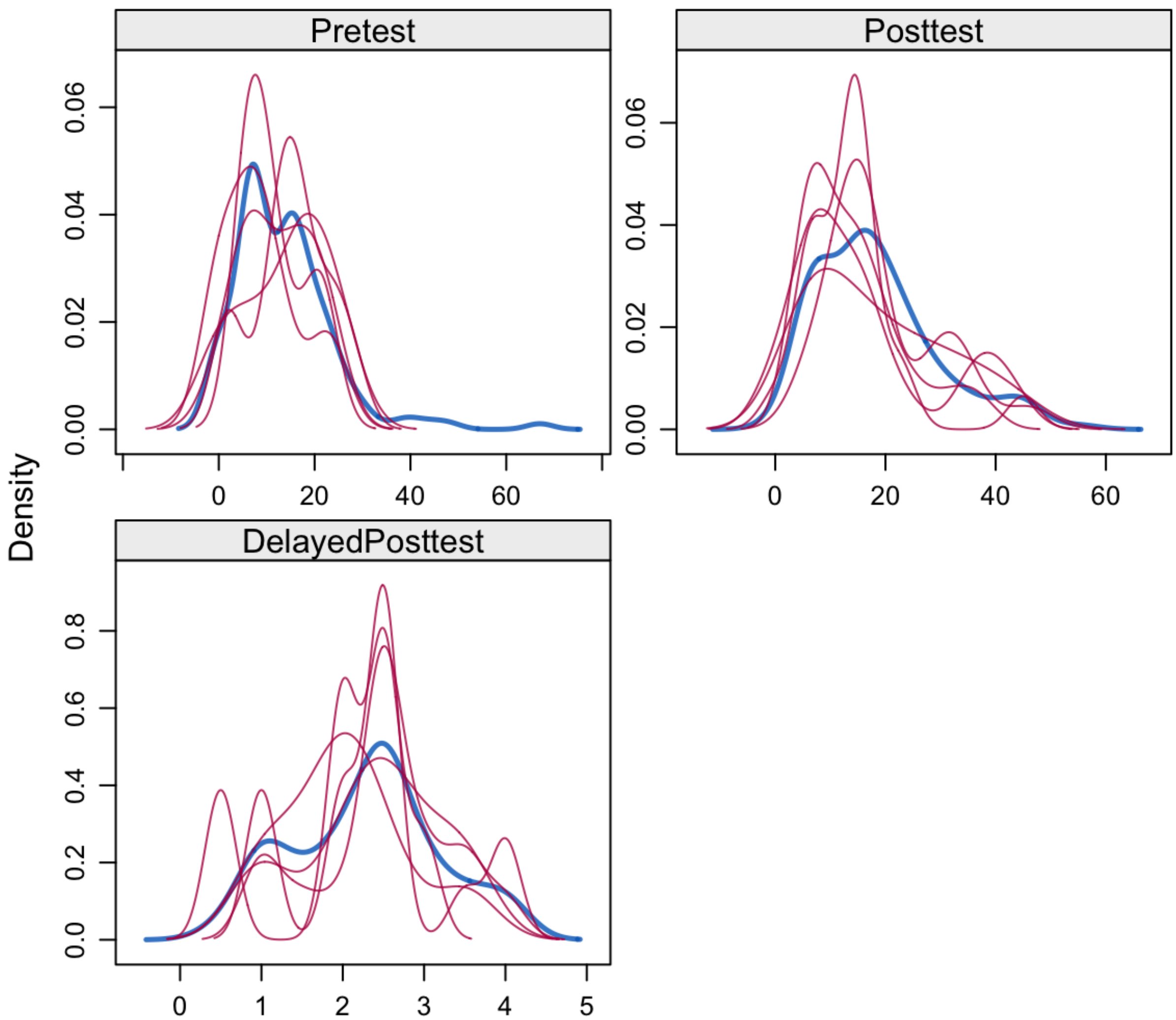


Figure 27

Scree Plots: PCA of Subjective Knowledge and Confidence Items at Time 1–5

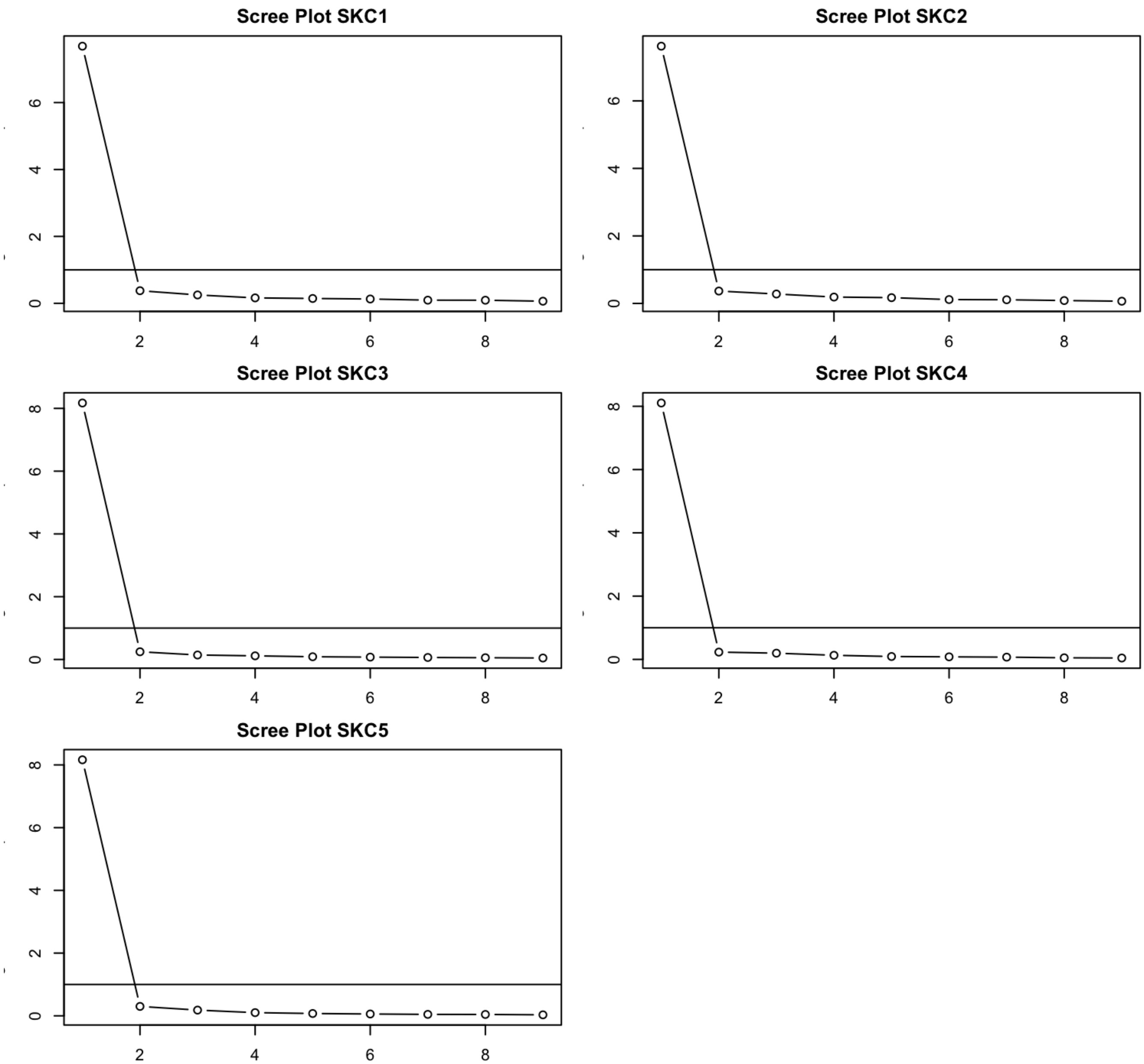
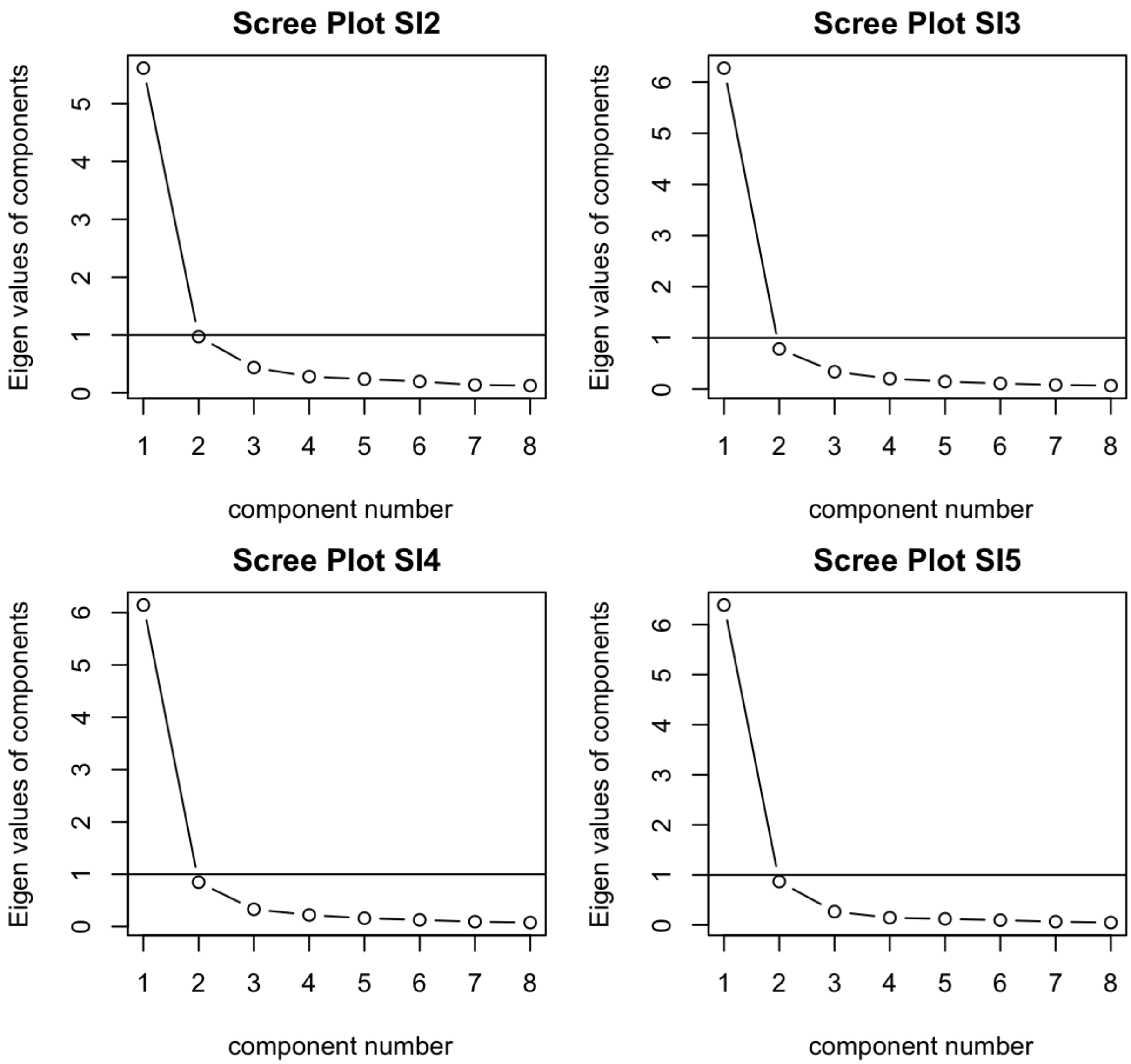


Figure 28

Scree Plots: PCA of Situational Interest Items at Time 2–5



REFERENCES

Ainley, M., & Patrick, L. (2006). Measuring self-regulated learning processes through tracking patterns of student interaction with achievement activities. *Educational Psychology Review*, *18*(3), 267–286. https://doi.org/10.1007/s10648-006-9018-z

Alexander, P. A. (1997). Mapping the multidimensional nature of domain learning: The interplay of cognitive, motivational, and strategic forces. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 213–250). JAI Press.

Alexander, P. A. (2018). Information management versus knowledge building: Implications for learning and assessment in higher education. In O. Zlatkin- Troitschanskaia, M. Toepper, H. Pant, C. Lautenbach, & C. Kuhn (Eds.), *Assessment of learning outcomes in higher education: Cross-national comparisons and perspectives*, 43-56. Springer. https://doi.org/10.1007/978-3- 319-74338-7_3

Alexander, P. A., & Murphy, P. K. (1998). Profiling the differences in students' knowledge, interest, and strategic processing. *Journal of Educational Psychology*, *90*(3), 435–447. https://doi.org/10.1037/0022-0663.90.3.435

Alexander, P. A., Fusenig, J., Schoute, E. C., Singh, A., Sun, Y., & van Meerten, J. E. (2023). Confronting the challenges of undergraduates' argumentation writing in a "Learning How to Learn" course. *Written Communication*, *40*(2), 482–517. https://doi.org/10.1177/07410883221148468

Apple Inc. (2024). *Numbers (Version 14.0)*. [Computer software]. https://www.apple.com/numbers/

Arnold, K. M., Umanath, S.,Thio, K., Reilly, W. B., McDaniel, M. A., & Marsh, E. J. (2017). Understanding the cognitive processes involved in writing-to-learn. *Journal of Experimental Psychology Applied, 23*(2), 115–127. https://doi.org/10.1037/xap0000119

Ausubel, D. P. (1968). Facilitating meaningful verbal learning in the classroom. *The Arithmetic Teacher*, *15*(2), 126–132. https://doi.org/10.5951/AT.15.2.0126

Baaijen, V. M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*, *36*(3), 199–223. https://doi.org/10.1080/07370008.2018.1456431

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*(1), 29–58. https://doi.org/10.3102/00346543 074001029

Bereiter, C., & Scardamalia, M. (1986). Levels of inquiry into the nature of expertise in writing. *Review of Research in Education*, *13*(1), 259–282. https://doi.org/10.3102/0091732X013001259

Besterfield‐Sacre, M., Gerchak, J., Lyons, M. R., Shuman, L. J., & Wolfe, H. (2004). Scoring concept maps: An integrated rubric for assessing engineering education. *Journal of Engineering Education, 93*(2), 105–115. https://doi.org/10.1002/j.2168-9830.2004.tb00795.x

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. Academic Press. https://doi.org/10.1016/C2013-0-10375-3

Biggs, J., & Collis, K. (1989). Towards a model of school-based curriculum development and assessment using the SOLO taxonomy. *Australian Journal of Education*, *33*(2), 151–163. https://doi.org/10.1177/168781408903300205

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction*, *20*(5), 372–382. https://doi.org/10.1016/j.learninstruc.2009.03.002

Booth, C. M., & Tannock, I. F. (2014). Randomised controlled trials and population-based observational research: Partners in the evolution of medical evidence. *British Journal of Cancer*, *110*(3), 551–555. https://doi.org/10.1038/bjc.2013.725

Boscolo, P., Arfé, B., & Quarisa, M. (2007). Improving the quality of students' academic writing: An intervention study. *Studies in Higher Education*, *32*(4), 419–438. https://doi.org/10.1080/03075070701476092

Brown, A. L. (1978). Knowing when, where, and how to remember: A problem of metacognition. In R. Glaser (Ed.), *Advances in Instructional Psychology* (Vol 1, pp. 77–165). Lawrence Erlbaum Associates.

Brown, K. W., Cozby, P. C., Kee, D. W., & Worden, P. E. (1999). *Research methods in human development* (2nd ed.). Mayfield Publishing Co.

Buchin, Z. L., & Mulligan, N. W. (2022). Retrieval-based learning and prior knowledge. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000773

Burnet, C. M. L. (2006). *Artificial Neural Network.* [Vector Image]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Artificial_neural_network.svg

Butler, J. A., & Britt, M. A. (2011). Investigating instruction for improving revision of argumentative essays. *Written Communication*, *28*(1), 70–96. https://doi.org/10.1177/0741088310387891

Chi, M. T. H. (1985). Changing conception of sources of memory development. *Human Development*, *28*(1), 50–56. https://doi.org/10.1159/000272935

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self‐explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1

Chi, M. T. H., De Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self‐explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1207/s15516709cog1803_3

De Bruin, A. B. H., Rikers, R. M. J. P., & Schmidt, H. G. (2007). The effect of self-explanation and prediction on the development of principled understanding of chess in novices. *Contemporary Educational Psychology*, *32*(2), 188–205. https://doi.org/10.1016/j.cedpsych.2006.01.001

Dewey, J. (1913). *Interest and effort in education*. Houghton Mifflin Company. https://doi.org/10.1037/14633-000

Dinsmore, D. L., & Alexander, P. A. (2016). A multidimensional investigation of deep-level and surface-level processing. *The Journal of Experimental Education*, *84*(2), 213–244. https://doi.org/10.1080/00220973.2014.979126

Dinsmore, D. L., & Hattan, C. (2020). Levels of strategies and strategic processing. In *Handbook of strategies and strategic processing* (pp. 29–46). Routledge.

Dowst, K. (1980). The epistemic approach: Writing, knowing, and learning. In T. R. Donovan & B. W. McClelland (Eds.), *Eight approaches to teaching composition* (pp. 65–86). National Council of Teachers of English.

Dunlosky, J. (2013). Strengthening the student toolbox: Study strategies to boost learning. *American Educator, 37*, p. 12–21.

Elbow, P. (1973). The pedagogy of the bamboozled. *Soundings: An Interdisciplinary Journal, 56*(2), 247–258. Penn State University Press.

Emig, J. (1971). *The composing processes of twelfth graders.* (NCTE Research Report No. 13) National Council of Teachers of English.

Faigley, L. (1986). Competing theories of process: A critique and a proposal. *College English*, *48*(6), 527–542. https://doi.org/10.2307/376707

Fallahi, C. R., Wood, R. M., Austad, C. S., & Fallahi, H. (2006). A program for improving undergraduate psychology students' basic writing skills. *Teaching of Psychology*, *33*(3), 171–175. https://doi.org/10.1207/s15328023top3303_3

Ferretti, R. P., & Lewis, W. E. (2018). Argumentative writing. In S. Graham, C. A. MacArthur, M. A. Hebert (Eds.), *Best practices in writing instruction* (pp. 135–161). The Guilford Press.

Flavell, J. H. (1971). First discussant's comments: What is memory development the development of? *Human Development, 14*(4)*,* 272–278. https://doi.org/10.1159/000271221

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), pp. 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Flower, L. (1989). Cognition, context, and theory building. *College Composition and Communication*, *40*(3), 282–311. https://doi.org/10.2307/357775

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387. https://doi.org/10.2307/356600

Fonseca, B. A., & Chi, M. T. H. (2011). Instruction based on self-explanation. In R. E. Mayer & P. A. Alexander (Eds.), *The handbook of research on learning and instruction* (pp. 296-321). Routledge. https://doi.org/10.4324/9780203839089-24

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false‐positive findings–a practical guide. *Biological Reviews, 92*(4), 1941-1968. https://doi.org/10.1111/brv.12315

Friend, R. (2001). Effects of strategy instruction on summary writing of college students. *Contemporary Educational Psychology*, *26*(1), 3–24. https://doi.org/10.1006/ceps.1999.1022

Galbraith, D. (1992). Conditions for discovery through writing. *Instructional Science*, *21*(1–3), 45–71. https://doi.org/10.1007/BF00119655

Galbraith, D. (1999). Writing as a knowledge-constituting process. In G. Rijlaarsdam, E. Espéret, M. Torrance, & D. Galbraith (Eds.), *Knowing what to write: Conceptual Processes in Text Production* (Vol. 4, pp. 139–160). Amsterdam University Press.

Galbraith, D. (2009). Writing as discovery. In V. Connelly, A. L. Barnett, J. E. Dockrell, & A. Tolmie (Eds.), *BJEP monograph series II: Part 6 teaching and learning writing*. British Psychological Society. https://doi.org/10.1348/978185409X421129

Galbraith, D., & Baaijen, V. M. (2018). The work of writing: Raiding the inarticulate. *Educational Psychologist*, *53*(4), 238–257. https://doi.org/10.1080/00461520.2018.1505515

Galbraith, D., Ford, S., Walker, G, & Ford, J. (2005). The contribution of different components of working memory to knowledge transformation during writing. *L1–Educational Studies in Language and Literature*, *5*(2), 113–145. https://doi.org/10.1007/s10674-005-0953-2

Galbraith, D., Peters, A., Hall, S., & Baaijen, V. (2023). Measuring discovery through writing. In R. Horowitz, *The Routledge international handbook of research on writing* (2nd ed., pp. 573–584). Routledge. https://doi.org/10.4324/9780429437991-49

Gibbons, M. (2019). Attaining landmark status: Rumelhart and McClelland's *PDP Volumes* and the connectionist paradigm. *Journal of the History of the Behavioral Sciences*, *55*(1), 54–70. https://doi.org/10.1002/jhbs.21946

Gourgey, A. F.. (2002). Metacognition in basic skills instruction. In H. J. Hartman (Ed.), *Metacognition in Learning and Instruction* (Vol. 19, pp. 17–32). Springer. https://doi.org/10.1007/978-94-017-2243-8

Graham, S., Kiuhara, S. A., & MacKay, M. (2020). The effects of writing on learning in science, social Studies, and mathematics: A meta-analysis. *Review of Educational Research*, *90*(2), 179–226. https://doi.org/10.3102/0034654320914744

Greenhouse, S. W., & Geisser, S. (1959a). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112. https://doi.org/10.1007/BF02289823

Greenleaf, C., Schoenbach, R., Friedrich, L., Murphy, L., & Hogan, N. (2023). *Reading for understanding: How reading apprenticeship improves disciplinary learning in secondary and college classrooms* (3rd ed.). John Wiley & Sons.

Hacker, D. J., & Bol, L. (2019). Calibration and self-regulated learning: Making the connections. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 647–677). Cambridge University Press. https://doi.org/10.1017/9781108235631.026

Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction*, *24*, 37–47. https://doi.org/10.1016/j.learninstruc.2012.10.001

Hall, S. M. (2023). *Estimating the effects of writing beliefs, writing processes and drafting strategies on the development of subjective understanding.* (Publication No. 476267) [Doctoral dissertation, University of Southampton]. University of Southampton Institutional Repository. https://eprints.soton.ac.uk/id/eprint/476267

Hartman, H. J. (Ed.). (2001). *Metacognition in Learning and Instruction* (Vol. 19). Springer Netherlands. https://doi.org/10.1007/978-94-017-2243-8

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hayes, J. R., & Flower, L. S. (1979, April 8–12). *Writing as problem solving*. Paper presented at the Annual Meeting of the American Educational Research Association in San Francisco, California.

Hayes, J. R., & Flower, L. S. (1986). Writing research and the writer. *American Psychologist, 41*(10), 1106–1113. https://doi.org/10.1037/0003-066X.41.10.1106

Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research, 60*(4), 549–571. https://doi.org/10.3102/00346543060004549

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

Hung, C.-H., & Lin, C.-Y. (2015). Using concept mapping to evaluate knowledge structure in problem-based learning. *BMC Medical Education*, *15*(212). https://doi.org/10.1186/s12909-015-0496-x

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational and Behavioral Statistics, 1*(1), 69–82. https://doi.org/10.3102/10769986001001069

Jitendra, A. K., DuPuis, D. N., Rodriguez, M. C., Zaslofsky, A. F., Slater, S., Cozine-Corroy, K., & Church, C. (2013). A randomized controlled trial of the impact of schema-based instruction on mathematical outcomes for third-grade students with mathematical difficulties. *The Elementary School Journal, 114*(2), 252–276. https://doi.org/10.1086/673199

Johnstone, K. M., Ashbaugh, H., & Warfield, T. D. (2002). Effects of repeated practice and contextual-writing experiences on college students' writing skills. *Journal of Educational Psychology, 94*(2), 305–315.

Kingir, S., Geban, O., & Günel, M. (2012). How does the Science Writing Heuristic approach affect students' performances of different academic achievement levels? A case for high school chemistry. *Chemistry Education Research and Practice, 13*(4), 428–436. https://doi.org/10.1039/C2RP20013A

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review, 11*(3), 203–270. https://doi.org/10.1023/A:1021913217147

Klein, P. D. (2004). Constructing scientific explanations through writing. *Instructional Science*, *32*(3), 191–231. https://doi.org/10.1023/B:TRUC.0000024189.74263.bd

Klein, P. D., & Boscolo, P. (2016). Trends in research on writing as a learning activity. *Journal of Writing Research*, *7*(3), 311–350. https://doi.org/10.17239/jowr-2016.07.03.01

Klein, P. D., & Rose, M. A. (2010). Teaching arguments and explanation to prepare junior students for writing to learn. *Reading Research Quarterly, 45*(4), 433–461. https://doi.org/10.1598/RRQ.45.4.4

Kwok, S. (2013). *The effects of freewriting on the quality and characteristics of short essays* (Publication No. 3564754) [Doctoral Dissertation, Temple University]. ProQuest LLC. https://doi.org/10.34944/dspace/1659

Lapp, D., Flood, J., & Farnan, N. (Eds.). (2016). *Content area reading and learning: Instructional strategies*. Routledge.

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement*, *70*(4), 647–671. https://doi.org/10.1177/0013164409355699

List, A. (2018). Strategies for comprehending and integrating texts and videos. *Learning and Instruction*, *57*, 34–46. https://doi.org/10.1016/j.learninstruc.2018.01.008

Mateos, M., Martín, E., Cuevas, I., Villalón, R., Martínez, I., & González-Lamas, J. (2018). Improving written argumentative synthesis by teaching the integration of conflicting information from multiple sources. *Cognition and Instruction, 36*(2), 119–138. https://doi.org/10.1080/07370008.2018.1425300

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*(2), 204–209. https://doi.org/10.1214/aoms/1177731915

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In J. L. McClelland, D.E. Rumelhart, and the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol 1, pp. 3-44). The MIT Press.

McNamara, D. S. (Ed.). (2007). *Reading comprehension strategies: Theories, interventions and technologies* (1st ed.). Taylor & Francis Group. https://doi.org/10.4324/9780203810033

Melby-Lervåg, M., & Lervåg, A. (2014). Effects of educational interventions targeting reading comprehension and underlying components. *Child Development Perspectives, 8*(2), 96–100. https://doi.org/10.1111/cdep.12068

Morphy, P. (2013). *Frederick Douglass and I: Writing to read and relate history with life among African American adolescents at a high-poverty urban school* (Publication No. 3575567) [Doctoral dissertation, Vanderbilt University]. ProQuest Dissertations and Theses.

Nückles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., & Renkl, A. (2020). The self-regulation-view in writing-to-learn: Using journal writing to optimize cognitive load in self-regulated learning. *Educational Psychology Review, 32*(4), 1089–1126. https://doi.org/10.1007/s10648-020-09541-1

Novak, J. D., & Cañas, A. J. (2006). The theory underlying concept maps and how to construct and use them. Technical Report IHMC CmapTools 2006-01, Florida Institute for Human and Machine Cognition

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn.* Cambridge University Press. https://doi.org/10.1017/CBO9781139173469

Office of Institutional Research, Planning & Assessment. (2023). *UMD undergraduate application summary.* https://irpa.umd.edu/CampusCounts/Admissions/apps_ug.pdf

Palincsar, A. S., & Brown, A. L. (1986). Interactive teaching to promote independent learning from text. *The Reading Teacher, 39*(8), 771–777. https://www.jstor.org/stable/20199221

Persky, A. M., & Dinsmore, D. L. (2019). Metacognitive changes and sources of confidence judgements in health professions classroom learning. *Currents in Pharmacy Teaching and Learning*, *11*(4), 338–345. https://doi.org/10.1016/j.cptl.2019.01.005

Phillips, N. E., Mills, C., D'Mello, S., & Risko, E. F. (2016). On the influence of re-reading on mind wandering. *Quarterly Journal of Experimental Psychology, 69*(12), 2338–2357. https://doi.org/10.1080/17470218.2015.1107109

Pillow, B. H., Mash, C., Aloian, S., & Hill, V. (2002). Facilitating children's understanding of misinterpretation: Explanatory efforts and improvements in perspective taking. *The Journal of Genetic Psychology*, *163*(2), 133–148. https://doi.org/10.1080/00221320209598673

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667. https://doi.org/10.1037/0022-0663.95.4.667

Posit Team (2022). *RStudio: Integrated Development Environment for R* (Version 2023.12.1+402) . Posit Software, PBC. http://www.posit.co/.

Raedts, M., Rijlaarsdam, G., van Waes, L., & Daems, F. (2007). Observational learning through video-based models: Impact on students' accuracy of self-efficacy beliefs, task knowledge and writing performances. In G. Rijlaarsdam (Series Ed.) and P. Boscolo & S. Hidi (Volume Eds.), *Studies in Writing: Writing and Motivation* (Vol. 19, pp. 219–238). Elsevier. https://doi.org/10.1163/9781849508216

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). *Psychology of reading* (2nd Ed.). Psychology Press.

Renninger, K. A., & Hidi, S. (2002). Student interest and achievement: Developmental issues raised by a case study. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 173–195). Academic Press. https://doi.org/10.1016/B978-012750053-9/50009-7

Renninger, K. A., & Hidi, S. (2017). *The power of interest for motivation and engagement*. Routledge. https://doi.org/10.4324/9781315771045

Rijlaarsdam, G., & Van den Bergh, H. (2006). Writing process theory: A Functional Dynamic Approach. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Writing Research* (pp. 41–53) Guildford Press.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*(2), 346–362. https://doi.org/10.1037/0022-0663.93.2.346

Rost, M. (2011). *Teaching and researching: Listening* (2nd Ed.) (C. N. Candlin & D. R. Hall, Series Eds.). Routledge. https://doi.org/10.4324/9781315833705

Rotgans, J. I., & Schmidt, H. G. (2011). Situational interest and academic achievement in the active-learning classroom. *Learning and Instruction*, *21*(1), 58–67. https://doi.org/10.1016/j.learninstruc.2009.11.001

Rouzinov, S., & Berchtold, A. (2022). *Regression-based approach for testing the type of missing data (*Version 1.1) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=RBATMD

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In J. L. McClelland, D.E. Rumelhart and the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol 1, pp. 45–76). The MIT Press. https://doi.org/10.7551/mitpress/5236.003.0005

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D.E. Rumelhart and the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol 2, pp. 7–57). The MIT Press. https://doi.org/10.7551/mitpress/5237.003.0004

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics* (Vol 2, pp. 142–175). Cambridge University Press.

Scardamalia, M., & Bereiter, C. (1991). Higher levels of agency for children in knowledge building: A challenge for the design of new knowledge media. *Journal of the Learning Sciences, 1*, 37–68. https://doi.org/10.1207/s15327809jls0101_3

Scardamalia, M., Bereiter, C., & Steinbach, R. (1984). Teachability of reflective processes in written composition. *Cognitive Science*, *8*(2), 173–190. https://doi.org/10.1207/s15516709cog0802_4

Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist, 26*(3–4), 299–323. https://doi.org/10.1207/s15326985ep2603&4_5

Schraw, G. (2002). Promoting general metacognitive awareness. In H. J. Hartman (Ed.), *Metacognition in Learning and Instruction* (Vol. 19, pp. 3–16). Springer.

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction*, *24*, 48–57. https://doi.org/10.1016/j.learninstruc.2012.08.007

Segalàs, J., Mulder, K. F., & Ferrer‐Balas, D. (2012). What do EESD "experts" think sustainability is? Which pedagogy is suitable to learn it?: Results from interviews and Cmaps analysis gathered at EESD 2008. *International Journal of Sustainability in Higher Education*, *13*(3), 293–304. https://doi.org/10.1108/14676371211242599

Siegler, R., DeLoache, J., & Eisenberg, N. (2003). *How children develop*. Worth Publishers.

Sinatra, G. M. (2005). The" warming trend" in conceptual change research: The legacy of Paul R. Pintrich. *Educational psychologist, 40*(2), 107-115. https://doi.org/10.1207/s15326985ep4002_5

Singer, L. M., & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education*, *85*(1), 155–172. https://doi.org/10.1080/00220973.2016.1143794

Snyder, M. (1987). *Public appearances, private realities: The psychology of self-monitoring*. W. H. Freeman & Co.

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review 12*, pp. 437–475. https://doi.org/10.1023/A:1009084430926

Tracey, D. H., & Morrow, L. M. (2024). *Lenses on reading* (4th Ed.). The Guilford Press.

University of Maryland, Human Development and Quantitative Methodology Department [UMD HDQM] (2024a). *Paradigms and perspectives in human development* [Course syllabus]. University of Maryland.

University of Maryland, Department of Human Development and Quantitative Methodology [UMD HDQM] (2024b). *Research methods in human development* [Course syllabus]. University of Maryland.

University of Maryland, Psychology Department [UMD PSYC] (2023). *Research methods laboratory* [Course syllabus]. University of Maryland.

van Buuren S., Groothuis‐Oudshoon K. (2011) MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

van der Loo, J. M., Krahmer, E. J., & van Amelsvoort, M. A. A. (2018). Learning how to write an academic text: A comparison of observational learning with learning by doing. *Journal of Writing Research*, *9*(3), 233–258. https://doi.org/10.17239/jowr-2018.09.03.01

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merriënboer, J. J. G. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction*, *24*, 15–25. https://doi.org/10.1016/j.learninstruc.2012.08.005

van Meerten, J. E., & Alexander, P. A. (2023). Classroom-based academic writing interventions in higher education: A systematic review [Manuscript in publication]. Department of Human Development and Quantitative Methodology, University of Maryland, College Park

Watson, M. K., Pelkey, J., Noyes, C. R., & Rodgers, M. O. (2016). Assessing conceptual knowledge using three concept map scoring methods. *Journal of Engineering Education, 105*(1), 118–146. https://doi.org/10.1002/jee.20111

Winne, P. H., & Jamieson-Noel, D. (n.d.). *Exploring students' calibration of self reports about study tactics and achievement*.

Winne, P. H., & Muis, K. R. (2011). Statistical estimates of learners' judgments about knowledge in calibration of achievement. *Metacognition and Learning*, *6*(2), 179–193. https://doi.org/10.1007/s11409-011-9074-8

Wolfe, C. R., & Britt, M. A. (2008). The locus of the myside bias in written argumentation. Thinking & Reasoning, 14(1), 1–27. https://doi.org/10.1080/13546780701527674

Wong, R. M. F., Lawson, M. J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction*, *12*(2), 233–262. https://doi.org/10.1016/S0959-4752(01)00027-5

Zwaal, W. , & Otting, H. (2012). The Impact of Concept Mapping on the Process of Problem-based Learning. *Interdisciplinary Journal of Problem-Based Learning, 6*(1), 105–128. https://doi.org/10.7771/1541-5015.1314