Use Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Idaho State University, I agree that the Library shall make it freely available for inspection. I further state that permission to download and/or print my thesis for scholarly purposes may be granted by the Dean of the Graduate School, Dean of my academic division, or by the University Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature _____

Date _____

Influence of Summary Modality on Metacomprehension Accuracy

by

Erin Madison

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in the Department of Psychology

Idaho State University

Summer 2019

Committee Approval

To the Graduate Faculty:

The members of the committee appointed to examine the thesis of ERIN MADISON find it satisfactory and recommend that it be accepted.

_____

Dr. Erika K. Fulton,

Major Advisor

_____

Dr. Kandi Turley-Ames,

Committee Member

_____

Dr. Elizabeth Brunner,

Graduate Faculty Representative

TABLE OF CONTENTS

List of Tables v

List of Figures vi

Abstract vii

Chapter 1: Influence of Summary Modality on Metacomprehension Accuracy 1

Importance of Metacomprehension 4

Immediate vs. Delayed Judgments 5

Metacognitive Cues 6

Are Speaking and Writing Summaries Equivalent? 9

Perceived Cognitive Load Differences Between Summary Types 12

The Present Study 14

Hypotheses 14

Chapter 2: Methods 17

Design 17

Participants 17

Materials 18

Procedure 18

Statistical Analyses 21

Chapter 3: Results 23

Demographics 23

Judgment Magnitude and Multiple-Choice Performance 23

Metacognitive Prediction Accuracy 24

Summary Characteristics 25

Cognitive Load 27

Chapter 4: Discussion 28

Absolute Accuracy 28

Relative Accuracy 30

Accessibility Hypothesis vs Situation Model 32

Perceived Cognitive Load 35

Limitations 37

Implications and Future Directions 38

References 41

Appendices 63

List of Tables

Table 1 Summary of Average Response and Performance Scores……..…….…….. 52

Table 2 Summary Characteristics Comparisons……..…….……..53

Table 3 Sequential Multiple Regression Analysis of Summary Characteristics and Prediction Magnitude ……..... .....……54

Table 4 Sequential Multiple Regression Analysis of Summary Characteristics and Relative Accuracy……...…….…….55

Table 5 Multiple Regression Analysis With Cognitive Load and Prediction Magnitude...…..…..56

Table 6 Multiple Regression Analysis With Summary Characteristics and Multiple-Choice Accuracy ……..…..……... …….……...……..……...….57

Table 7 Correlation Matrix for Summary Characteristics…….…….….58

List of Figures

Figure 1 Model of cognitive monitoring and control proposed by Nelson & Narens (1990).……....59

Figure 2 Mean gamma correlation as a function of condition ……..…….60

Figure 3 Mean intrinsic and extraneous cognitive load as a function of condition …….……...61

Abstract

Metacomprehension, the ability to monitor and regulate reading, can facilitate efficient studying. Individuals have low relative accuracy, meaning they cannot adequately differentiate well known from less well-known information. Delayed written summarization increases relative accuracy, so this study compared written summaries to oral summaries to test summary modality's impact on metacomprehension accuracy. Summary modality can lead to differences in summary characteristics, and thus metacomprehension cues, which can influence relative accuracy. Participants in an oral, written, or no summary condition read and summarized passages, judged their comprehension, and took a multiple-choice test on the passages. Only the written condition exhibited relative accuracy significantly greater than zero. Out of the summary characteristics measured, word count and summary quality related to prediction magnitude, whereas word count and total time influenced relative accuracy. The results have implications for the accessibility and situation model hypotheses, and practical applications for study habits.

Keywords: metacomprehension, summary modality, cue utilization, accessibility hypothesis, situation model hypothesis

Chapter 1: Influence of Summary Modality on Metacomprehension Accuracy

Metacognition occurs in almost every facet of cognitive experience (Flavell, 1979). Metacognition is a term coined by Flavell that describes the awareness and thoughts of one's personal cognitions (Flavell, 1979), and includes monitoring and controlling cognitions (Lin, Moore, & Zabrucky, 2000; Maki, Shields, Wheeler, & Zacchilli, 2005). Two commonly studied aspects of this field include metamemory and metacomprehension, which are thoughts and behaviors regarding memory and reading, respectively. Research in metacognition can be readily applied to school and work environments (Hacker, Horgan, & Rakow, 2000), so it is deeply important to understand.

Nelson and Narens (1990) proposed a model that is commonly used to describe metacognitive activity (see Figure 1). According to this model, there are two levels of any cognitive task: an object level and a meta-level. Cognitive tasks occur in the object level. By monitoring a current task, information from the object level is shuttled to the meta-level. Judgments about the information occur at the meta-level first, then can regulate behaviors or cognitions occurring at the object-level. Metacognitive monitoring is the process of assessing current cognitions. For example, a student might realize their mind has wandered while reading. Metacognitive control allows for active adjustment of thoughts or behaviors depending on the input from monitoring, ensuring that the individual performs their current cognitive task accurately and efficiently. A student who rereads parts of a textbook because they did not understand the content displays metacognitive control.

A person can use their metacognitive ability to assess and regulate their actions and beliefs so that they can study and learn efficiently, a process formally termed self-regulated learning (Dunlosky, Hertzog, Kennedy, & Thiede, 2005; Schunk & Zimmerman, 1998). Self-

regulated learning is the ability to set and reach learning goals during study (Efklides, 2011). Students create self-regulation rules as they learn and improve their reading abilities, such as spending more time on sentences that are unclear (Winne, 1996). The effectiveness of self-regulated learning depends on the individual's planning and strategy use. Metacognitive judgments are important in this process because they affect behaviors during the task. During self-regulated study, metacognitive judgments can determine whether to begin, continue, or end a task, as well as determine which information to study further (Metcalfe & Finn, 2008; Nelson & Narens, 1990). For example, a person judges whether they will continue or end their study session depending on their beliefs about how well they learned the information. If their metacognitive judgments do not match reality, the inaccurate judgment might lead to poor cognitive control, such as the termination of the study session before the information is sufficiently understood (Ariel, 2013; Kornell & Metcalfe, 2006; Wiley, Griffin, & Thiede, 2005).

To assess judgment accuracy, researchers have developed a paradigm that requires participants to study information, make judgments about their learning, and then demonstrate evidence of learning (Epstein, Glenburg, & Bradly, 1984; Glenberg & Epstein, 1985; Glenberg, Wilkinson, & Epstein; 1982; Maki & Berry, 1984; Weaver, 1990). Judgments can then be compared to performance to determine metacognitive accuracy. There are two main types of metacognitive accuracy: relative accuracy and absolute accuracy. From a monitoring perspective, relative accuracy measures the ability of a person to distinguish between that which is well learned and less well learned. For example, an individual is considered highly accurate when they can predict on which tests they will achieve high scores and on which they will achieve low scores. Typically, gamma correlations are used to calculate relative accuracy in metacognitive research (Nelson, 1984), but Stuart's Tau-c has been argued to be a more correct correlation

calculation (Fulton, 2015). When calculating gamma correlations, “ties” in the data are not accounted for, while Stuart’s tau-c accounts for ties. Thus, Stuart’s tau-c is statistically more precise than gamma correlations but also more conservative. Both methods were used in the current research to assure that the results are precise but also comparable to prior research. Absolute accuracy, or calibration, is the degree of correspondence between objective and subjective performance (Dunlosky & Lipko, 2007; Kwon & Linderholm, 2014), and can measure over- and under-confidence. Bias is a type of calibration that is calculated by finding the magnitude and direction of the difference between the average judgment and comprehension scores. For example, an individual might predict that they will score 90% on a test, but they actually score 80%. This prediction shows overconfidence by a fairly high magnitude. In comparison, their classmate might predict they will score an 85% when they actually earn a 90%. This student shows under-confidence but with a smaller magnitude difference between predicted and actual scores. It should be noted that accuracy and magnitude are two different concepts. Magnitude simply refers to how high or low judgments are, while accuracy indicates the relationship between the judgment and the performance on a test.

Accuracy can be measured at different times in the metacognitive process. A prediction judgment measures how an individual believes they will score on a future test (Maki & Serra, 1992). Predictions are generally inaccurate, in part, because the specifics of the test information are unknown, and thus individuals must base judgments on information such as their subjective feelings about the material, or their past-experience (Koriat, 1997). These are called cues and will be discussed in more detail later. Judgments are based on cues; therefore, prediction judgments can be used to uncover which cues are present and utilized during summarization. For example, if a high word count is correlated with a high prediction judgment, then judgment may

be based on word count. Post-diction judgments, made after each test question or after the full test, are called confidence judgments (Maki & Serra, 1992). Post-dictions are more accurate than predictions on average, as during the test they can judge the plausibility of their answer and the quality of distractors (i.e., incorrect answer choices) on the test (Maki, 1998b; Pierce & Smith, 2001). Because post-diction judgments can be based on the test and subjective feelings about the test, these tell less about the cues present in the summarization process and will not be a focus of this study.

Importance of Metacomprehension

Metacomprehension, a type of metacognition, is broadly defined as a person's thoughts, assessment, and regulation of their reading (Dunlosky & Lipko, 2007; Dunlosky et al., 2005; Maki & Berry, 1984) and was the focus of the current study. As a person reads, they can monitor the input, assuring it is understood (Nelson & Narens, 1990). Importantly, if the individual realizes they lost focus or misunderstood the information, they might implement techniques to improve their understanding, such as rereading (Dunlosky & Lipko, 2007; Rawson, Dunlosky, & Thiede, 2000). In the context of an educational setting, if a person misjudges a text that they read, they might be overconfident in their comprehension of the text and fail to regulate their learning properly (Dunlosky et al., 2005; Wiley et al., 2016). In general, students with higher metacomprehension monitoring accuracy have a higher likelihood of implementing techniques to better understand information (Rawson, Dunlosky, & Thiede, 2000). However, students tend to over-estimate their reading comprehension (Dunlosky & Lipko, 2007) and underutilize reading strategies (Bjork, Dunlosky, & Kornell, 2013). Students also tend to have poor relative accuracy. Across studies, the relative accuracy for reading comprehension predictions tends to have a gamma correlation of about .27 (Dunlosky & Lipko, 2007; Lin & Zubrucky, 1998; Maki, 1998b;

Reid, Morrison, & Bol, 2017). Given that metacomprehension accuracy is poor, improving metacomprehension accuracy is crucial for efficient studying and increasing academic performance, and fortunately there is evidence that these skills can be improved (Anderson & Thiede, 2008; Reid, Morrison, & Bol, 2017; Wiley et al., 2016). Metacomprehension training successfully leads to higher quiz or exam grades in the context of an undergraduate classroom (Nietfiled, Coa, & Osborne, 2006; Wiley et al., 2016). For example, students were instructed to ask themselves questions during reading, such as "What new information does this paragraph add?" They later showed an increase in metacognitive control as demonstrated by participants studying the texts in an order that was tailored to their own studying needs. In contrast, the controls studied the texts in the order that the texts were given, even though they had the choice to study the texts in their preferred order (Wiley et al., 2016). The metacognitive group received higher grades on the quizzes, suggesting that test performance can be improved by increasing metacognitive monitoring (Wiley et al., 2016). These studies indicate that metacomprehension can be trained relatively easily and that improving metacomprehension will, in turn, improve test grades.

Immediate vs. Delayed Judgments

Although judgment accuracy tends to be low, metacognitive accuracy can be enhanced with relative ease, and not just with training. Receiving feedback about one's metacomprehension helps with metacognitive accuracy (Lee, Lim, & Grabowski, 2010). Furthermore, rereading (Rawson, Dunlosky, & Thiede, 2000), self-explanation during reading (Griffin, Wiley, & Thiede, 2008), and self-generation techniques have all been found to increase metacomprehension accuracy. Self-generation techniques include creating a title for the text (Morris, 1990), generating keywords (de Bruin, Thiede, Camp, & Redford, 2011; Shiu & Chen,

2013; Thiede, Dunlosky, Griffin, & Wiley, 2005), or generating summaries of the text (Anderson & Thiede, 2008; Thiede & Anderson, 2003). However, unlike other techniques, self-generation techniques such as summarizing only increase metacognitive accuracy if produced at a delay (Anderson & Thiede, 2008). When producing delayed summaries before making a metacomprehension judgment, relative accuracy increases from a gamma correlation of .27 to about .6 (Anderson & Thiede, 2008). Delayed summaries allow the individual to account for forgetting that may occur between a study session and a test (Anderson & Thiede, 2008). Immediate summaries, as well as delayed judgments without summarizing, are approximately equivalent in judgment accuracy as a control group, so delayed summaries were used in the current experiment to compare modalities (Anderson & Thiede, 2008; Maki, 1998a).

Metacognitive Cues

Another aim of metacognitive research is to reveal *how* individuals make judgments and was a pivotal question in the current study. The cue-utilization hypothesis is widely accepted in the field, and states that people cannot directly judge the strength of their memories and therefore must use a set of cues, or heuristics, to estimate the amount of information they know (Koriat, 1997). For example, if a word feels highly familiar, a person might predict that they will recognize it later. Researchers have identified multiple cues that influence metacognitive judgments, such as cue familiarity, self-efficacy, and fluency (Koriat, 1997; Metcalfe & Dunlosky, 2008), as well as study repetition and even font size (Pyc, Rawson, Aschenbrenner, & 2014; Sanchez & Jaeger, 2015). The current study will look at two cues known to influence judgments, the accessibility to information at retrieval and the situation model; these cues may account for differences in metacomprehension accuracy between modalities. Theoretical

accounts for the accessibility hypothesis and situation model hypothesis will each be described in turn.

The accessibility hypothesis explains how individuals use the amount of information they recall as a cue to make judgments. While making a judgment, recalling a large quantity of information might increase how confident an individual feels about their knowledge. Individuals who access a large amount of information at retrieval tend to have a higher judgment magnitude. Importantly, higher confidence does not necessarily mean higher metacognitive accuracy (Baker & Dunlosky, 2006; Koriat, 1993; Maki, Willmon, & Pietan, 2009; Morris, 1990). If the retrieved information is both correct and relevant, then the individual will likely make an accurate judgment because the heuristic used reflects knowledge accurately. In other contexts, a person might generate information that is incorrect, repetitive, or irrelevant, leading to overconfidence because the cue provides an inflated representation of their actual knowledge. The accessibility hypothesis predicts that while generating a summary, a high word count or number of total ideas will lead to a high prediction magnitude (Maki, Willmon, & Pietan, 2009), but not necessarily high accuracy.

Metacognition researchers have described the situation model as an additional cue that can be used to judge comprehension (Anderson & Thiede, 2008; Fukaya, 2013; Thiede & Anderson, 2003; Thiede et al., 2005). The situation model is one of three mental models created when reading text; the others are verbatim memories and propositional models (Kintsch, 1998; Kintsch & van Dijk, 1978; Perrig & Kintsch, 1985). Each of these contribute to understanding in a unique way. Verbatim memories represent the words that were used, and the propositional model describes the relationships between ideas. The situation model incorporates the overarching ideas of the text with inferences and past knowledge to create meaning from the

information read (Kintsch, 1998). When creating summaries immediately after reading, people may draw information from each type of mental representation: verbatim, propositional, and situation. However, the three representations together include redundant and irrelevant information, leading to a poor cue on which to base judgments and, therefore, inaccurate judgments. However, after time has passed, the verbatim and propositional models fade from memory (Kintsch, Welsch, Schmalhofer, & Zimny, 1990). The situation model creates meaning associated with past knowledge, so it is most resistant to forgetting compared to the other models discussed (Kintsch et al., 1990; Thiede et al., 2005). After a short time has passed, the verbatim and propositional models fade from memory; therefore, individuals who create a summary at a delay have access to this condensed, gist-based representation of the text and can make a more accurate judgment of their understanding.

Anderson and Thiede (2008) provided evidence for the situation model hypothesis by asking participants to write summaries of a text after a delay. The summaries were scored for number of gist-based ideas, number of details, and total ideas, with gist-based ideas indicating the use of the situation model. The number of gist-based ideas was the strongest predictor of relative accuracy. The situation model hypothesis predicts that high quality summaries in the current study will lead to higher metacognitive accuracy. The question remains whether the situation model is the strongest predictor of metacognitive accuracy for oral summaries as well.

Fluency is an additional cue that can influence prediction magnitude and accuracy (Benjamin, Bjork, & Schwartz, 1998), and fluency might differ between spoken and written summaries. Although there are different types of fluency, the focus was on retrieval fluency, or the ease in which a person retrieves information. Because fluency is closely related to response time (Bjork, Dunlosky, & Kornell, 2013), total summary time and latency to begin summarizing

were used to approximate fluency. Fluency may be a shallow cue to base judgments on, meaning that it does not necessarily represent information at the level of comprehension. As such, total time and shorter latency to begin summarizing would be associated with higher prediction magnitude, but not necessarily a higher prediction accuracy. Fluency was statistically controlled so that the effects of accessibility and situation model could be isolated.

It is important to note that multiple cues can impact metacognitive judgments and accuracy simultaneously (Morris, 1990), which might impact judgments in concurrence with the accessibility of information and the situation model. It was assumed that cues would not influence judgments equally, and that the same cue would be the biggest predictor of judgment magnitude for most people. Cues with a strong relationship with prediction magnitude are considered salient, as those are likely the cues on which participants base their judgments. Therefore, in this analysis the cue that accounted for the most variance in judgments will be the cue considered the most salient during summarization. The most salient cues may differ between conditions.

Are Speaking and Writing Summaries Equivalent?

Both written and oral summaries have been used in past research, but to my knowledge, they have never been compared in the context of metacomprehension accuracy. Written summaries are often used as a tool to improve metacomprehension accuracy (Anderson & Thiede, 2008; Maki, Willmon, & Pietan, 2009; Reid, Morrison, & Bol, 2017; Thiede & Anderson, 2003), but oral summaries are less commonly used in the literature (Baker & Dunlosky, 2006; Fukaya, 2013; Fulton, 2015). There are differences between the oral and written modalities that might lead to different metacognitive accuracy and a difference in preference as a study tool, so the differences between modalities should be investigated further. For example,

oral summaries contain more gist-based ideas and have more total words; additionally, repetitions and distortions are more characteristic of oral summaries (Kellogg, 2007; Vieiro & García-Madruga, 1997). Writing a summary tends to take longer (Fulton, 2015) and requires more effort than speaking a summary (Kellogg, 2007; McPhee, Paterson, & Kemp, 2014). Because the contents of the summaries are not equivalent, different cues are accessible at the time of comprehension judgments. If one summary type contains superior cues, it may facilitate improved metacomprehension accuracy.

Oral summaries provide a host of cues that may positively or negatively impact a person's judgment accuracy. Oral summaries produce a greater number of inference, or gist-based ideas (Kellogg, 2007; Vieiro & García-Madruga, 1997). The situation model hypothesis predicts summaries containing more inference-based ideas will lead to greater metacomprehension accuracy (Anderson & Thiede, 2008), so an oral summary with a greater number of gist-based ideas would lead to greater metacomprehension accuracy compared to a written summary. Although inferences as a cue might increase metacognitive accuracy, there seem to be multiple cues in oral summaries that may lead to lower judgment accuracy, including the length and fluency. Oral summaries were found to contain higher levels of distortions and more idea units in general (Kellogg, 2007). In this situation, length as a cue will likely induce overconfidence because the unrelated or unnecessary content does not indicate a higher level of understanding.

The accessibility hypothesis suggests that increasing the number of ideas correlates positively with judgment magnitude, but not necessarily accuracy; consequently, increasing word count and distortions can induce overconfidence. Additionally, spoken summaries take less time (Fulton, 2015) and may feel more automatic (Kellogg, 2007). The greater speed of summarizing

might feel more fluent than written summaries, which may also induce overconfidence (Benjamin, Bjork, & Schwartz, 1998) because greater summarizing fluency does not necessarily correlate with greater comprehension. Oral summaries include cues that promote higher metacognitive accuracy, such as a high amount of gist-based ideas, but they also contain cues that might induce overconfidence, such as distortions and overall length. Therefore, it was uncertain if oral summaries would be associated with more or less accurate judgments compared to written summaries.

Although oral summaries have cues that can allow for an accurate metacognitive judgment, written summaries have certain characteristics that might lead to a metacognitive advantage. Because written summaries take more time (Fulton, 2015; Kellogg, 2007), they might feel less fluent and lead to a less overconfident, or even under confident prediction. Unlike oral summaries, written summaries can be reread, allowing participants to review the details of what they wrote. The additional processing of their summary might benefit their memory or allow the writer to better evaluate their own knowledge, which can both uniquely and jointly affect metacomprehension accuracy. Furthermore, studies found that writing words may benefit memory more compared to speaking words aloud (Mama & Icht, 2016). However, these studies looked solely at word memory, not comprehension, and memory for a text is not equivalent to text comprehension (Kintsch, 1994). It is possible, but unlikely, that written summaries increase comprehension scores, which would influence metacomprehension accuracy. Rather, there are no hypothesized differences between comprehension scores.

In short, oral and written summaries differ in their characteristics, which may lead to a difference in the availability or salience of the cues. If one summary type contains cues that provide greater predictive validity, that modality will likely induce better metacognitive

monitoring. Although there is an expected difference in the metacomprehension accuracy between conditions, the direction and nature of that difference is yet unknown.

Perceived Cognitive Load Differences Between Summary Types

Kellogg (2007) proposed that speaking and writing differ in terms of working memory demand, which will influence a person's perceived cognitive load, a measurement in the current study. Working memory integrates and manipulates information required to complete the task at hand (Baddeley & Hitch, 1974). Working memory capacity is limited, meaning that one can only maintain a certain amount of information at a given time (Kane & Engle, 2003). When a greater amount of working memory is used to complete a task, the task feels more effortful (Paas, Renkl, & Sweller, 2003). Because the two summary types differ in the demands that they place on working memory, they will also differ in how effortful they feel.

Working memory has three primary components, including the visual spatial sketchpad, the phonological loop, and the central executive (Baddeley & Hitch, 1974). Speaking relies heavily on the central executive and the phonological loop, whereas writing relies on all three components (Galbrainth, Ford, Walker, & Ford, 2005). Thus, writing may tax working memory more than speaking (Bourdin & Fayol, 1994). Additionally, there is evidence that working memory supports both macrostructures (e.g. generating speech) and microstructures (e.g. spelling) of language (Vanderber & Swanson, 2007). Spoken summaries have a lower tax on working memory because some of the microstructures, such as spelling, are less prominent during speaking. Furthermore, the writing process is slower than speaking, so text representations may remain in working memory for longer, using more resources (Kellogg, 2007). Prior research indicates that writing taxes working memory more than speaking for both children (Bourdin & Fayol, 1994) and adults (Grabowski, 2010); participants prefer to speak

because writing is more effortful (McPhee, Paterson, & Kemp, 2014). The additional effort exerted during writing due to a higher working memory may influence predictions, acting as an additional cue.

Working memory demand is a form of cognitive load, defined as the amount of cognitive resources required to complete a task (Chandler & Sweller, 1991). Paas, Renkl, and Sweller (2003) conceptualized cognitive load as having three separate components: intrinsic, extraneous, and germane. These components describe respectively how the actual difficulty of the task, the presentation or environment of the information, and the motivation to complete the task all influence the perceived effort required. For example, a tricky puzzle would increase intrinsic cognitive load, but trying to complete it in a noisy café would lead to high extraneous cognitive load, and a love of puzzles would increase motivation and decrease germane load (Paas, Renkl, & Sweller, 2003). In the current study, the demand placed on working memory should differ between participants asked to produce spoken summaries and those asked to produce written ones. Therefore, the effort (perceived and actual) required to complete the summary should also differ. In the current research, perceived cognitive load, defined as the perceived mental effort required by the individual (Klepsch, Schmitz, & Seufert, 2017), was measured to assess the difference in perceived cognitive load between summary modalities.

Ultimately, the level of perceived cognitive load might act as an additional cue to influence metacognitive judgments. When a task is experienced as more difficult, students are less confident in their judgments (Maki et al., 2005, although see Moore, Lin, & Zabrucky, 2005 for counter evidence), and a lower confidence judgment might be more realistic. One study found that written summaries increased cognitive load in comparison to a control group but did not increase relative accuracy nor calibration (Reid, Morrison, & Bol, 2017). Unfortunately, this

study only analyzed written summaries, so no assumptions can be made about the level of perceived cognitive load for oral summaries. Even if perceived cognitive load does not act as a judgment cue, or not a major explanatory one, the difference in perceived cognitive load between conditions is still valuable to measure. If the results of the current experiment show that the oral and written conditions lead to equivalent metacognitive accuracy, but one modality leads to lower perceived cognitive load, then summarizing in the modality with lower perceived cognitive load could have an obvious advantage as a study tool.

The Present Study

The delayed summary technique has been found to increase metacomprehension (Anderson & Thiede, 2008), but is one modality superior? It was anticipated that summary modality would impact which cues were available or salient, which could then influence prediction magnitude. Both the situation model and the availability of information were expected to differ between spoken and written summaries, which would influence the accuracy of the predictions. The situation model and accessibility of information, as well as perceived cognitive load were measured so they could be compared across conditions and used as predictors of metacomprehension accuracy. Comparing the summary characteristics of the two modalities can shed light on which cues are more salient for participants in making prediction judgments, as well as suggest which summarization type might be more useful for students while studying.

Hypotheses

Hypothesis 1: The three groups will differ in their metacognitive accuracy.

1a: Metacomprehension prediction magnitude, but not comprehension scores, will differ between the oral, written, and control conditions.

The utilization of different cues is predicted to affect level of confidence while making judgments, which will lead to prediction magnitude differing between conditions. However, there are no manipulations between groups that are predicted to affect text processing, so overall comprehension should not differ between groups. Because the comprehension scores are not predicted to change, any difference in metacomprehension accuracy between groups will be driven by the difference in prediction magnitude.

1b: Metacomprehension prediction accuracy will differ between the oral, written, and control conditions.

If prediction magnitude differs between groups but comprehension does not, prediction accuracy will differ between groups. Both relative accuracy and absolute accuracy are predicted to differ between conditions. Because it is unknown which cues will be most salient for each modality, it is difficult to make a prediction about which modality will lead to greater accuracy.

Hypothesis 2: Summary characteristics will predict metacomprehension accuracy

2a: The oral and written conditions will differ in summary length, quality, latency to begin summarizing, and total time.

It is predicted that the oral and written summaries will differ in summary length (total words), quality as assessed by latent semantic analysis, latency to begin summarizing, and total time to complete the summary. Specifically, oral summaries will have higher summary length and quality, whereas written summaries will have greater total time, and longer latency.

2b: Summary characteristics will predict judgement magnitude.

Consistent with the cue utilization hypothesis, characteristics of the summary will influence prediction judgments. The accessibility hypothesis predicts that amount of information in the summary, measured by word count, should correlate highly with the judgment magnitude. According to this view, a high word count will lead to a high magnitude prediction that is not necessarily more accurate. If it is true that summary characteristics predict magnitude, they should predict magnitude regardless of modality, so participant summaries will be aggregated. However, given that modality may change the cues available to participants, summary modality was added to the model to test moderation of this effect.

2c: Summary characteristics will predict judgement accuracy.

If the situation model is a more salient cue, the summary quality will correlate with higher prediction accuracy. Similar to Hypothesis 2b, the summaries will be aggregated, and modality will be added to the model to test moderation of this effect.

Hypothesis 3: Perceived cognitive load will influence metacognitive judgment.

3a: Perceived cognitive load will be unequal between groups.

This will be an exploratory analysis. Because working memory will likely increase the most in the written condition (Kellogg, 2007), it is predicted that the control group will exhibit the lowest levels of perceived cognitive load, and the written summary group will exhibit the highest levels of perceived cognitive load.

3b: Perceived cognitive load will influence metacognitive judgment.

This will be an exploratory analysis that will probe the relationship between perceived cognitive load and prediction magnitude as well as prediction accuracy. Given that each

condition may change the participant's perceived cognitive load, participant condition will be added to the model to test moderation of this effect.

Chapter 2: Methods

Design

This experiment used a one-factor between subject design. Participants were randomly assigned to one of three groups, an oral summary group, a written summary group, and a control group.

Participants

Participants were recruited from classes at Idaho State University, using a program called SONA. For compensation, students received one SONA credit for each 30 minutes of participation. The students were required to be at least 18 years of age. Participants were excluded if they were diagnosed with a learning disability, an intellectual disability, or autism, because there is evidence that these populations have atypical metacognitive ability (Girli & Ozturk, 2017; Grainger, Williams, & Lind, 2014; Holzer, Madaus, Bray, & Kehle, 2009; Nader-Grosbois, 2014; Trainin & Swanson, 2005). Additionally, participants were screened for native language. If participants did not identify English as their first language, they were allowed to complete the study to receive SONA credits for their class, but their data were replaced to protect from differences due to language.

Two studies were used to estimate the sample size needed for sufficient power: one study compared metacomprehension accuracy with delayed summaries, and the other looked at recall differences between modalities. The effect size for a delayed summary approximates η^2=.21 (Anderson & Thiede, 2007). With this effect size and .80 power, the study will require 77

participants total. Using the effect size of recall in different modalities, *d*=.45, (Putnam & Roediger, 2013), the current study requires a total sample size of 95 participants at .80 power, and 120 participants at .90 power (Bausell & Li, 2002). Because the current study might be more nuanced, and to account for possible missing data, I aimed to recruit 105 participants.

Materials

The experiment was run on the program Eprime, which will display the texts and record participant predictions and post-dictions. Six texts that have been used in similar metacomprehension research were used (Fulton, 2015; Rawson & Dunlosky, 2002). These texts come from the Scholastic Aptitude Test (Board, 1997) and are at a Flesch-Kinsaid grade-level of 9.8-12.0 (*M*= 11.6). The titles of the texts are: Television Newscast, Precision of Science, Women in the Workplace, Zoo Habitats, American Indians, and Real vs Fake Art (see Appendix A for sample). Participants in the written condition typed their summaries in Eprime, which has been used in metacomprehension research in the past (Anderson & Thiede, 2008). Oral summaries were recorded with Audacity (http://audacity.sourceforge.net/) and transcribed for analysis. The participants then took the multiple-choice comprehension test, cognitive load survey, and demographic survey on Qualtrics. The cognitive load survey has been validated and shows strong reliability (Cronbach's α=0.81, Klepsch, Schmitz, & Seufert, 2017; see Appendix B). To score the cognitive load survey, participant answers were averaged across a 7-point Likert-type scale for each subscale (intrinsic, extraneous, germane).

Procedure

As an overview of the current procedure, students began by reading six different texts. Next, they produced summaries of the texts and made a prediction after each summary about

their performance on a multiple-choice test of that text. Finally, they took the multiple-choice test. Before debriefing, participants filled out a cognitive load survey and a demographic questionnaire.

Participants first read each of the six texts in which the order of presentation was randomized for each participant. The instructions stated that participants may be asked to summarize the texts. By having identical instructions for each group, average reading strategy should not differ between groups. The texts were displayed so that only one sentence appears at a time. Participants pressed a key after they have completed reading each sentence, blocking them from rereading. Rereading can lead to an increase in accuracy (Rawson, Dunlosky, & Thiede, 2000); preventing rereading assures that any difference in accuracy between groups is due to the summarizing modality. At the completion of each text, participants pressed a key to move to the title slide of the next text. The reading task had no time limit, but the time that it took to read each text was recorded with Eprime.

Once participants finished reading, they were asked to summarize one text at a time. The title of each text appeared as a prompt for them to begin summarizing. The summarizing task did not have a time limit, but completion time was recorded. After each summary was completed, participants made their multiple-choice comprehension predictions in Eprime. A prediction question was presented for each passage, which asked, “How many questions out of eight do you think you will answer correctly about this passage?” A key press presented the next title for them to summarize. For both conditions, the summary order did not necessarily match reading order; both reading order and summarizing order were randomized. This is considered summarizing at a delay, because the participants summarized after reading all the texts, instead of summarizing after each individual text (Thiede & Anderson, 2003).

The summaries were measured on four dimensions: length, quality, latency, and total time. Length was measured by a word count. Summary quality was measured using a technique called latent semantic analysis (LSA), which measures how closely the summary matches the semantics of an ideal summary (http://lsa.colorado.edu/; Landauer, 1998). LSA has been used in metacognitive research in the past (Maki, Willmon, & Pietan, 2009; Thiede & Anderson, 2003) and found to be comparable to a trained scorer (Landauer, 1998). Because this program measures semantic relatedness of words, if main ideas are used in the comparison summary, LSA can measure gist-based ideas in participant's summaries (Kintsch, 1998). Latency is defined as the time it takes for the student to begin summarizing. Total time is the amount of time it takes from the presentation of the summarizing prompt to the time it takes to finish summarizing and go to the next screen. Both total time and latency were recorded by Eprime.

The control condition did not summarize the text. The purpose of this condition was to assure that the delayed summarization manipulation successfully improves prediction accuracy, for if neither summary condition increases accuracy more than the control condition, then delayed summarizing loses its practical and possibly theoretical implications. The control participants read the texts as in the experimental conditions but were given a word search as an easy distraction task in place of generating summaries. The distraction task prevents the individuals from rehearsing information, which could influence their comprehension and metacomprehension. The duration of the distraction task was 15 minutes, which was determined by the length of the summarizing task during pilot testing. After an approximately equivalent time as the summarizing task, the control group made their multiple-choice comprehension predictions. Participants were shown the title of the text and asked to predict their multiple-choice performance, just like the two experimental conditions.

After completing predictions, all participants completed the multiple-choice comprehension test. The test was composed of 8 questions for each text. Once they finish each set of questions, participants took the cognitive load questionnaire. Finally, they completed a demographic survey and were debriefed about the study.

Statistical Analyses

Hypothesis 1: The three groups will differ in their metacognitive accuracy.

1a: Metacomprehension prediction magnitude, but not comprehension scores, will differ between the oral, written, and control conditions.

Both the judgment magnitude and comprehension scores were compared across groups using a one-way ANOVA. Then, a Tukey test was used to distinguish which groups differed in magnitude and in comprehension scores.

1b: Metacomprehension prediction accuracy will differ between the oral, written, and control conditions.

Relative accuracy was calculated by correlating (gamma and Stuart's tau-c) predictions and comprehension scores for each participant. The correlations were then averaged for each group and compared with a one-way ANOVA. In addition to this measure of relative accuracy, bias scores were analyzed. The predicted amount correct was subtracted from the actual amount correct for every summary of each participant, then averaged within each group. A one-way ANOVA was used to analyze differences in bias scores across groups, and a Tukey test was used to determine which groups differed from each other.

Hypothesis 2: Summary characteristics will predict metacomprehension accuracy

2a: The oral and written conditions will differ in summary characteristics.

Two-tailed t-tests were conducted for each measurement to assess differences between the conditions, including summary length, quality, latency, and total time (see page 18 for measurement details). A Bonferroni correction was used to adjust the inflation of the p-value.

2b: Summary characteristics will predict judgment magnitude.

A linear regression was used to measure the relationship between summary characteristics and judgment magnitude aggregated across modality. Length, LSA quality, latency, and total time were used to predict summary magnitude. Summary modality was an additional variable added in the regression to test if it is a moderator. A hierarchical regression was used, with latency and total time put into the model first to control for the effects of fluency, and length and quality put into the model second.

2c: Summary characteristics will predict judgment accuracy.

A linear regression was used to measure the relationship between summary characteristics and judgment accuracy. The summaries were aggregated, regardless of modality. Length, LSA quality, latency, and total time were used to predict summary accuracy. Summary modality was an additional variable added in the regression to test if it is a moderator. A hierarchical regression was used, with latency and total time put into the model first to control for the effects of fluency, and length and quality put into the model second.

Hypothesis 3: Perceived cognitive load will influence metacognitive judgment.

3a: Perceived cognitive load will be unequal between groups.

Perceived cognitive load was compared between groups with a one-way ANOVA.

3b: Perceived cognitive load will influence metacognitive judgment.

This was an exploratory analysis to use perceived cognitive load as a predictor for metacognitive judgments. Two regressions were used with cognitive load as a predictor; one regression used prediction magnitude as an outcome variable while the other used metacognitive accuracy as an outcome variable.

Chapter 3: Results

Demographics

Overall, 120 individuals participated in this study. Eleven people who did not speak English as their first language were replaced, and two individuals were replaced due to lack of audio. Four additional participants were excluded because of a technical difficulty, being highly distracted during the summarization task, and making the same prediction for every text. No participant had a diagnosed learning disability, intellectual disability, or autism. There were 103 participants included in the final analysis.

The sample was primarily white (86%) and female (70%). Ten percent identified as Hispanic. The mean age of the sample was 22.11 (*SD*=5.08), and participants were typically early in their college career, with 41.76% in their first year and 38.46% in their second year.

Judgment Magnitude and Multiple-Choice Performance

The grand mean for prediction magnitude was 4.99 (*SE*= .07) out of eight (see Table 1 for group means). On average, the participants scored 3.80 (*SE*= .07) on the multiple-choice assessment out of eight total questions. No group differences were found for prediction magnitude ($F(2, 615)=0.16$, $p=.85$, $\eta^2=.00$) nor for multiple-choice score ($F(2, 615)=0.23$, $p=.79$,

η^2=.00). These results show mixed support for our hypotheses, as group differences were expected for prediction magnitude, but multiple-choice performance was expected to be consistent across groups. This analysis suggests that summary modality does not influence comprehension judgments or comprehension.

Metacognitive Prediction Accuracy

Absolute accuracy. Bias scores were calculated by subtracting prediction magnitude from multiple-choice performance. Participants were somewhat overconfident with an overall average bias score of 1.19 (*SE*= .08; see Table 1 for group means). This was significantly different from zero ($t(617)= 14.16$, $p<.01$), suggesting there is room for improvement in calibration. Bias scores were statistically equal across all groups ($F(2, 615)=0.19$, $p=.83$, $\eta^2=.00$). Modality did not impact absolute accuracy in participants, contrary to the hypothesized effect.

Relative accuracy. The average gamma correlation was .13 (*SE*= .05) for participants, and although small, is significantly different from zero ($t(102)= 2.40$, $p=.02$), suggesting that on average, participants were above chance at distinguishing on which texts they would score well. Similarly, Stuart's tau-c was .11 (*SE*= .04), and significantly different from zero ($t(102)= 2.74$, $p=.01$). Because gamma and Stuart's tau-c lead to the same conclusion for all analyses, I have focused on gamma correlations because it is more widely accepted.

There was a difference in average gamma correlations between conditions ($F(2, 100)=3.61$, $p=.03$, $\eta=.07$), supporting the hypothesis that relative accuracy would differ between groups. Using a Tukey test, the written condition had the highest average gamma correlation (*M*=.31, *SE*=.09; Figure 2), which differed significantly from the control condition (*M*=.005, *SE*=.09, 95% CI [.008, .618]; *p*=.043). The written condition only marginally differed from the

oral condition (*M*= .04 *SE*=.10, 95% CI [-.031, .579]; *p*= .088). The oral and control condition did not differ from each other (95% CI [-.353, .274]; *p*= .95). Another way to view these data is to see which group averages differ from zero, as this indicates ability to discriminate between texts more and less well understood. The written condition gamma correlation differed significantly from zero ($t(36)= 2.44$, $p=.02$), but neither the oral ($t(32)= 0.30$, $p=.66$) nor control ($t(32)= 0.05$, $p=.95$) conditions differed from zero. Thus, the written summary condition was the only group that showed some ability to discriminate between texts more and less well understood.

Summary Characteristics

Most summary characteristics differed between the written and oral conditions, as hypothesized (Table 2). The written summaries on average were higher quality ($t(418)= 2.95$, $p<.01$, $g=.28$), took longer to complete ($t(418)= 18.05$, $p<.01$, $g=1.77$), and were quicker to begin ($t(418)= -8.02$, $p<.01$, $g=.59$). All tests were Bonferroni corrected, with $p= .0125$. The oral and written summaries did not differ in word count ($t(418)= -0.84$, $p=.40$, $g=0.04$), contrary to the hypothesis.

As hypothesized, summary characteristics significantly predicted judgment magnitude, accounting for about 15% of the variance ($R^2=.15$, $F(4, 402)=17.67$, $p<.01$, $\eta^2=.15$; Table 3). To control for fluency effects, only total time and latency were added to the first model. They accounted for a significant amount of variance in prediction magnitude ($R^2=.02$, $F(4, 402)=4.87$, $p=.01$). When word count and LSA score were added to the second model, the model significantly improved ($R^2=.15$, $\Delta R^2= .13$, $\Delta F(4, 402)=30.80$, $p<.01$), showing that summary quality and amount of information account for a significant amount of the variance above and beyond total time and latency. However, only word count ($\beta=.28$, $p<.01$) and LSA score ($\beta=.21$,

p<.01) were significant independent predictors of judgment magnitude. Total summary time (β=-.07, p=.17) and latency to begin summarizing (β=-.05, p=.27) were not significant. Added to the third model, condition did not account for additional variance (R^2=.15, ΔR^2= .00, ΔF(4, 402)=0.01, p=.94). Furthermore, when condition was used as a moderator, the model was not improved, (R^2=.16, ΔR^2= .01, ΔF(4, 402)=2.08, p=.08), suggesting that modality did not change how the participants used cues.

Summary characteristics did not influence bias scores (R^2=.01, F(4, 402)=1.20, p=.32, η^2=.01), which is counter to the hypothesis. They did, however, influence relative accuracy (R^2=.10, F(4, 402)=9.28, p<.01, η^2=.09). Using a hierarchical regression, total time and latency were first added to the model, which was significant (R^2=.06, F(4, 402)=13.25, p<.01; Table 4). When word count and LSA score were added to the model, the model improved (R^2=.09, ΔR^2= .03, ΔF(4, 402)=6.50, p<.01), indicating that these two variables accounted for a significant amount of the variance in relative accuracy above and beyond total time and latency. Interestingly, the two characteristics that accounted for most of the variance were word count (β=-.19, p<.01), which was negatively related to relative accuracy, and total time (β=.24, p<.01), which was positively related to relative accuracy. Therefore, whereas relative accuracy increases as total time increases, relative accuracy decreases as word count increases. Neither LSA score (β=.06, p=.23) nor latency (β=.03, p=.59) were significant predictors of relative accuracy. Condition was then added to the model, and the model was significantly improved (R^2= .10, ΔR^2= .01, ΔF(4, 402)=5.00, p=.026), reflecting the group differences discussed earlier. However, when condition was used as a moderator, the model was not improved, (R^2=.11, ΔR^2= .01, ΔF(4, 402)=1.11, p=.35), suggesting that modality did not affect the way summary characteristics influenced relative accuracy.

Cognitive Load

Three different types of cognitive load were measured using the Cognitive Load scale. They were not added to create an aggregate cognitive load scale because the original authors view the subscales as qualitatively different (Klepsch, Schmitz, & Seufert, 2017); therefore, the subscales were run separate for the analysis. There were found to be differences for intrinsic ($F(2, 99)=3.14$, $p=.048$, $\eta^2=.06$) and extraneous ($F(2, 99)=4.08$ $p=.02$, $\eta^2=.08$) cognitive load, but not germane load ($F(2, 99)=1.06$, $p=.35$). Counter to the prediction, a Tukey test showed that the oral condition ($M= 5.77$, $SE=.18$; Figure 3) reported a significantly higher intrinsic cognitive load than the written condition ($M=5.16$, $SE=.17$; 95% CI [.03, 1.19], $p=.01$). The control condition did not differ significantly from either the oral condition ($M=5.41$, $SE=.18$; 95% CI [-.23, .97]) or the written condition (95% CI [-.34, .83], $p=.01$). Extraneous cognitive load showed similar results, with the oral condition ($M= 4.30$, $SE=.24$) reporting significantly higher cognitive load than the written condition ($M= 3.37$, $SE=.23$; 95% CI [.14, 1.70], $p=.01$). Again, the control condition did not differ from the oral condition ($M= 3.95$, $SE=.24$; 95% CI [-.46, 1.16]) nor control (95% CI [-.21, 1.37]).

Regression analyses were used to measure the correspondence of cognitive load to prediction magnitude and the two types of metacomprehension accuracy. Three simultaneous regressions were run with the cognitive load subscales included as criterion variables (intrinsic, extraneous, and germane), with a different regression for each of the outcome variables (prediction magnitude, relative accuracy, absolute accuracy). Cognitive load was associated with prediction magnitude ($R^2=.12$, $F(3, 98)=4.47$, $p<.01$; Table 5). Intrinsic load was the only significant predictor of magnitude ($\beta=-.33$, $p<.01$). Condition did not moderate this effect ($R^2=.12$, $\Delta R^2= .00$, $\Delta F(3, 98)=0.30$, $p=.83$), suggesting that modality did not impact the

relationship between cognitive load and prediction magnitude. Cognitive load did not predict either bias scores (R^2=.06, F(3, 98)=2.15, p=.10, η^2=.06) or gamma correlations (R^2=.02, F(3, 98)=0.54, p=.65, η^2=.02). The moderation analysis was not used because the model was not significant. To conclude, cognitive load seems to be a cue for predictions but not a diagnostic one.

Chapter 4: Discussion

The results of the present study provide valuable information about how the modality of delayed summaries influences metacognitive accuracy. Summary modality did not affect prediction magnitude, multiple-choice performance, or bias scores. This suggests that modality does not influence the level of over-confidence or under-confidence at prediction. However, summary modality impacted relative accuracy. It should be noted that no correlation was found between absolute and relative accuracy in past research (Kelemen, Frost, Weaver, 2000; Maki et al., 2005), so it is not improbable that group differences were found for relative but not absolute accuracy. Given that only the average gamma correlation of the written condition differed from zero and the control condition, there must be different cues available during written summarization leading to improved accuracy, or cues are more salient or diagnostic for written compared to oral summaries. This assumption is supported by the analysis of summary characteristics. Additionally, cognitive load differed between the oral and written conditions, seemingly influencing prediction magnitude. Each of the results will be further evaluated in turn, and the applications and future directions of this study will be discussed.

Absolute Accuracy

Participant calibration did not vary between conditions. From a statistical perspective, this is unsurprising because neither prediction magnitude nor comprehension performance

differed on average between groups, so mathematically it would be unlikely for bias scores to differ between groups. Yet, it was still counter to the hypothesis. One previous study found similar results, with no differences in bias scores between a delayed written summary and a no summary condition (Dunlosky, Rawson, & Middleton, 2005). However, it was expected that there would be differences in bias scores between the oral and written condition because the cue availability differed between groups. Word count and summary quality will be used to explain the non-significant difference in prediction magnitude between groups as fluency (total time and latency to begin) did not relate to prediction magnitude. Based on the accessibility hypothesis, it was suspected that a higher word count in the oral condition would drive overconfidence compared to the written condition (Kellogg, 2007; Koriat, 1995); however, word count did not differ between groups, and, therefore, could not lead to differences in bias scores between groups. Although LSA did relate to higher prediction magnitude, it did not lead to differences in absolute accuracy. Most studies that measured summary quality only related it to relative accuracy, so although summary quality has been shown to relate to prediction magnitude (Anderson & Thiede, 2008; Maki, Willmon, & Pietan, 2009), there is less evidence to support summary quality affecting calibration. Moreover, many studies have found calibration to be resistant to change (Foster, Was, Dulosky, & Isaacson, 2017; Kelemen, Frost, & Weaver, 2000; Maki et al., 2005), so perhaps the similarity in absolute accuracy between groups should not be surprising.

Some researchers found they can manipulate average absolute accuracy (Callender, Franco-Watkins, & Roberts, 2016; Dunlosky, Rawson, & Middleton, 2005; Koriat, 1997), but still others suggest that absolute accuracy is stable compared to relative accuracy (Foster et al., 2017; Kelemen, Frost, & Weaver, 2000; Maki et al., 2005; Nietfeld & Schraw, 2002). Absolute

accuracy was shown to have strong test-retest reliability over a two-week period, while this was not the case for relative accuracy (Keleman, Frost, & Weaver, 2000). It is possible that cues must be very conspicuous to affect absolute accuracy. Although the average LSA score differed between the oral and written conditions, the difference in means was small (.54 and .59, respectively), so maybe the difference in summary quality as a cue was not pronounced enough to shape calibration.

Relative Accuracy

Despite the non-significant impact of summary modality on bias scores, summary modality did influence relative accuracy of the participants. Only the written modality condition exhibited relative accuracy above chance, suggesting there are fundamental differences between oral and written summaries that lead to this difference. To understand why only written summaries improved relative accuracy above chance, summary characteristics were analyzed in relation to both prediction magnitude and relative accuracy. Although LSA and word count related to prediction magnitude, neither of these were diagnostic cues as LSA did not predict relative accuracy and word count's relationship to relative accuracy was negative. Total summary time did not affect prediction magnitude, but it *did* predict relative accuracy. This suggests that total time was not a cue per se, as participants were not actively basing their judgments on total time; rather, total summary time is a characteristic that allowed for higher relative accuracy.

Summary modality did not moderate the relationship between cues and relative accuracy, but regressions for the oral and written condition were run separately to further assess the possibility that the two groups used cues differently. This analysis revealed that relative accuracy related to the cues differently between conditions, as word count and total time predicted relative

accuracy in the written condition, yet none of the measured cues related significantly to relative accuracy in the oral condition. This suggests that the cues were less diagnostic or more difficult to judge in the oral condition. Total summary time can provide a possible explanation of why participants in the written condition were better able to judge their comprehension.

Total summary time was the summary characteristic that accounted for the most variance in relative accuracy. There are several reasons why this might be the case. Because written summaries are slower than oral summaries, it was expected that the lower fluency in the written condition would lead to greater relative accuracy than the oral condition. However, there was mixed evidence that the written condition was disfluent—they took longer to summarize but were quicker to begin, potentially because those in the oral condition used more time to plan what they wanted to say, but produced their summaries more quickly. And, the written condition did not have a lower prediction magnitude than the other groups, so if they were experiencing disfluency they were not using it as a cue. It is uncertain that total time and latency measured fluency in the intended way because of its lack of relationship with prediction magnitude; future studies might use words per minute as a better measure of fluency. The rate of word production will give a more precise measure of fluency rather than total time, as fluency is related to speed of response. Total time may capture other factors, like rereading in the written condition. Additionally, the literature on disfluency is mixed; sometimes the disfluency effect is not replicated (Bjork & Yue, 2016). As the disfluency effect is unlikely to account for the results, a different interpretation needs to be considered.

As disfluency is not the most viable explanation, other factors are likely causing the increase in total time and improvement in relative accuracy for the written condition. Those in the written condition could have spent more time evaluating whether they have typed all the

information that they knew, or spent time comparing how they felt on each summary in comparison to the others. However, if participants were spending more time evaluating their summaries, then one would expect that total time would also predict relative accuracy in the oral condition, but that is not what was found. A clear advantage for the written condition is that participants had the ability to reread their summaries. Research shows that rereading texts increases relative accuracy (Griffin, Wiley, & Thiede, 2008; Rawson, Dunlosky, Thiede, 2000). One hypothesis proposes that rereading increases the available attentional resources, so more resources are shunted to monitoring rather than comprehension (Dunlosky & Rawson, 2005). Although there is no study to my knowledge that tests rereading during summarization, this hypothesis can explain the increased relative accuracy of the written condition. Whereas the participants from the oral condition must use their attentional resources to produce summaries, those in the written condition can first focus on producing summaries and then focus on making accurate judgements upon re-reading their summaries. This might also explain why the written condition reported the lowest cognitive load; the ability to assess the summaries after rereading may have decreased perceived cognitive load compared to the oral condition, in which participants had to produce and monitor their summaries concurrently. Unfortunately, rereading during the summarization process was not measured; future research might restrict participants to summarize one sentence at a time, so that rereading could be controlled or measured.

Accessibility Hypothesis vs Situation Model

Which hypothesis has more support, accessibility or situation model? Although word count did not differ between the two groups, both LSA score and word count related to prediction magnitude regardless of condition. This demonstrates that both the accessibility of information and situation model are cues that influence prediction magnitude. Notably, the larger

standardized beta for word count suggests participants weighed the accessibility of information greater as a cue. Out of the two cues, only word count predicted gamma correlation. However, the relationship was negative, suggesting that fewer words led to greater relative accuracy. Thus, although both cues contributed to participant predictions, neither cue was diagnostic of metacognitive accuracy in our sample.

The negative relationship between word count and relative accuracy supports the idea that the accessibility of information is not a valid cue, as higher number of words led to higher predictions, but lower relative accuracy, supporting past research (Baker & Dunlosky, 2006; Koriat, 1993; Morris, 1990). Some might interpret these results to mean that word count is not beneficial to the participant, but this is an oversimplification. Those with higher word count generally had higher predictions, as well as higher multiple-choice scores (Table 6). It could be that longer summaries may have decreased the utility of word count as a diagnostic cue. For example, in a low word count scenario, a person may easily judge that their 40 word summary is larger than their 20 word summary, as this is twice the amount of words, and a huge relative increase. In comparison, the difference between 100 and 120 is a much smaller relative change, even though both examples have the same absolute difference. Therefore, those who wrote in abundance might have had a more difficult time judging summaries based on word count.

It was unanticipated that LSA did not predict relative accuracy because LSA measures the situation model, and the situation model should be a valid cue as it approximates comprehension (Anderson & Thiede, 2008; Fukaya, 2013; Kintsch, 1998; Thiede & Anderson, 2003; Thiede et al., 2005). Yet, this study provides evidence that summary quality as measured by LSA is not a valid cue to base judgments on. There are several reasons that this could be the case. First, summary quality and multiple-choice questions both measure comprehension, but in

different ways (Fulton, 2015; Ko, 2010). For additional evidence, a regression was run using multiple-choice score as the outcome variable and the four summary characteristics as the predictors. Although the regression was significant (R^2=.10, F(4, 402)=11.11, p<.01), LSA did not predict multiple choice comprehension (B=.72 (.49), p=.14). Therefore, it could be that LSA did not predict relative accuracy because it does not reliably predict multiple-choice comprehension. This conflicts with past research, as the situation model is the proposed mechanism for the increase in relative accuracy in a number of studies (Anderson & Thiede, 2008; Fukaya, 2013; Thiede & Anderson, 2003; Thiede et al., 2005). A different reason could be that participants were not using the situation model to make their judgments. The relationship between LSA and prediction magnitude could be caused by a third variable, which could be word count or a cue that was not measured. There was a correlation between word count and LSA (Table 7), which could have led to the relationship between LSA and prediction magnitude. Students do not always use the most diagnostic cues. Thiede and colleagues (2010) found that when students were asked explicitly about cue use during judgments, they were more likely to report cues like level of interest or memory for the text rather than comprehension (Thiede, Griffin, Wiley, & Anderson, 2010). This supports the possibility that the correlation between LSA and a different cue (e.g.word count) is spurious, rather than students using summary quality as a cue.

Many studies show the positive benefit of the situation model, so it is possible that this study simply failed to find an effect, and that the situation model is generally a diagnostic cue despite the current findings. Participants use multiple cues to make their judgments (Morris, 1990; Undorf, Söllner, & Bröder, 2018); perhaps participants relied more heavily on word count, or other cues not measured. Therefore, even though LSA may have been considered in the

judgment, it might not have been the biggest contributor to participants' judgments. Finally, it must be acknowledged that even the situation model is a heuristic and heuristics are not a direct analysis of comprehension (Koriat, 1997). Cues that are generally diagnostic can also be misleading; an individual relying on the situation model might have incorrect or irrelevant information that leads to inaccurate judgments. Participants may account for the situation model while making judgments, but when this imperfect cue is used in conjunction with other imperfect cues, especially if participants weigh the non-diagnostic cues more heavily, it is certainly possible for LSA to have a weak relationship to relative accuracy.

Perceived Cognitive Load

Perceived cognitive load was another unreliable cue on which individuals seem to base their judgment. As cognitive load increased, judgment magnitude decreased, but perceived cognitive load did not relate to either relative or absolute accuracy. This is true for both intrinsic and extraneous cognitive load. Although this was the first time perceived cognitive load was assessed as a cue, this finding parallels past research regarding difficulty and metacomprehension. Difficult texts can lead to less confident and, therefore, more accurate judgments (Maki et al., 2005), although this effect can be moderated by working memory capacity (WMC) and verbal ability (Ikeda & Kitagami, 2013; Maki et al., 2005). Perceived cognitive load measured in the current study is a function of both WMC and difficulty (Paas, Renkl, & Sweller, 2003), and so it could be that perceived cognitive load could be a better cue to use for some participants. The metacognitive task seemed difficult for our participants, as they did fairly poorly on average on the multiple-choice test and in judging their accuracy. Therefore, maybe only those with high verbal ability or WMC were successfully able to utilize perceived cognitive load as a cue, while cognitive load may have been a less diagnostic cue for low

performing students. Future studies should look specifically at cognitive load to determine its validity as a cue, as well as the influence of individual differences in WMC.

Although cognitive load differed between conditions, the direction was opposite of the predicted effect, with oral summaries having the highest cognitive load. This contradicts past research that suggests written summaries have higher cognitive load due to grammar, spelling, and slower processing, and written summaries are reported as more effortful (Grabowski, 2010; Kellogg, 2007; McPhee, Paterson, & Kemp, 2014). Anecdotally, participants were not fond of the spoken summaries. Many of them were nervous, and they seemed more likely to ask questions about the summaries (although this was not measured). For example, multiple participants asked the researcher if it was "weird" to listen to them summarize aloud. One participant asked if the researcher could leave the room while they summarized. This condition may have unintentionally evoked anxiety, which has been shown to increase cognitive load (Eysenck, Derakshan, Santos, & Calvo, 2007). Further research could disentangle the actual difference between oral and written summaries without the influence of anxiety. A follow-up to the current study could test the difference between a researcher being present or absent during oral summarization.

Although cognitive load likely acts as a metacognitive cue, these results should be interpreted with caution. Cognitive load was probed at the end of the task, once the multiple-choice test was complete. When answering the cognitive load survey, participants were asked to focus on the overall task rather than one specific section, so it is unknown if participants were basing their cognitive load responses on their summary performance, their performance on the multiple-choice test, etc. Their memories of their cognitive load during summarization could be biased, and, therefore, it is uncertain if cognitive load acts as a cue for predictions, or if factors

such as how the participant thinks they scored on the test actually influenced perceived cognitive load. Thus, this measure might reflect their postdictions of cognitive load. Follow up studies should consider using cognitive load after each summary in the procedure to test if cognitive load operates as a cue, or if it is a byproduct of hindsight.

Limitations

This study is limited in several ways. First, some of the oral summaries may be compromised due to technological complications. Although this would not affect latency to begin summarizing or total time (because they were measured with Eprime), it may have influenced word count and LSA score (measured with Audacity), which could affect the credibility of the regressions that include summary characteristics. Additionally, there is no way to know if the cues that were measured are the primary cues on which our participants based their judgments. In fact, assuming that the regression perfectly captured how participants used the cues that were measured, it only accounted for 15% of the variance, suggesting there are other cues on which students base their judgments. Last, our measure of cognitive load occurred at the end of the study and may not have captured cognitive load of the summaries. This procedure can benefit from measuring cognitive load immediately after summarization to get a better idea of how cognitive load influences judgments.

Additionally, the average relative accuracy of participants was remarkably low. Typically, relative accuracy is above chance, even if poor, with a gamma correlation of .27. However, many of the participants had a relative accuracy that was not only poor, with the average gamma correlation for the oral and control conditions being .04 and .01, respectively. Furthermore, many participants had gamma correlations in the negative direction, meaning that they predicted higher scores for texts that they did more poorly on. In this study, written

summaries increased relative accuracy better than chance, but only to the level that is typical in the literature without interventions. Two ideas might explain this remarkably below-average relative accuracy. First, the state in which this sample was collected has one of the lowest ranked K-12 education systems in the United States, ranked 45 in the nation (Education Week, 2018), and according to collegeboard.org, about 91% of the students at Idaho State University are from in state. The average multiple-choice score in this study was 48%. Previous studies that used this same multiple-choice test have shown average multiple-choice scores of at least 60% (Fulton, 2015; Miele & Molden, 2010; Rawson & Dunlosky, 2002). Higher verbal ability is correlated with better relative accuracy (Hacker et al., 2000; Maki et al., 2005), so although neither verbal ability nor working memory were measured, it could be that poor relative accuracy in this sample is driven by low performing students. Additionally, the current study prevented participants from rereading the passages. This may have made the task unintentionally difficult and lowered relative accuracy. Rereading tends to increase relative accuracy and is more naturalistic for students. If the participants were not limited on rereading, perhaps they would have demonstrated a more typical relative accuracy.

Implications and Future Directions

This study provides support for the availability heuristic influencing prediction magnitude with summary quality as a secondary cue. However, with the current analysis, it did not appear that either the availability of information or the situation model were valid cues on which to base judgments. Additionally, this study provides evidence that students should implement written summaries to increase their relative accuracy. Improved relative accuracy can increase the effectiveness of self-regulated learning (Dunlosky et al., 2005). By having a stronger sense of what information they learned best, students can allocate their study time to the less

well-learned information to increase academic performance (Metcalfe & Finn, 2008). However, students should be aware that spoken summaries do not have this same effect. Even though they are faster, oral summaries did not show a metacognitive benefit, and might be perceived as more difficult by students.

Although this study shed light on the research question, it led to many more questions that are worth investigating. One possible extension of this study could look at more naturalistic settings for oral summaries, which may reduce anxiety during oral summarization. For example, a student might be instructed to summarize the passage as if teaching another student, as if they were in a study group. This added "teaching" component might make students more comfortable and can potentially alter anxiety level and the content of the summary, and thus the available cues. Alternatively, state anxiety can be measured and correlated with metacognitive judgments and accuracy. Additionally, using a within-subjects design to test both modalities for each participant may improve this study. A within-subjects design might increase sensitivity to the manipulation, as participants can compare cues across modalities. This comparison could make the difference in cues between modalities very salient, possibly affecting accuracy for better or for worse.

Further, the cognitive load aspect of the current study is very interesting, but the method can be improved. Changing the timing of the cognitive load measurement can provide higher quality data on perceived cognitive load. By inserting the cognitive load survey immediately after the summaries, a more precise measure of cognitive load can be derived. A different extension can investigate the effect of oral summaries without the possible influence of anxiety. Anxiety can possibly be reduced by not having a researcher present during summarization, or giving students practice with the microphone. Given that students are typically studying in a

comfortable environment, if oral summaries lead to an adequate increase in relative accuracy when students are relaxed, then perhaps oral summaries could be a suitable technique for some students. Further research can benefit from using summary modality as a tool to manipulate cue availability, which can expand the cue utilization hypothesis by demonstrating the contexts in which certain cues are valid. Learning more about when and why cues are diagnostic of comprehension may increase study efficiency and academic performance for students.

References

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Pyschologica, 128,* 110-118. doi: 10.1016/j.actpsy.2007.10.006

Ariel, R. (2013). Learning what to learn: the effects of task experience on strategy shifts in the allocation of study time. *Journal of Experimental Psychology, 39(6),* 1697-1711. doi: 10.1037/a0033091

Baddeley, A., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.) (Vol. 8, pp. 47–89). Academic Press.

Baker, J. M., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgements? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review, 13(1),* 60-65.

Bausell, R. B., & Li, Y. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. Cambridge: Cambridge University Press.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). This mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127 (1),* 55-68.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. Annual Review of Psychology, 64, 417–444.doi: 10.1146/annurev-psych-113011-143823.

Bjork, R. A., Yue, C. L. (2016). Commentary: Is disfluency desirable? *Metacognition and Learning, 11*(1), 133-137.

Board, T. C. (1997). 10 real SATs. New York: College Entrance Examination Board.

Bourdin, B., & Fayol, M. (1994). Is written language production more difficult than oral language production? A working memory approach. *International Journal of Psychology, 29*(5), 591-620.

de Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109,* 294-310.

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*, 215-235.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction,* 8(4), 293-332.

Bigfuture.collegeboard.org. (2019). *Idaho State University - ISU - The College Board*. [online] Available at: https://bigfuture.collegeboard.org/college-university-search/idaho-state-university?searchtype=college&q=Idaho%2BState%2BUniversity [Accessed 20 Jun. 2019].

Dunlosky, J., Hertzog, C., Kennedy, M. R. T., & Thiede, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology, 10(1),* 4-11.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: a brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16(4),* 228-223.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgements? Testing the transfer-appropriate-monitoring and accessibility hypothesis. *Journal of Memory and Language, 52,* 551-565.

Education Week. (2018). *State Grades on Chance for Success: Map and Rankings*. [online] Available at: https://www.edweek.org/ew/collections/quality-counts-2018-state-achievement-success/state-grades-on-chance-for-success-map.html [Accessed 20 Jun. 2019].

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46(1),* 6-25. doi: 10.1080/00461520.2011.538645

Epstein, W., Glenberg, A. M., & Bradley, M. M. (1984). Coactivation and comprehension: Contribution of text variable to the illusion of knowing. *Memory & Cognition, 12(4),* 355-360.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion, 7*(2), 336-353.

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

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: the role of memory for past exam performance. *Metacognition and Learning, 12,* 1-19.

Fukaya, T. (2013). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition Learning, 8,* 1-18. doi: 10.1007/s11409-012-9093-0

Fulton, E. K. (2015). *Can younger and older adults judge the quality of their text summaries?* (Unpublished doctoral dissertation). Georgia Institute of Technology, Atlanta, Georgia.

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,* 113–145. http://dx.doi.org/10.1007/s10674-005-0119-2

Girli, A., & Oztruk, H. (2017). Metacognitive reading strategies in learning disability: relations between usage level, academic self-efficacy and self-concept. *International Journal of Elementary Education, 10*(1). doi: 10.26822/iejee.2017131890

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11(4),* 702-718.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10(6),* 597-602.

Grabowski, J. (2010). Speaking, writing, and memory span in children: output modality affects cognitive performance. *International Journal of Psychology,45*(1), 28-39.

Grainger, C., Williams, D. M., & Lind, S. E. (2014). Metacognition, metamemory, and mindreading in high-functioning adults with Autism Spectrum Disorder. *Journal of Abnormal Psychology, 123*(3), 650-659. doi: 10.1037/a0036531

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36(1),* 93-103. doi: 10.3758/MC.361.93

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92(1),* 160-170. doi: 10.1037//0022-0663.92.1.160

Holzer, M. L., Madaus, J. W., Bray, M. A., & Kehle, T. J. (2009). The test-taking strategy intervention for college students with learning disabilities. *Learning Disabilities Research and Practice, 24(1),* 44-56.

Ikeda, K., & Kitagami, S. (2013). The interactive effect of working memory and text difficulty on metacomprehension accuracy. Journal of Cognitive Psychology, 25(1), 94-106.

Keleman, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28*(1), 92-107.

Kellogg, R. (2007) Are written and spoken recall of text equivalent? *American Journal of Psychology, 120 (3),* 415-428.

Kintsch, W. (1994). Text comprehension, memory, and learning. American Psychologist, 49(4), 294-303. doi: 10.1037/0003-066x.49.4.294

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85(5), 363-394. doi: 10.1037/0033- 295x.85.5.363

Kintsch, W., Welsch, D.M., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29,* 133-159.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8.* doi: 10.3389/fpsyg.2017.01997

Ko, M. H. (2010). A comparison of reading comprehension tests: Multiple-choice vs. open-ended. *English Teaching, 65*(1), 137-159.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychology Review, 100(4),* 609-639. doi: 10.1037/0033-295X.100.4.609

Koriat, A., (1995). Dissociating knowing and the feeling of knowing: further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*(3), 311-333.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgements of learning. *Journal of Experimental Psychology: General, 126(4),* 349-370.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(3),* 609-622.

Kwon, H., & Linderholm, T. (2014). Effects of self-perception of reading skill on absolute accuracy of metacomprehension judgements. *Current Psychology, 33,* 73-88. doi: 10.1007/s12144-013-9198-x

Landauer, T. K. (1998). Learning and representing verbal meaning: the latent semantic analysis theory. *Current Directions in Psychological Science, 7,* 161

Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Education Technology Research Development, 58,* 629-648. doi: 10.1007/s11423-010-9153-6

Lin, L.-M., & Zabrucky, K. M. (2000). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23,* 345-391.

Lin, L.-M., Moore, D., & Zabrucky, K. M. (2000). Metacomprehension knowledge and comprehension of expository and narrative text among younger and older adults. *Educational Gerontology, 26,* 737-749.

Mama, Y., & Icht, M. (2016). Auditioning the distinctiveness account: expanding the production effect to the auditory modality reveals the superiority of writing over vocalizing. *Memory, 24(1),* 98-113.

Maki, R. H. (1998a). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition, 26(5),* 959-964

Maki, R. H. (1998b). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Metacognition in educational theory and practice (pp. 117–144). Mahwah, NJ: Erlbaum.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of Text Material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(4),* 663-679.

Maki, R. H., & Serra, M. (1992). The basis of test prediction for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(1),* 116-126.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Education Psychology, 97(4),* 723-731.

Maki, R. H., Willmon, C., & Pietan, A. (2009). Basis of metamemory judgments for text with multiple-choice, essay and recall tests. *Applied Cognitive Psychology, 23,* 204-222. doi: 10.1002/acp.1440

McPhee, I., Paterson, H. M., & Kemp, R. I. (2014). The power of the spoken word: can spoken-recall enhance eye-witness evidence? *Psychiatry, Psychology, and Law, 21(4),* 551-556. doi: 10.1080/13218719.2013.848001

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15(1),* 174-179. doi: 10.3758/PBR.15.1.174

Miele, D. B., & Molden, D. C. (2010). Naïve theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General, 139*(3), 535-557. http://dx.doi.org/10.1037/a0019745

Moore, D., Lin, L.-M., & Zabrucky, K. M. (2005). A source of metacomprehension inaccuracy. *Reading Psychology, 26,* 251-265.

Morris, C. C. (1990). Retrieval Processes Underlying Confidence in Judgements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(2),* 223-232.

Nader-Grosbois, N. (2014). Self-perception, self-regulation and metacognition in adolescents with intellectual disability. *Research in Developmental Disabilities, 35,*1334-1348.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological Bulletin, 95, 109–133.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation, 26,* 125-173.

Nietfiled, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition Learning, 1,* 159-179. doi: 10.1007/s10409-006-9595-6

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*(3), 131-142.

Paas, F., Renkl, A. & Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educational Psychologist, 38*(1), 1-4.

Perrig, W., & Kintsch, W. (1985). Propositional and situational representations of text. *Journal of Memory and Language, 24,* 503-518.

Pierce, B. F., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition, 29(1),* 62-67. doi: 10.3758/BF03195741

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41,* 36-48. doi: 10.3758/s13421-012-0245-x

Pyc, M. A., Rawson, K. A., & Aschenbrenner, A. J. (2014). Metacognitive monitoring during criterion learning: when and why are judgments accurate? *Memory & Cognition, 42,* 886-897. doi: 10.3758/s13421-014-0403-4

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28 (1),* 69-80. doi: 10.1037//0278-7393.28.1.69

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28(6),* 1004-1010.

Reid, A. J., Morrison, G. R., & Bol, L. (2017). Knowing what you know: improving metacognition and calibration accuracy in digital text. *Education Technology Research Development, 65,* 29-45. doi: 10.1007/s11423-016-9454-5

Sanchez, C. A., & Jaeger, A. J. (2015). If it's hard to read, it changes how long you do it: Reading time as an explanation for perceptual fluency effects on judgment. *Psychonomic Bulletin & Review, 22*(1), 206-211.

Schunk, D. H., & Zimmerman, B. J. (1998). *Self-regulated learning: From teaching to self-reflective practice.* New York: Guilford.

Shiu, L., & Chen, Q. (2013). Self and external monitoring of reading comprehension. *Journal of Educational Psychology, 105(1),* 78-88. doi: 10.1037/a0029378

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28,* 129-160. doi: 10.1016/S0361-476X(02)00011-5

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95(1),* 66-73. doi: 10.1037/0022-0663.95.1.66

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(6),* 1267-1280. doi: 10.1037/0278-7393.31.6.1267

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331-362.

Trainin, G., & Swanson, H. L. (2005). Cognition, metacognition, and achievement of college students with learning disabilities. *Learning Disabilities Quarterly, 28,* 261- 272.

Undorf, M. Sollner, A., & Broder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition, 46*(4), 507-519.

Vanderberg, R., & Swanson. H. L. (2007). Which components of working memory are important in the writing process? *Reading & Writing, 20,* 721-752.

Verhaegen, P., & Salthouse, T. A. (1997). Meta-analysis of age-cognition relation in adulthood: estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin, 122*(3), 231-249.

Vieiro, P., & García-Madruga, J. A. (1997). An analysis of story comprehension through spoken and written summaries in school-age children. *Reading and Writing: An Interdisciplinary Journal, 9,* 41-53. doi: 10.1023/a:1007932429184.

Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(2),* 214-222.

Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P. J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied, 22(4),* 393-405.

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology*, *132*(4), 408-428.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*(4), 327-353. doi: 10.1016/S1041-6080(96)90022-9

Table 1

Summary of Average Response and Performance Scores

Condition	Prediction Magnitude	Multiple Choice Score	Bias Score	Relative Accuracy (Gamma)	Relative Accuracy (Stuart's Tau-C)
Written	4.99 (.12)	3.74 (.12)	1.25 (.14)	.31 (.09)	.25 (.06)
Oral	5.05 (.12)	3.85 (.12)	1.20 (.15)	.04 (.10)	.04 (.07)
Control	4.94 (.12)	3.82 (.12)	1.30 (.15)	.01 (.09)	.01 (.07)

Note. Parentheses indicate standard error.

Table 2

Summary Characteristics Comparisons

Condition	Word Count	LSA	Total Time (sec)	Summary Latency (sec)
Written	61.31 (1.98)	.59 (.01)	127.30 (4.19)	7.24 (0.38)
Oral	59.94 (2.66)	.54 (.01)	44.24 (1.49)	19.64 (1.57)

Note. Parentheses indicate standard error.

Table 3

Sequential Multiple Regression Analysis of Summary Characteristics and Prediction Magnitude

Predictor	Model 1	Model 2	Model 3
Total Time	.10 (.001)*	-.07 (.002)	-.06 (.002)
Summary Latency	-.09 (.005)	-.05 (.005)	-.06 (.005)
Word Count		.27 (.003)*	.27 (.003)*
LSA		.21 (.480)*	.21 (.480)*
Session		.02 (.270)	

Note. Represents standardized beta and standard error in parentheses. *$p < .05$.

Table 4

Sequential Multiple Regression Analysis of Summary Characteristics and Relative Accuracy

Predictor	Model 1	Model 2	Model 3
Total Time	.26 (.0004)*	.33 (.0004)*	.18 (.001)*
Summary Latency	.03 (.0020)	.03 (.0020)	.07 (.002)
Word Count		-.20 (.0010)*	- .11 (.001)*
LSA		.06 (.1500)	.05 (.160)
Session			-.19 (.080)*

Note. Represents standardized beta and standard error in parentheses. *$p < .05$.

Table 5

Multiple Regression Analysis with Cognitive Load and Prediction Magnitude

Predictors	β (*SE)*
Intrinsic CL	-.33 (.13)*
Extraneous CL	.02 (.10)
Germane CL	.12 (.15)

Note. *$p < .05$.

Table 6

Multiple Regression Analysis with Summary Characteristics and Multiple-Choice Accuracy

Predictors	β (*SE*)
Total Time	-.04 (.002)
Summary Latency	-.04 (.005)
Word Count	.29 (.003)*
LSA	.08 (.490)

Note. *$p < .05$.

Table 7

Correlation Matrix for Summary Characteristics

	1.	2.	3.
1. Word Count			
2. LSA Score	.34*		
3. Total Time	.47*	.26*	
4. Summary Latency	-.20*	-.14*	.27*

Note. *$p < .05$.

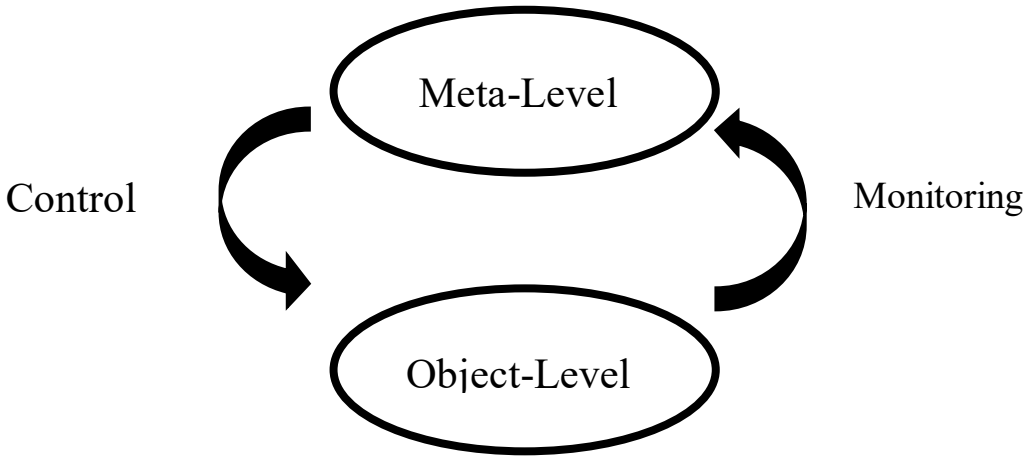

Figure 1. Model of cognitive monitoring and control proposed by Nelson & Narens (1990).

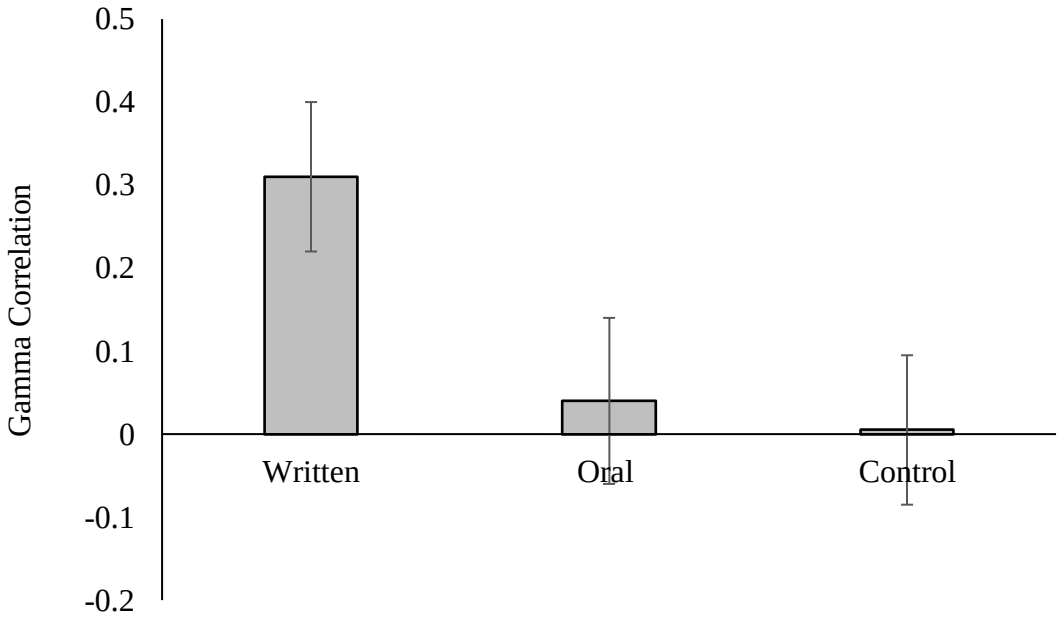

Figure 2. Mean gamma correlation as a function of condition. Error bars represent standard error.

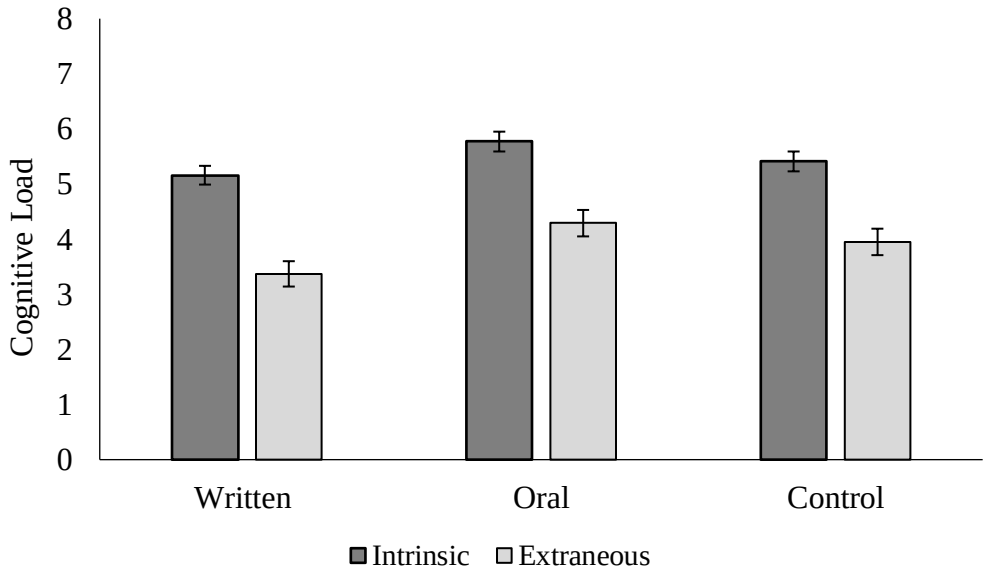

Figure 3. Mean intrinsic and extraneous cognitive load as a function of condition. Error bars represent standard error.

APPENDIX A: SAMPLE PASSAGE AND QUESTIONS

Television Newscasts

Relaying information and images instantly, television newscasts have allowed viewers to form their own opinions about various political events and political leaders. In many instances, television newscasts have even fostered active dissent from established government policies. It is no coincidence that, in the 1960's, the civil rights movement took hold in the United States with the advent of television, which was able to convey both factual information and such visceral elements as outrage and determination. Only when all of America could see, on the nightly newscasts, the civil disobedience occurring in places like Selma and Montgomery did the issue of civil rights become a national concern rather than a series of isolated local events. By relaying reports from cities involved to an entire nation of watchers, television showed viewers the scope of the discontent and informed the disenfranchised that they were not alone. The ability of television news to foster dissent has also been affected by increasingly widespread access to video cameras, so that the news presented on television now comes from the bottom up as well as from the top down. Across the world, dissidents have used video equipment to gather visual evidence of human rights abuses. Uncensored images and information have then been transmitted across otherwise closed borders by television newscasts. One professor of popular culture, Jack Nachbar, views the personal video camera as a "truth- telling device that can cut through lies." That claim presumes, though, that the television viewer can believe what he or she sees. But the motivation of the photographer must be taken into account, and the videotape that appears on television can, like still photography, be staged and even faked. When and if propagandists for some government utilize computer-generated effects, viewers will have more trouble believing what they see. However, even if seeing is not automatically believing, at least seeing is seeing--and in some repressive regimes, seeing is the fastest road to freedom

1. The passage is primarily concerned with ways in which
    a) television newscasts deliberately distort information
    b) television affects viewers by its presentation of news
    c) truth frustrates efforts by the media to constrain it
    d) viewers of television newscasts cannot sort out fact from fiction
    e) governments manage to control television newscasts

2. Which of the following, if true, would most strengthen the assertion about television and the American civil rights movement?
    a) Many filmed reports of civil disobedience were censored by television executives during the 1960s
    b) Recent studies have questioned the objectivity with which television newscasts presented reports of civil disobedience during the 1960s
    c) A biography of a major civil rights leader describes in detail the occasions on which the leader was featured in television newscasts in the 1960s
    d) A 1960s poll shows that those Americans who considered civil rights a national priority had seen television newscasts of civil disobedience
    e) Many of the reporting techniques used today originated in newscasts covering the 1960s civil rights movement

3. It can be inferred from the passage that television newscasts would be better at informing public opinion if
    a) newscasts presented only competing views and not one-sided views
    b) personal videos were banned from television newscasts
    c) technology was developed to detect when videos had been tampered with
    d) highly visceral information were not presented during television newscasts
    e) only factual information were presented during television newscasts

4. The author suggests a major reason why television newscasts are effective at influencing public opinion. Based on this argument, which medium below would be the most effective at influencing public opinion?
    a) daily newspapers
    b) radio broadcasts
    c) classroom instruction
    d) grassroots movements based on word of mouth
    e) witnessing newsworthy events first hand

5. According to the passage, television coverage of the civil rights movement did all of the following EXCEPT
    a) inform dissenters that they were not alone
    b) convey factual information
    c) present emotional elements such as anger
    d) portray the scope of the dissent
    e) express opinions of the political leaders

6. Jack Nachbar, who is quoted in the passage, is
    a) a popular culture professor
    b) a government propagandist
    c) a reporter for a professional news agency
    d) a civil rights activist
    e) a prominent political figure

7. The author explicitly states that the believability of television news may be compromised by
    a) effects produced by computers
    b) videos from personal cameras
    c) photographers for professional news agencies
    d) established government policies
    e) reports that are transmitted across closed borders

8. The passage states that when nightly newscasts portrayed civil dissent in the 1960s,
    a) it incited dissent in places like Selma and Montgomery
    b) it created a national concern for civil rights
    c) it started a series of isolated local events
    d) viewers formed opinions about political leaders
    e) interest in personal video cameras increased

APPENDIX B

Perceived Cognitive Load Questionnaire

1. For this task, many things needed to be kept in mind simultaneously.

Absolutely Wrong						Absolutely Right
1	2	3	4	5	6	7

2. This task was very complex.

Absolutely Wrong						Absolutely Right
1	2	3	4	5	6	7

3. I made an effort, not only to understand several details, but to understand the overall context.

Absolutely Wrong						Absolutely Right
1	2	3	4	5	6	7

4. My point while dealing with the task was to understand everything correctly.

Absolutely Wrong						Absolutely Right
1	2	3	4	5	6	7

5. During this task, it was exhausting to find the important information.

Absolutely Wrong						Absolutely Right
1	2	3	4	5	6	7

6. The design of this task was very inconvenient for learning.

Absolutely Wrong						Absolutely Right
1	2	3	4	5	6	7

7. During this task, it was difficult to recognize and link the crucial information.

Absolutely Wrong						Absolutely Right
1	2	3	4	5	6	7