Metacognition and Learning (2022) 17:117–138
https://doi.org/10.1007/s11409-021-09277-5

The influence of summary modality on metacomprehension accuracy

Erin M. Madison[1] · Erika K. Fulton[1]

Received: 13 July 2020 / Accepted: 20 August 2021 / Published online: 1 September 2021

Abstract

Metacomprehension refers to the ability to monitor and control reading comprehension. It is important for individuals to be accurate in their judgments of comprehension, as this can affect academic performance. One type of accuracy, relative accuracy, tends to be low, meaning individuals cannot adequately differentiate well-known from less well-known information. Fortunately, past research has shown that relative accuracy increases with delayed summarization. The literature has only assessed written summaries as an intervention, but oral summaries tend to be faster and easier and therefore may be a better study tool. Individuals use cues to make judgments, which may differ between modalities. This study investigated whether modality impacts relative accuracy and if differences in cue use might explain these effects. We found that written summaries benefitted relative accuracy compared to a control group, with relative accuracy greater than chance. In contrast, oral summarizers only marginally differed from chance accuracy and did not differ from the control group. An analysis of summary characteristics suggests that participants use multiple cues in order to make judgments. We conclude that spoken summaries are likely better than not summarizing at all, but the written modality is the better summary technique to increase relative accuracy. By increasing relative accuracy, delayed written summaries may increase effectiveness of studying, thereby maximizing a student's academic potential.

Keywords Metacomprehension · Summary modality · Relative accuracy · Situation model hypothesis · Accessibility hypothesis

Metacomprehension describes thoughts and ideas about reading comprehension and includes monitoring and control of the comprehension process (Dunlosky & Lipko, 2007; Dunlosky et al., 2005; Maki & Berry, 1984). A common metacognitive framework demonstrates that a person can monitor or evaluate their basic cognitions then, based on this evaluation, control the basic cognitive task (Nelson & Narens, 1990). Using metacomprehension as an example, a person may read a paragraph and realize that they lost focus and do not comprehend a text (monitoring), then reread that section of the text (controlling).

✉ Erin M. Madison
madierin@isu.edu

[1] Department of Psychology, Idaho State University, 921 S 8th Ave, Stop 8021, Pocatello, ID 83209, USA

Metacomprehension benefits all readers, but is particularly important for students tested on expository readings (Dunlosky & Lipko, 2007) and those who read for work because a misperception of comprehension could result in poor performance on exams or projects.

Efficient readers tend to adjust reading behaviors to reach learning goals, which requires them to have accurate judgments of their comprehension (Dunlosky et al., 2005; Efklides, 2014; Schunk & Zimmerman, 1998). During this process, called self-regulated learning (SRL), an individual will plan and set study goals and then monitor their learning to assess whether they have met their goals. As monitoring learning is an integral part of SRL, students should be accurate in their judgments for successful studying (Lee et al., 2010). However, metacognitive monitoring accuracy tends to be poor, leading many students to struggle with SRL (Lee et al., 2010). Specifically, relative accuracy, or the ability to distinguish between what is learned and not learned, helps learners evaluate what to spend time studying, a crucial part of SRL (Dunlosky et al., 2005; Wiley et al., 2016). For example, if a student is studying for an exam, they might determine that chapter two is learned well and does not need further studying, but that chapter four is not well-learned and they must study it further in order to be successful on an exam. If that judgment is correct, then the student demonstrates high relative accuracy. To measure relative accuracy in the context of metacomprehension, a person is typically given multiple texts to read, and then asked to predict their future performance on each text on some scale predetermined by the experimenter. This prediction magnitude is compared to actual test performance using a gamma correlation. When relative accuracy (gamma correlation) is high, students tend to study efficiently by, for example, allocating more time for passages that are challenging but within their reach (Kornell & Metcalfe, 2006; Metcalfe, 2009).

Relative accuracy is typically low but implementing metacognitive strategies may lead to significant improvements. Without any strategy, students' mean relative accuracy is a gamma correlation of 0.27 (Dunlosky & Lipko, 2007). Although this correlation is significantly greater than zero, there is still much room for improvement. Interventions can greatly improve relative accuracy, especially when there is a generative component; for example, generating keywords after a short delay (de Bruin et al., 2011; Thiede et al., 2003, 2005), mental self-explanation of the text during reading (Griffin et al., 2008), and delayed written summaries (Anderson & Thiede, 2008; Fukaya, 2013; Thiede & Anderson, 2003) all seem to increase relative accuracy. For each of these interventions, the participants' relative accuracy leaped to a gamma correlation of approximately 0.6! It is generally believed that these interventions increase accuracy by bringing attention to the situation model, which is a gist-based mental representation of the text (Kintsch, 1998) that helps participants better monitor their comprehension (Anderson & Thiede, 2008). The purpose of the present study was to further explore the benefit of written summarization to relative accuracy by comparing it to oral summarization. We reasoned that summary modality may affect the cue basis for comprehension judgments. For example, oral summaries typically have more inference-based ideas (Anderson & Thiede, 2008; Kellogg, 2007), which might lead the situation model to be especially salient in the oral condition, benefitting relative accuracy. The use of judgment cues and how they may differ between summary modalities is discussed next.

Cue utilization

Individuals use their experiences with the material to predict their comprehension; thus, inequitable summarizing experiences between speakers and writers may alter the cues that individuals might use to make judgments. The cue-utilization hypothesis states that people cannot directly judge the accuracy of their cognitions, and therefore must use cues, or heuristics, to estimate the information they know (Koriat, 1997). For example, a person might estimate they will score well on a comprehension exam because they enjoy the topic, or because they read the information multiple times. Certain cues allow for better predictions, particularly those that represent a deeper understanding of the material (Griffin et al., 2008) and are referred to as having higher cue validity. Using the earlier example, if a person thinks they will pass a comprehension test because they enjoy the topic but they do not pass, then topic enjoyment is not a valid judgment cue. The cue utilization hypothesis provides a framework to describe why oral and written summaries may contribute to potential differences in prediction accuracy. The accuracy of oral summaries may depend on which cues the participants are using in their judgment; by comparing two similar but distinct metacognitive strategies, we can start to uncover when students use certain cues, and whether the use of cues is influenced by summary modality. For example, if oral summaries are longer and associated with less accurate metacomprehension judgments compared to written summaries, we might conclude that oral summarizers were basing their judgments on less valid cues, such as summary length. There are two cues already well described in the literature that may differ between oral and written summaries: the accessibility of information at retrieval and the situation model. Theoretical and empirical accounts for each will be described in turn.

The accessibility hypothesis explains how individuals use the amount of information they recall from a task as a cue to make prediction judgments. While making a judgment, recalling a large quantity of information tends to increase an individual's confidence in their knowledge, and therefore increases prediction magnitude (Koriat, 1993, 1997). Importantly, higher prediction magnitude does not equate to higher metacognitive accuracy (Baker & Dunlosky, 2006; Koriat, 1993; Maki et al., 2009; Morris, 1990). Accessibility of information can be a valid cue if the retrieved information is both correct and relevant, but incorrect, repetitive, or irrelevant information may erroneously inflate prediction magnitude. The accessibility hypothesis predicts that a high word count or number of total ideas during a summary will lead to a high prediction magnitude (Maki et al., 2009) but not necessarily high accuracy.

The situation model, or a gist-based mental representation of the text (Kintsch, 1998), is another cue used to judge comprehension (Anderson & Thiede, 2008; Fukaya, 2013; Thiede & Anderson, 2003; Thiede et al., 2005). The situation model connects information within the text to previous knowledge about the topic. This deep level of processing is resistant to forgetting and therefore considered a valid cue on which to base judgments of comprehension (Kintsch et al., 1990; Thiede et al., 2005). The situation model may be related to summary length, but summary length includes gist-based ideas and details as well as distorted or irrelevant information (Anderson & Thiede, 2008). Therefore, although the accessibility of information and the situation model are related, they are distinct concepts, and not equally valid judgment cues. Although evidence suggests that delayed summaries increase salience of the situation model, which improves relative accuracy (Anderson & Thiede, 2008; Thiede & Anderson, 2003), the situation model is not always the most salient cue. For instance, students are more likely to use other cues like the accessibility of information, or surface level cues, like the readability of the text. This is especially true when students make judgments without using a metacognitive intervention such as delayed summaries, as these interventions may work in part by bringing attention to the situation model (Anderson & Thiede, 2008; Thiede et al., 2010).

Summarization modality and comprehension

The ability to summarize a text depends on one's ability to comprehend it (Alterman, 1991), which is likely why summarizing increases metacomprehension accuracy. There are generally two types of summaries that students construct: one more surface level, and one that is deeper, connecting it to past information, representative of the situation model. College students tend to draw from their prior knowledge when summarizing, therefore tapping the situation model (Leon et al., 2006). When creating a summary, individuals typically form a group of key ideas that represent the main ideas of the text (León & Escudero, 2015). Summarizing is more complicated than text comprehension, as this process includes identifying and generalizing the most important ideas, and conveying them in a coherent and concise manner (León & Escudero, 2015). Being able to construct a good summary is an active process rather than a passive one. Although this core process is similar in oral and written modalities, there are some processes that differ between the two.

Written and oral summarizing are similar in that both involve condensing texts to their key components in a cohesive manner, and result in similar output quality (Hidi & Hildyard, 1983; Scardamalia et al., 1982). However, studies have shown that several summary characteristics can differ between modalities. For example, oral production (recall, summaries, narratives) tends to have higher idea units and word count, yet take less time (Hidi & Hildyard, 1983; Kellogg, 2007; Viero & García-Madruga, 1997). Also, they have been shown to include more gist-based ideas, whereas written summaries tend to be more verbatim (Viero & García-Madruga, 1997). Written output tends to be slightly more cohesive, which is believed to be a result of being able to pause during the writing process and assess what is already written (Hidi & Hildyard, 1983). Additionally, there are potential differences in the demand placed on working memory during oral and written summarizing. First, while speaking has a small motor component, writing has a much larger motor component, and therefore must be considered in the working memory process (Kellogg, 2007). Furthermore, although both modalities include the macrostructures of language (e.g. generating speech), spoken summaries should tax working memory less, as some microstructures, such as spelling, are less prominent during speaking (Vanderber & Swanson, 2007). On the other hand, oral summarizers must rely on their memory of what they have already said. Because written summarizers can reread what they wrote, they may have less strain on their working memories.

In the metacomprehension literature, oral and written summaries seem to differ in ways that affect the situation model and the accessibility of information. The oral modality tends to produce a greater number of inference or gist-based ideas (Kellogg, 2007; Vieiro & García-Madruga, 1997), an indicator of a strong situation model (Anderson & Thiede, 2008), which may increase metacomprehension accuracy. Unfortunately, oral production may also contain higher levels of distortions and more idea units in general (Hidi & Hildyard; Kellogg, 2007), and therefore have a higher word count. When word count is driven in part by distortions, the accessibility hypothesis would predict that individuals inflate prediction judgments, but have potentially lower accuracy because summary length is not always a valid cue (Dunlosky et al., 2005b; Koriat, 1993). The accessibility of information is only a valid cue when the amount of information recalled is both correct and relevant. Another issue with oral summarization revolves around retrieval fluency as a judgment cue. Retrieval fluency, or the ease with which a person retrieves information, is closely related to response time (Bjork et al., 2013), and is associated with higher prediction magnitude but not necessarily comprehension (Benjamin et al., 1998). Because oral production takes much less time than written production (Kellogg, 2007), it could lead to lower metacomprehension accuracy if retrieval fluency is used as a cue.

Although students may prefer the oral modality because it is faster and feels easier (Kellogg, 2007; McPhee et al., 2014), it may not afford the same metacomprehension accuracy as written summarization. If oral summaries lead to worse metacognitive outcomes, we can still gain vital information about *why* the outcome was worse. During oral summarization, participants may use heuristics that sometimes work but are not valid cues in this scenario (such as length as a cue), due to oral summaries generally having more distortions. Alternatively, there might be explanations outside of cue use that contribute to group differences. For example, unlike written summaries, oral summaries cannot be reread. If participants are unable to review their summaries, they lose an additional opportunity to evaluate their knowledge and improve their metacomprehension accuracy.

Perceived cognitive load differences between summarizing modality

Summary modality may also affect perceived cognitive load, which may be another cue affecting judgments and therefore judgment accuracy. Kellogg (2007) proposed that speaking and writing differ in terms of working memory demand, which can influence cognitive load, defined as the amount of cognitive resources required to complete a task (Chandler & Sweller, 1991). Bourdin and Fayol (1994) proposed that writing may tax working memory more than speaking. This may primarily be because writing requires spelling, whereas speaking, obviously, does not (Vanderber & Swanson, 2007). Furthermore, the writing process is slower than speaking, so text representations may remain in working memory for longer, using more resources (Kellogg, 2007). Currently there is evidence that writing taxes working memory more than speaking for both children (Bourdin & Fayol, 1994) and adults (Grabowski, 2010). There is also evidence that participants prefer to speak because writing is more effortful (McPhee et al., 2014). For these reasons, it is possible that perceived cognitive load may be higher when summarizing in writing, and act as another cue that could influence comprehension judgments and group differences.

The way in which perceived cognitive load may affect metacomprehension accuracy is unclear. For example, when a task is experienced as more difficult, students are less confident in their judgments (Maki et al., 2005, although see Moore et al., 2005 for counter evidence), which could *decrease overconfidence*, making judgments more accurate. However, one study found that written summaries increased cognitive load in comparison to a control group without increasing relative accuracy (Reid et al., 2017), but there was no oral summarization condition, so no modality comparisons were made. Even if perceived cognitive load does not act as a judgment cue, or is not a major explanatory one, the difference in perceived cognitive load between conditions is still worth measuring. If metacomprehension monitoring accuracy is equivalent between summarizing orally and in writing, but one modality leads to lower perceived cognitive load, then summarizing in that modality would have an obvious study benefit: students should summarize in the “easier” modality if the harder one has no metacomprehension benefits. Another reason it is difficult to predict the effect of perceived cognitive load on metacomprehension judgment accuracy is that cognitive load is not a unitary construct. Paas and colleagues (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 motivation 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 et al., 2003). Summarizing modality could affect one or more of these three components, so all three were measured.

The present study

The delayed summary technique has been found to increase metacomprehension accuracy (Anderson & Thiede, 2008), but is summarizing in one modality superior to the other? Both written (Anderson & Thiede, 2008; Maki et al., 2009; Reid et al., 2017; Thiede & Anderson, 2003), and oral summaries (Baker & Dunlosky, 2006; Fukaya, 2013; Fulton, 2021) have been used in past metacomprehension research but, to our knowledge, they have never been compared in the context of metacomprehension monitoring accuracy. The present study was conducted to address this gap in the research literature. With this study, we can learn (a) whether oral summaries seem to improve relative accuracy compared to written summaries and (b) whether summary characteristics drive potential differences in relative accuracy. Both the situation model and the availability of information were expected to differ between spoken and written summaries, which could influence which cues were available or salient and thus the prediction magnitude and the accuracy of the predictions. Comparing the accuracy for three conditions (written, oral, and no summary), as well as their summary characteristics, informs metacognitive theory and could elucidate which summarization type might be most useful for students while studying. We included a control condition to assure that the delayed summarization manipulation reliably improves metacognitive accuracy. If neither summary condition increases accuracy more than the control condition, then delayed summarization loses its practical and possibly theoretical implications.

It is important to note that the cues measured in the study do not encompass all cues that a person can use, and that a person can use multiple cues to make their judgments (Morris, 1990; Undorf et al., 2018). We focused on cues that past literature suggested might change with summary modality, but we acknowledge that other cues, such as familiarity or interest in the topic, may play an important role in prediction magnitude (Koriat, 1997; Thiede et al., 2010). We also note that because multiple cues can be used simultaneously, we are not pitting the accessibility hypothesis directly against the situation model hypothesis, per se. We assessed relative reliance on each cue first by correlating each participants' cues with their predictions; the cue(s) with the strongest relationship to prediction magnitude were considered the most salient to the participants. Next, we assessed which cue has the highest validity by correlating cues with multiple-choice accuracy. We believed that summary modality would likely alter the cues in the summaries (e.g. amount of information recalled), and that the salience and validity of each cue might differ between modalities.

> Hypothesis 1: We predicted that the three groups would a.) differ in their prediction magnitude, b) not differ in their comprehension, but c) differ in their relative accuracy.

Written summaries generally increase metacomprehension accuracy, likely due to the increased attention to the situation model (Anderson & Thiede, 2008; Thiede & Anderson, 2003). Because oral summaries tend to have more valid cues (gist-based ideas) and more invalid cues (more total words, faster summary time; Kellogg, 2007; Vieiro & García-Madruga, 1997), it was unclear how oral summarizing would impact metacomprehension accuracy. Thus, hypotheses did not specify whether oral summaries would be associated with more or less accurate judgments compared to written summaries, just that there would be a difference.

> Hypothesis 2: Each summary characteristic was expected to differ between conditions.

We were interested in length (word count), the situation model (latent semantic analysis), and retrieval fluency (latency to begin summarizing and total time). Past research suggests higher word count and latent semantic analysis scores in the oral condition (Kellogg, 2007; Vieiro & García-Madruga, 1997), and higher summary time and latency to begin summarizing in the written condition (Kellogg, 2007). Although these predictions are directional, we planned two-tailed tests because we felt the presence of any differences was more important than the direction of that difference.

> Hypothesis 3: (a) We expected summary characteristics (word count, latent semantic analysis, latency to begin summarizing, and total summary time) to relate to prediction magnitude and (b) comprehension. (c) Further, it was hypothesized that these relationships would differ between cues, and by condition.

In order to assess cue use and cue validity, intra-individual gamma correlations were calculated for each participant, which correlated summary characteristics and prediction magnitude, as well as summary characteristics and multiple-choice scores (Anderson & Thiede, 2008; Maki et al., 2009). Higher correlations represent higher cue use (when correlated with prediction) and cue validity (when correlated with comprehension). Differences between modalities can suggest which cues are used to a greater extent, or are more valid, in one condition compared to another.

> Hypothesis 4: We expected perceived cognitive load to differ between groups.

Because working memory will likely increase the most in the written condition (Kellogg, 2007), it was predicted that the control group would exhibit the lowest levels of perceived cognitive load, and the written summary group would exhibit the highest levels of perceived cognitive load. However, we planned two-tailed tests because we felt the presence of any differences was more important than the direction of that difference.

This hypothesis was exploratory, as our main focus was on summary modality and differences in cue use. Although perceived mental effort can certainly act as a cue, the timing of the cognitive load measure makes it unclear whether this measure reflects perceptions of load during the summarization process, per se. This was an intentional design choice to prevent artificially increasing the salience of cognitive load before making prediction judgments. We believe our cognitive load measure has merit in the current study, but results should be viewed as preliminary evidence.

Methods

Participants

To estimate the target sample size, we used an effect size of $d = 0.45$, derived from a study which assessed modality differences in recall ability (Putnam & Roediger, 2013); the current study required 95 participants at 0.80 power (Bausell & Li, 2002). Overall, 116 individuals over the age of 18 participated in this study, recruited from the SONA system at Idaho State University. Students were compensated with course credits. Ten people who did not speak English as their first language were excluded. Three additional participants were excluded for not following directions, and one was excluded for making uniform predictions, so a gamma correlation could not be calculated. It should be noted that there was a problem in the original audio recordings. Unfortunately, the oral summaries could not be transcribed accurately, so the oral condition was completely replaced. Both the original and new samples were similar in their demographics and performance. There were 102 participants included in the final analysis. There were 35 participants in the written condition, 34 participants in the oral condition, and 33 in the control condition.

The sample was primarily white (91.2%) and female (70.6%). Of our population, 11.8% identified as Hispanic. The mean age of the sample was 21.76 ($SD=5.20$) years, and most participants were early in their college career, with 68.6% in their first year and second year.

Design

This experiment used a one-factor, between-subjects design. Participants were randomly assigned to one of three groups, an oral summary group, a written summary group, or a no summary control group.

Materials

Eprime 3.0 (Psychology Software Tools; www.pstnet.com) was used to display each of the texts in this study. We used six texts that have been used in similar metacomprehension experiments (Fulton, 2021; 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 ($M=11.6$). Each text was between 337 and 398 words. 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). There were eight multiple choice questions for each text. Four could be answered from information that was explicitly stated in the text and four required making inferences from the text. Inference-based questions best assess comprehension of the situational model.

Additionally, Eprime recorded prediction and postdiction judgments, the multiple-choice answers, latency to begin summarizing and summary time for the oral and written conditions, and the typed summaries for the written condition. The oral condition summaries were recorded with Audacity (http://audacity.sourceforge.net/) and transcribed for analysis. Qualtrics was used to administer the demographic questionnaire, as well as an exploratory cognitive load survey. The cognitive load survey has been validated and shows strong reliability (Cronbach's $\alpha=0.81$, Klepsch et al., 2017). To score the cognitive load survey, participant answers were averaged across a 7-point Likert-type scale for each subscale (intrinsic, extraneous, germane).

Procedure

Participants first read each of the six texts in a random order. The texts were displayed so that only one sentence appeared at a time to control for rereading, as rereading can lead to an increase in accuracy (Rawson et al., 2000). The reading task had no time limit, but the

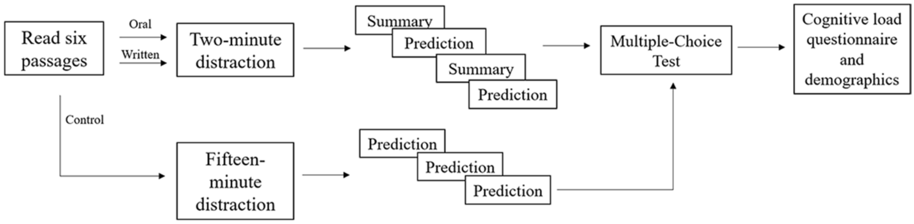

Fig. 1 Summary of methods

time that it took to read each text was recorded with Eprime. A two-minute word search was presented to the oral and written conditions after the texts to assure that they were summarizing at a delay and not rehearsing the information that they read. After the two-minute delay, participants in the oral and written conditions were asked to summarize one text at a time. The title of each text appeared as a prompt for them to begin summarizing. After each summary was completed, a prediction question was presented, 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. Participants were told to read carefully for a future test, but they were not informed of the nature of the test.

The control condition did not summarize the texts. The control participants read the texts as in the experimental conditions but were a given a 15-minute word search as an easy distraction task in place of generating summaries. The distraction task prevents the individuals from rehearsing information from the passages, which could influence their comprehension and metacomprehension. After spending 15 min on the word search, 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, as in the two experimental conditions.

After completing predictions, all participants completed the multiple-choice comprehension test. The test was composed of eight questions for each text. Once they finished each set of questions, participants took the cognitive load questionnaire. Finally, they completed a demographic survey and were debriefed about the study. See Fig. 1 for a procedural summary.

Data analysis

The summaries were measured on four dimensions: length, situation model, latency, and total time. Length was measured by a word count. The situation model was measured using a technique called latent semantic analysis (LSA; http://lsa.colorado.edu/; Landauer, 1998). Latency to begin summarizing and total time were measured using Eprime. To measure latency, participants were instructed to remain on a screen until they were ready to summarize. The time that they spent on this slide was considered latency to begin summarizing. Total time was measured by the amount of time it takes from the presentation of the summarizing prompt to the time it takes to finish summarizing and moving to the next screen. Total time did not include latency to begin summarizing.

LSA, the current measure for the situation model, measures how closely a summary relates to an ideal target summary, using a cosine that measures the semantic relatedness of the two texts. The cosine is comparable to a correlation coefficient. LSA does not assess synonyms; rather, it compares how words are used in similar contexts. Because of this feature, LSA is able to measure the gist of the text and can therefore potentially measure the situation model (Landauer et al., 2007). LSA has been shown to measure both comprehension and cohesion (Landauer & Dumais, 1997). The target summaries were adapted from the grading rubric in Fulton (2021), which described the main ideas and important details of each text, as agreed upon by two judges. LSA has been used in metacognitive research in the past (Maki et al., 2009; Thiede & Anderson, 2003) and found to be comparable to a trained scorer (Landauer, 1998).

Relative accuracy was calculated for each participant using a gamma-correlation between participant prediction magnitude and multiple-choice performance. A one-way ANOVA was used to compare the three groups on prediction magnitude, multiple-choice accuracy, and relative accuracy. A Tukey test was run after each significant ANOVA to assess which groups differed from each other.

The summary characteristics in the oral and written conditions were compared using t-tests, and a Bonferroni correction was used to account for increased error rate. After differences in summary characteristics were established, summary characteristics were correlated to prediction magnitude and multiple-choice scores to establish cue use and cue validity. To assess whether individuals were using our measured summary characteristics to make predictions of their performance, intra-individual gamma correlations were calculated between each summary characteristic and prediction magnitude. To measure whether these cues were valid, intra-individual gamma correlations between each summary characteristic and multiple-choice scores were calculated. There were eight gamma correlations calculated for each individual: four correlations between summary cues and predictions, and four between summary cues and multiple-choice scores. Finally, mean gamma correlations were compared across groups using a 2 (oral vs. written) x 4 (summary characteristic) repeated measures ANOVA. Type of summary characteristic was considered a repeated measure variable, as each person had four measures of summary characteristics, and these were compared between groups. A separate 2 (oral vs. written) x 4 (summary characteristic) repeated measures ANOVA was conducted for gammas relating summary characteristics to prediction magnitude and to multiple-choice scores.

Results

Prediction magnitude and multiple-choice performance

No group differences were found for prediction magnitude [$F(2, 99) = 0.04$, $p = .96$, $\eta^2 = 0.00$] nor for multiple-choice score [$F(2, 99) = 0.68$, $p = .51$, $\eta^2 = 0.01$; see Table 1 for means]. 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 mean prediction magnitude or comprehension.

Table 1 Mean prediction magnitude and multiple-choice performance by condition

	Written	Oral	Control
Prediction Magnitude	5.03(0.12)	5.00(0.12)	4.94(0.12)
Multiple-Choice Score	3.73(0.12)	4.00(0.12)	3.82(0.13)

Both prediction magnitude and multiple-choice scores range from 0–8. Standard error in parentheses

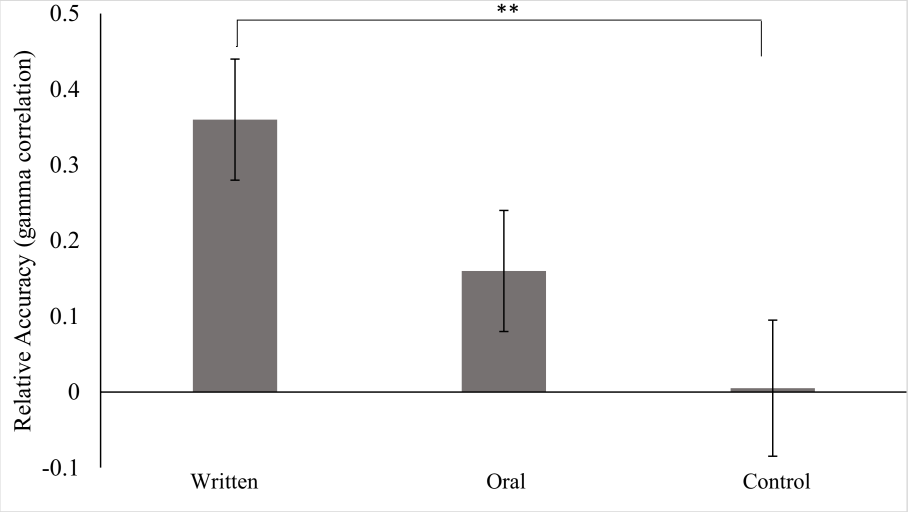

Fig. 2 Mean relative accuracy by condition. Note: Error bars represent standard error. $^{**}p < .05$

Relative accuracy

The average gamma correlation across all conditions was small, at 0.18 ($SE = 0.05$), but significantly different from zero [$t(101) = 3.55$ $p < .01$]. This suggests participants were, on average, above chance at distinguishing on which texts they would score well. An ANOVA showed differences in average gamma correlations between conditions [$F(2, 99) = 4.46$, $p = .01$, η^2 =0.08], supporting the hypothesis that relative accuracy would differ between groups. A Tukey test revealed that the written condition had the highest average gamma correlation (Fig. 2), which differed significantly from the control condition, which had the lowest relative accuracy of the three groups (95% CI [0.07, 0.63]; $p = .01$). The oral condition had an intermediate relative accuracy; it did not differ significantly from either the written condition (95% CI [-0.08, 0.47]; $p = .22$), or the control condition (95% CI [-0.47, 0.08]; $p = .40$). However, the written condition was the only condition to have a gamma correlation significantly different from zero ($t(34) = 4.53$, $p < .01$). The average gamma correlation for the oral condition was marginally different than zero [$t(33) = 2.01$, $p = .052$], but the control condition failed to differ [$t(32) = 0.05$, $p = .95$]. Thus, we are confident that delayed written summaries effectively increased relative accuracy, as this is the only condition to really differ from both zero and from the control condition.

Table 2 Differences in averages of summary characteristics by modality

	Written	Oral	t(67)	*p*-value
Word Count	60.92 (4.07)	73.66 (6.64)	-1.64	0.10
LSA Score	0.59 (0.01)	0.60 (0.01)	-0.63	0.50
Summary Time (sec)	127.60 (8.66)	41.65 (2.95)	9.28	< 0.01
Summary Latency (sec)	7.27 (0.52)	16.70 (2.12)	-4.38	< 0.01

Standard error in parentheses

Summary characteristics between modalities

Some, but not all summary characteristics differed between the written and oral conditions (Table 2). The written summaries took longer to complete on average [t(67) = 9.28, p < .01, d = 2.25], which we anticipated. The written condition was also quicker to begin summarizing [t(67) = -4.38, p < .01, d = 1.05]. The oral and written summaries did not differ in LSA score [(t(67) = -0.68, p = .50, d = 0.17] or word count [t(67) = -1.64, p = .10, d = 0.39], contrary to the hypothesis. All tests were Bonferroni corrected, with p = .0125. Overall, the groups did not differ in their situation models or word count, but oral summarizers took less time to summarize, and more time to begin summarizing.

Relation between summary characteristics, prediction magnitude, and MC accuracy

As hypothesized, summary characteristics significantly related to prediction magnitude, with each gamma correlation between summary characteristics and prediction magnitude differing significantly from zero (Table 3), indicating that each summary characteristic measured was related to prediction magnitude. Next, in order to measure whether some characteristics were more related to prediction magnitude, a 2×4 repeated measures ANOVA was conducted. There was found to be a significant main effect of summary cues [F(3, 65) = 15.17, p < .001], suggesting that cues did not equally influence prediction magnitude. Specifically, latency to begin summarizing was significantly lower than each other cue, with each other cue found to be different from latency at the p < .001 level. Additionally, word count was marginally larger than LSA (p = .051) and total time (p = .094). Although approaching traditional significance levels, group differences in cue use were not significant [F(1, 67) = 2.91 p = .09], and there were no significant interaction terms [F(3, 65) = 0.33 p = .80].

In regards to comprehension, there were significant mean differences in the relationship between summary characteristics and comprehension scores [F(3, 65) = 7.00, p < .001; Table 3]. Again, latency to begin summarizing was significantly lower than each other cue, with each other cue found to be different at the p < .001 level. No other cues were different from one another. The interaction term (cues by condition) was found to be significant [F(3, 65) = 4.16, p < .009]. For the interaction term, total time differs in cue validity between written and oral conditions (p = .045), with time only as a valid cue in the *oral* condition. Finally, there were no main effects for condition [F(1, 67) = 0.26, p = .98].

Table 3 Gamma correlations between summary characteristics and predictions, as well as between summary characteristics and multiple-choice accuracy

	Gammas between characteristics and predictions				Gammas between characteristics and multiple-choice scores			
	LSA	Word Count	Total time	Latency to begin	LSA	Word Count	Total time	Latency to begin
Written	0.35 (0.07)	0.48 (0.07)	0.36 (0.07)	−0.13 (0.08)	0.13 (0.07)	0.15 (0.07)	0.01 (0.07)	−0.17 (0.07)
Oral	0.24 (0.07)	0.30 (0.07)	0.26 (0.07)	−0.16 (0.08)	0.02 (0.07)	0.09 (0.07)	0.20 (0.07)	−0.19 (0.07)

Standard error in parentheses

Table 4 Group means for perceived cognitive load

	Intrinsic	Extraneous	Germane
Written	5.16 (0.22)	3.33 (0.23)	6.21 (0.13)
Oral	5.60 (0.17)	4.08 (0.23)	5.94 (0.14)
Control	5.41 (0.13)	3.91 (0.24)	5.88 (0.18)

Standard error in parentheses

Cognitive load

There were three subscales of cognitive load, and each were compared individually using a one-way ANOVA. First, intrinsic load was not found to be different between groups [$F(2, 101) = 1.55$, $p = .21$ (see Table 4 for group means)]. Extraneous load was marginally different [$F(2, 101) = 2.77$, $p = .067$], with a Tukey test revealing a marginal difference between the oral condition and written condition ($p = .066$) and no other differences. Finally, germane load did not differ between conditions [$F(2, 101) = 1.40$, $p = .25$]. Therefore, extraneous load, or the cognitive load dependent on environment, may be higher in the oral modality compared to the written modality.

Discussion

Our findings provide the first evidence of an effect of summary modality on metacomprehension relative accuracy and evidence of multiple cue use in a metacomprehension context. The results suggest that written summaries, but not oral summaries, benefit relative accuracy in metacomprehension, as the written summary condition was the only group whose relative accuracy was greater than chance and differed from the control group. Explanations for this effect remain unclear, but we believe the findings have implications for metacognitive theory and SRL, and we discuss possible explanations and implications below. We also discuss the novel evidence for multiple cue use in metacomprehension predictions and the extent to which they are valid cues.

Summarizing in writing appears to benefit relative accuracy in metacomprehension predictions, as the written summary group was the only group whose judgment accuracy was significantly different than the control group and significantly different from zero. We are aware that no strong statement can be made about oral summarizing as that group was not significantly different from either the written or the control condition. However, we note that Griffin and colleagues (2019) argued that gamma correlations, which are assumed to be an ordinal variable, experience a reduction in variation and thus statistical power, which can inflate type II error rates. For this reason, it is possible that a replication with greater statistical power could show that oral summarizing is significantly better than the control and/or worse than written summarizing. Regardless, we believe our results provide evidence that summarizing, particularly in writing, may improve SRL through its impact on metacomprehension. In particular, students who *write* summaries of texts they read may be better judges of which of those texts are more or less well understood. This can allow them to make informed choices about continued study, such as which texts need to be restudied or need the most time allocated to them during either study or test (Metcalfe & Finn, 2008). We had planned to use analyses of cue use to help interpret why summarizing and/or a particular summary modality might afford greater advantages for metacomprehension judgment accuracy, but the complex and nuanced nature of the findings prevents strong conclusions about mechanism. Nonetheless, we provide some possible explanations below.

We originally hypothesized that group differences in cue use could help explain why one summary modality might lead to better metacomprehension relative accuracy. As such, the written condition may have outperformed the oral condition because they were better able to use the cues (i.e., summary characteristics we measured) at their disposal. However, the group difference in the relationship between prediction judgments and cues was only marginally significant ($p = .09$), with no significant interaction between condition and summary characteristic. It is interesting, though, that all gamma correlations between cues and predictions (barring latency) were higher in magnitude for the written condition than the oral condition, particularly for word count ($g = 0.48$ versus $g = 0.31$). The gamma between word count and predictions was also marginally larger than between predictions and two other summary characteristics, LSA and total time; importantly, this was only the case in the written condition. Again, although we must very cautiously interpret null effects, it could mean that word count is a stronger judgment cue than the others, especially when considering the conservative nature of gamma correlations (Griffin et al., 2019). If so, one way this might have occurred is that the visual nature of written summaries could have increased the salience of word count as a cue for participants in the written condition, making it easier to monitor word count when one can see how much space it takes up on the page, a visual that is absent while speaking. Higher salience of word count might lead those in the written condition to incorporate word count into their judgments to a greater degree, leading the written condition to be more accurate. Nonetheless, we fully acknowledge that this reasoning is speculative given the null/marginally significant differences in cue use, but we do believe that this possibility is worth exploring.

In addition to showing that summary modality can affect metacomprehension relative accuracy, our study provides the first experimental evidence, to our knowledge, that people use multiple cues when making metacomprehension judgments (see Undorf et al., 2018 for evidence in metamemory). Each of the cues measured (word count, LSA, summary time, latency to begin summaries) were related to predictions of future comprehension performance. The well-established cues, accessibility of information and the situation model, seemed to both be utilized by participants to approximately the same extent, expanding our knowledge of how these cues are used. Thus, similar to a recent metamemory study (Undorf et al., 2018), we argue that participants use multiple cues in order to make predictions about their comprehension performance, but the cues can vary in validity, and some may be weighted more than others. Most of the cues (LSA was the exception) were valid, as they were significantly related to comprehension performance, but the gamma correlations were fairly low, indicating that there may be other more valid cues that should be used for prediction judgments. One possible explanation invokes the transfer-appropriate monitoring theory (Dunlosky et al., 2005), which posits that encoding and retrieval are more successful when the processes required at test match those employed at study. Because summarization involves some different cognitive processes than those used to successfully complete a multiple-choice comprehension test, high performance in one does not necessarily transfer to high performance in the other (Head et al., 1989). However, this explanation cannot fully account for current and previous findings. In our current study, the summarizing strategy, particularly in the written condition, still afforded greater metacognitive accuracy than not summarizing, so it seems that some part of the summary is representative of comprehension performance. Furthermore, Anderson and Thiede (2008) found a rather high correlation between gist-based ideas from summaries and multiple-choice performance. Even total ideas in that study were more highly correlated with multiple choice performance than in the present study, with their participants achieving much higher relative accuracy (Anderson & Thiede, 2008). Perhaps, then, the mismatch between processes involved in summarization and those involved in multiple-choice test performance is not fully to blame. Rather, poor summarizing and/or comprehension test performance in our sample may have diminished the relationship between the two, in part due to restriction of range.

The variation in cue validity deserves further interpretation. First, the average gamma correlation between comprehension and latency to begin summarizing was significantly greater than zero: the longer people paused before summarizing the worse they did on the comprehension test. Perhaps, this pause is an indicator of difficulty retrieving information about the text. This sense of disfluency was a fairly accurate cue, replicating some other disfluency findings (Pieger et al., 2016), although disfluency is not always the most beneficial for learning (Kühl & Eitel, 2016; Yue et al., 2013). Second, some cue validity depended on the summary modality. Summary time was related to comprehension in the oral summary condition but not in the written summary condition. Although we cannot be sure why, we conjecture that those in the written condition took time to reread or organize their summaries, such that time in the written condition was less correlated with actual content output than it was in the oral condition. In the oral condition, on the other hand, summary time was a decent indicator of how much one actually understood because people would stop summarizing when they could not recall or articulate more. Thus, summary time was more correlated with accessibility of information in the oral condition than it was in the written condition, making it a valid cue for oral summarizers.

Some describe fluency (at encoding and retrieval) as the most prevalent cue individuals use in making judgments (Koriat et al., 2004), but most fluency research is limited to the metamemory field. Typically, fluency is measured using reading speed, or recalling words and sentences (Benjamin et al., 1998; Pieger et al., 2016), and we do not believe others have assessed summary time as a cue for fluency. Most often, it is found that faster recall or processing leads to higher prediction judgment (Benjamin et al., 1998; Pieger et al., 2016; Rawson & Dunlosky, 2002); however, the current study finds that slower summary times are associated with higher judgments. Total summary time likely relates to greater accessibility of information; if the participants took a long time to summarize, they likely knew more about the subject. Another aspect to consider in the future is whether the participants took breaks while summarizing. One study conceptualized the fluency of processing as the regularity of task timing, rather than the speed of the task, and demonstrated that this consistency affects metacognitive judgments more than speed (Stevenson & Carlson, 2020). Measuring consistency of summary production may enhance our understanding of fluency's role in metacomprehension judgments in future research. Latency is likely a more traditional proxy for processing speed: when information comes quickly to mind, participants tend to believe they know the material better. However, according to pairwise comparisons, the gamma correlation between predictions and latency to begin summarizing was the only one that differed significantly from the gamma correlations between predictions and the other three cues. This suggests latency was not as strong of a cue as the other summary characteristics.

Based on the differences in fluency-related cues across conditions, it is surprising that average prediction magnitude did not differ between the summary modality conditions. However, past research has shown that average prediction magnitude tends to remain consistent and a within-subjects measure shows that judgments of performance are highly correlated after a week (Kelemen et al., 2000). It is possible that participants were using an unmeasured anchor while making judgments, and then based on cues during their summarization experience, they adjusted their predictions from this anchor (Zhao & Linderholm, 2008). So, even though written summaries were faster to begin summarizing, and took longer to complete (both associated with higher judgment magnitude), the average prediction magnitude was not higher because participants could have been adjusting from an anchor.

It was surprising that the situation model, as measured by LSA, was not correlated with multiple-choice performance. First, it was numerically only slightly smaller than the other measured cues. While the other cues were significantly different than zero, they were not much larger than the correlation between LSA and performance. Second, it is possible that participants are simply quite poor at judging the situation model. In Maki et al. (2009), the correlations between metacognitive judgments and LSA scores were quite low, although their method differed substantially from the one in the current study. Another possibility is that LSA is a poor measure of the situation model. As suggested by a reviewer, we assessed the relationship between LSA and inference questions, as the situation model is associated with a greater ability to make inferences. Surprisingly, despite a higher mean correlation, LSA was not more highly correlated with inference based questions ($g = 0.10$) on the comprehension test than questions that could be answered by what was explicitly stated in the texts ($g = -0.09$, $t(67) = -1.59$, $p = .11$), suggesting that those with higher LSA scores did not have better situational models. Thus, LSA may tap a slightly different construct, potentially summary quality, as our analysis suggests that LSA is a cue. Finally, it could be that participants' expectations about the comprehension test did not match the actual difficulty of the test (we did not offer a practice test). So, perhaps LSA did not correlate with performance because the comprehension test was unexpectedly challenging for the participants.

It should be noted that our participants had noticeably lower relative accuracy compared to past research; the average gamma correlation in the control condition was almost zero, while in the past, it has been 0.27 without intervention (Dunlosky & Lipko, 2007). Participants also scored more poorly on the multiple-choice comprehension test compared to other students on the same task (Fulton, 2021). This lower comprehension ability may have led to lower relative accuracy (Maki et al., 2005). Optimistically, relative accuracy in the written condition was similar to the average found in the literature (Dunlosky & Lipko, 2007). The written summary intervention appears to be effective at increasing relative accuracy, even when comprehension performance is low, and may help those with lower comprehension ability to maximize their learning potential.

Cognitive load

Our cognitive load results revealed a very different pattern than we predicted, with those in the written summarization condition reporting the lowest cognitive load, and oral summarizers reporting the highest. While initially surprising to some extent, there are several reasons why this pattern makes sense in hindsight. Those in the written condition had the ability to reread summaries, which could have allowed them to summarize the passages and monitor their comprehension separately in time, something those in the oral condition could not have done. Offloading in this way can free up cognitive resources (Risko & Dunn, 2015), which may then be allocated to metacomprehension monitoring, potentially allowing for more accurate judgments (Griffin et al., 2008). This possibility is partially supported by the cognitive load results. Whereas the intrinsic and germane loads reported were approximately the same between groups, extraneous load, the type of cognitive load that relates to presentation/environment, appears to be lowest in the written condition. Although differences between all groups in reported cognitive load only approached statistical significance ($p = .07$), a post hoc analysis comparing the two summary conditions shows that the written condition had a marginally significantly lower cognitive load than the oral condition, supporting the idea that cognitive offloading might explain differences between groups. Opportunities for offloading is a mechanism that may be worth exploring in the future to help explain group differences in metacomprehension monitoring accuracy. One study (Reid et al., 2017) found that cognitive load was higher in the summarizing condition than a control condition, contrary to the data in the current study, but they measured cognitive load before the comprehension test. It is unclear whether the type of measure is driving differences, as we used a survey that differentiated between the types of load, or due to the timing of the cognitive load questionnaire, as our measure occurred at the end of the multiple-choice test. Because our measure occurred after the comprehension test, we do not know if cognitive load acted as cue for participants' judgments, or whether our cognitive load survey measured perceived difficulty of the summaries, the multiple-choice test, both, or perhaps the entire experiment experience. However, differences seem to be due to summarization modality, as that was the only manipulated variable.

Limitations

First, we recognize that there may be other cues (e.g. prior knowledge) that influence prediction judgments (Koriat, 1997). Second, LSA was used to measure the situation model, because it has been used to measure summary quality in the past (Kintch, 1998; Maki et al., 2009; Thiede & Anderson, 2003). We used LSA to avoid a high correlation between word count and number of gist-based ideas and reduce potential bias of human scorers, but we acknowledge that there are other approaches to measuring the situation model (e.g., number of gist-based ideas; Anderson & Thiede, 2008), and each approach has its strengths and limitations. Third, some of our results were only marginally significant, and despite the conservative nature of gamma, replication is certainly necessary. A fourth limitation may be the relative lack of practice and experience with oral summarization among our participants. Although most people speak more than they write, summarization is more often done in writing. With the exception of presentations, oral assignments and examinations are relatively rare (Huxham et al., 2012). Although we do not know the exact experience that our participants have with oral summarization, it is likely a novel task for many. If students primarily summarize in the written modality, they may be more attuned to the cues available in written summaries and thus relative metacomprehension accuracy under those conditions could be largely an artifact of practice. Importantly, past research suggests that oral test anxiety is related in part to social anxiety (Laurin-Barantke et al., 2016) and, anecdotally, participants in the oral condition expressed nervousness when summarizing aloud in front of a researcher. As such, it may be the case that social anxiety was unintentionally evoked in some participants, and may be in part driving the marginal differences in cognitive load. The effects of social anxiety on metacomprehension monitoring accuracy is unknown, but is the topic of a study currently underway in our lab.

Future directions

Future studies could confirm whether writing summaries truly benefits relative accuracy in metacomprehension monitoring, whether summarization leads to utilization of multiple cues, and whether cognitive offloading (perhaps from summarizing in writing) is driving and benefiting accuracy. Some of our findings, that the accessibility of information and fluency were valid cues, is inconsistent with past findings (Dunlosky et al., 2005b; Koriat, 1993), so it is important to first replicate these effects and then understand which conditions allow the accessibility of information to be a valid cue. Finally, very few studies have addressed whether multiple cues increase accuracy, which is vital to our understanding and use of metacognitive strategies, so future research should confirm the benefit of multiple cue utilization to relative accuracy in metacomprehension monitoring.

Conclusions

Summarizing in writing may offer better relative metacomprehension accuracy than summarizing orally or not at all. Perhaps this is due to greater salience of cues, or cognitive offloading. Because higher relative accuracy improves study practices, such as a better allocation of study time, instructing students to summarize in writing, rather than orally, may improve their study efficiency, and therefore academic performance. With more research, we may find that instructing students to pay attention to multiple cues may additionally benefit their metacognitive accuracy. Although researchers have previously used summaries to assess the situation model and the accessibility of information (Anderson & Thiede, 2008; Maki et al., 2009), we believe that studying cues from a multi-cue perspective will advance our understanding of monitoring strategies.

Data availability Data will be made available on Open Science Framework.

Code availability Code will be made available on Open Science Framework.

Declarations

Ethics approval This study was approved by Idaho State University's Institutional Review Board.

Consent to participate Each participant signed an informed consent form before any part of the procedure.

Consent for publication Each author read and approved of final document before submission.

Conflicts of interest/Competing interests The authors declare no conflicts of interest.

References

Alterman, R. (1991). Understanding and summarization. *Artificial Intelligence Review, 5*(4), 239–254. https://doi.org/10.1007/bf00141756

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Pyschologica, 128*(1), 110–118. https://doi.org/10.1016/j.actpsy.2007.10.006

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. https://doi.org/10.3758/BF03193813

Bausell, R. B., & Li, Y. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. 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), 5–68. https://doi.org/10.1037/0096-3445.127.1.55

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology, 64,* 417–444. https://doi.org/10.1146/annurev-psych113011-143823

Board, T. C. (1997). *10 real SATs*. 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. https://doi.org/10.1080/00207599408248175

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2

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. https://doi.org/10.1016/j.jecp.2011.02.005

Dunlosky, J., Hertzog, C., Kennedy, M. R. F., & 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. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypothesis. *Journal of Memory and Language, 52*(4), 551–565. https://doi.org/10.1016/j.jml.2005.01.011

Efklides, A. (2014). How does metacognition contribute to the regulation of learning? An integrative approach. *Psychological Topics, 23*(1), 1–30.

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

Fulton, E. K. (2021). How well do you think you summarize? Metacomprehension accuracy in younger and older adults. *Journal of Gerontology: Series B, 76*(4), 732–740. https://doi.org/10.1093/geronb/gbz142

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

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. https://doi.org/10.3758/MC.36.1.93

Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(6), 1066–1092. https://doi.org/10.1037/xlm0000634

Head, M. H., Readence, J. E., & Buss, R. R. (1989). An examination of summary writing as a measure of reading comprehension. *Reading Research and Instruction, 28*(4), 1–11. https://doi.org/10.1080/19388078909557982

Hidi, S. E., & Hildyard, A. (1983). The comparison of oral and written productions in two discourse types. *Discourse Processes, 6*(2), 91–105. https://doi.org/10.1080/01638538309544557

Huxham, M., Campbell, F., & Westwood, J. (2012). Oral versus written assessments: A test of student performance and attitudes. *Assessment and Evaluation in Higher Education, 37*(1), 125–136. https://doi.org/10.1080/02602938.2010.515012

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. https://doi.org/10.3758/BF03211579

Kellogg, R. (2007). Are written and spoken recall of text equivalent? *American Journal of Psychology, 120*(3), 415–428. https://doi.org/10.2307/20445412

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

Kintsch, W., Welsch, D. M., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29*(2), 133–159. https://doi.org/10.1016/0749-596X(90)90069-C

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

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

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. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*(4), 643–656. https://doi.org/10.1037/0096-3445.133.4.643

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. https://doi.org/10.1037/0278-7393.32.3.609

Kühl, T., & Eitel, A. (2016). Effects of disfluency on cognitive and metacognitive processes and outcomes. *Metacognition and Learning, 11*(1), 1–13. https://doi.org/10.1007/s11409-016-9154-x

Landauer, T. K. (1998). Learning and representing verbal meaning: the latent semantic analysis theory. *Current Directions in Psychological Science, 7*(5), 161–164. https://doi.org/10.1111/1467-8721.ep10836862

Landauer, T. K. (2007). LSA as a theory of meaning. In Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.), *Handbook of latent semantic analysis* (pp. 3–34). Lawrence Erlbaum Associates Publishers.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Laurin-Barantke, L., Hoyer, J., Fehm, L., & Knappe, S. (2016). Oral but not written test anxiety is related to social anxiety. *World Journal of Psychiatry, 6*(3), 351–357. https://doi.org/10.5498/wjp.v6.i3.351

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*(6), 629–648. https://doi.org/10.1007/s11423-010-9153-6

León, J. A., & Escudero, I. (2015). Understanding causality in science discourse for middle and high school students. Summary task as a strategy for improving comprehension. In K. L. Santi & D. Reed (Eds.), *Improving comprehension for middle and high school students* (pp. 75–98). Springer International Publishing.

León, J. A., Olmos, R., Escudero, I., Cañas, J. J., & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods, 38*(4), 616–627. https://doi.org/10.3758/BF03193894

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 663–679. https://doi.org/10.1037//0278-7393.10.4.663

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. https://doi.org/10.1037/0022-0663.97.4.723

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*(2), 204–222. https://doi.org/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. https://doi.org/10.1080/13218719.2013.848001

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*(3), 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

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

Moore, D., Lin, L. M., & Zabrucky, K. M. (2005). A source of metacomprehension inaccuracy. *Reading Psychology, 26*(3), 251–265. https://doi.org/10.1080/02702710590962578

Morris, C. C. (1990). Retrieval Processes Underlying Confidence in Judgements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 223–232. https://doi.org/10.1037/0278-7393.16.2.223

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation, 26,* 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educational Psychologist, 38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency—Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction, 44,* 31–40. https://doi.org/10.1016/j.learninstruc.2016.01.012

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41*(1), 36–48. https://doi.org/10.3758/s13421-012-0245-x

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. https://doi.org/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. https://doi.org/10.3758/BF03209348

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. https://doi.org/10.1007/s11423-016-9454-5

Risko, E. F., & Dunn, T. L. (2015). Storing information in-the world: Metacognition and cognitive offloading in a short-term memory task. *Consciousness and Cognition, 36*, 61–74. https://doi.org/10.1016/j.concog.2015.05.014

Scardamalia, M., Bereiter, C., & Goelman, H. (1982). The role of production factors in writing ability. In M. Nystrand (Ed.), *What writers know: The language process and structure of written discourse*. Academic.

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

Stevenson, L. M., & Carlson, R. A. (2020). Consistency, not speed: temporal regularity as a metacognitive cue. *Psychological Research Psychologische Forschung, 84*(3), 88–98. https://doi.org/10.1007/s00426-018-0973-z

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129–160. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1080/01638530902959927

Undorf, M., Sollner, A., & Broder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition, 46*(4), 507–519. https://doi.org/10.3758/s13421-017-0780-6

Vanderberg, R., & Swanson, H. L. (2007). Which components of working memory are important in the writing process? *Reading & Writing, 20,* 721–752. https://doi.org/10.1007/s11145-006-9046-6

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. https://doi.org/10.1023/a:1007932429184

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. https://doi.org/10.1037/xap0000096

Yue, C., Castel, L., & Bjork, R. A. (2013). When disfluency is–and is not–a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory and Cognition, 41*(2), 229–241. https://doi.org/10.3758/s13421-012-0255-8

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychology Review, 20*(2), 191–206. https://doi.org/10.1007/s10648-008-9073-8