

(AI)nterventions: Exploring the effects of Gen-AI Support and prompting on Self-Regulated Learning

Lisa Smets (1774441)

Faculty of Social Sciences, Utrecht University

Master's Thesis Educational Sciences (201600025)

First Assessor: Dr. Jacqueline Wong

Second Assessor: Dr. Moniek Schaars

Juni, 10, 2024

Word Count: 6817

Abstract

In facilitating effective learning processes, self-regulated learning (SRL) is a widely acknowledged concept in the educational field. When applying learning from texts, learners are expected to self-regulate by employing various learning strategies, such as summarization. However, summarization can be challenging for learners who are not skilled nor trained. The availability of Gen-AI could potentially contribute to the design of innovative interventions supporting learners' SRL when learning from texts by summarizing complex texts to aid text comprehension. This study aims to gain insights into the effects of Gen-AI supported interventions on self-regulated learning (especially metacomprehension accuracy) and perceived cognitive load. Participants ($N = 105$) were randomly assigned to four conditions: 1) self-summary only, 2) self-summary with prompts, 3) Gen-AI summary only, 4) Gen-AI summary with prompts. No significant effects were found for conditions on perceived cognitive load or metacomprehension accuracy. By incorporating both SRL and cognitive load theory, this study sets the stage for the design of future Gen-AI interventions to promote SRL.

Keywords: Gen-AI, self-regulated learning, metacomprehension, cognitive load theory

Self-regulated learning (SRL) is considered necessary when individuals need to learn independently and handle diverse demands of a task (Roth et al., 2016). SRL's significance for academic achievement is demonstrated in several meta-analyses across education levels, ranging from higher education (Robbins et al., 2004) to adults' work-related learning (Sitzmann & Ely, 2011). To effectively self-regulate their learning, learners must set goals, monitor if that goal has been met, and regulate study activities to achieve it (Dunlosky & Thiede, 1998; Nelson & Narens, 1990; Thiede & Dunlosky, 1999; Winne & Hadwin, 1998). Despite the widely acknowledged importance of accurate monitoring for successful SRL (Butler & Winne, 1995; Dunlosky & Rawson, 2012), research shows learners are poor at self-monitoring and are often inaccurate in judging how well they have learned (Dunlosky & Lipko, 2007; Thiede et al., 2009). Specifically, monitoring learning and comprehension of texts, called metacomprehension accuracy (Maki & Berry, 1984), has typically been quite low across different educational levels and age groups (Yang et al., 2023; Thiede & Anderson, 2003; Maki, 1998). Learning from texts is crucial because texts serve as primary sources of information across various subjects, disciplines, and professional fields. If learners cannot monitor their text comprehension well, it undermines their ability to learn, think critically, make informed decisions, and ultimately, impacts their academic and professional success. Therefore, it is crucial to examine how learners can be supported when learning from texts.

For supporting metacomprehension, delayed generative learning strategies were found to be effective, such as delayed summary writing (Anderson & Thiede, 2008), delayed keyword listing (de Bruin et al., 2011) and delayed diagram completion (van de Pol et al., 2019). Despite their effectiveness, there are some factors that hinder learners' widespread use of these learning strategies (Bjork et al., 2013). To begin with, numerous effective generative learning strategies require learners to apply extra mental effort, which may debilitate learners from effectively utilizing them during SRL (Kirk-Johnson et al., 2019). Also, learners often have difficulties in spontaneously applying SRL strategies during learning activities (Azevedo et al., 2011). Considering these

complexities, on-demand, personalised, or adaptive support might be needed to support learners' independent implementation of SRL strategies (Wong et al., 2021).

As technologies become more advanced, like the recently launched generative AI (Gen-AI) chatbots, they can be leveraged as on-demand and personalized support to enhance SRL (Labadze et al., 2023). Gen-AI chatbots can perform language-related tasks based on large language models that leverage deep learning and advanced algorithms (Cotton et al., 2023). Unlike other tools, they offer real-time feedback and adapt responses based on users' progress, creating a personalized learning experience. Leveraging the capabilities of Gen-AI chatbots in SRL interventions holds promise in facilitating more efficient cognitive processing (Labadze et al., 2023). To this end, this study aims to explore how Gen-AI chatbots can be leveraged to support learning strategies when learning from texts. The study's result will contribute to the enrichment of SRL interventions by exploring the relationship between AI-assisted learning, metacognitive monitoring, and cognitive load.

Self-Regulated Learning and Metacognition

SRL involves cognitive, metacognitive, behavioral, and emotional aspects of learning, and refers to the extent to which individuals take control of their own learning and adapt their strategies to achieve academic goals (Panadero, 2017). Metacognition, described as the awareness and control of mental thoughts (Flavell, 1979), is a key component of SRL. It involves being aware of and understanding one's own thoughts, knowledge, and learning strategies. Learners' use of metacognitive strategies is significantly correlated with academic performance (Broadbent & Poon, 2015; Dent & Koenka, 2016). A key metacognitive strategy is metacognitive monitoring. Nelson and Narens (1990) described metacognitive monitoring as "the flow of information from the object-level to the meta-level", which involves the continuous assessment and evaluation of one's own learning progress. The accuracy of this monitoring (i.e., monitoring accuracy) affects the regulation of study activities (metacognitive control) and, therefore, influences learning gains (Thiede et al., 2009). In the context of reading comprehension, monitoring accuracy is called metacomprehension accuracy (Maki & Berry, 1984).

The relation between metacomprehension accuracy, metacognitive control, and SRL can be illustrated with an example of a student: Peter is reading a complex chapter in a history textbook (object-level). Peter starts reading; he notices that he reads faster than usual, and that there are no difficult words. Based on the reading experience, Peter judges his comprehension of the text as high because “it felt easy”. Peter does not base his judgment on predictive cues (e.g. “Can I summarize the main idea of this paragraph?”) for actual comprehension, but he uses ease of processing the text, a surface-level cue, for judging comprehension. Due to the faulty cue-use, he assumes that he comprehends the text well even though his actual comprehension is low. Peter’s overestimation of his comprehension (i.e., low metacomprehension accuracy) does not activate him to change his strategies to improve learning outcomes (metacognitive control), for example seeking additional clarification or revisiting challenging concepts (SRL strategies) (Bjork et al, 2013; Nelson & Narens, 1990). Metacomprehension accuracy can be measured by assessing one’s absolute metacomprehension accuracy, that is investigating whether metacomprehension judgments match performance exactly (Maki et al., 2005; Schraw, 2009).

Self-Management of Cognitive Load

Cognitive load theory emphasizes the necessity of managing cognitive load in complex and information rich environments given the limitations of one’s working memory (Sweller et al., 1998). When learners self-regulate their learning, taking control of their cognitive processes (e.g., metacognitive monitoring) demands cognitive resources (Seufert, 2018). Research shows that cognitive load plays a critical role in metacognitive monitoring (Wang et al., 2023; Van Gog et al., 2023). Van Gog et al. (2023) showed that, in more complex tasks, monitoring increases cognitive load and decreases performance. However, empirical research examining the relationship between cognitive load and SRL are still limited, and even fewer examining the impact of SRL interventions on cognitive load (Wang et al., 2023). Therefore, it is of interest in this study to examine cognitive load in SRL interventions.

Cognitive load can be categorized into three types: intrinsic, extraneous, and germane cognitive load (Sweller, 1994). Intrinsic cognitive load refers to the complexity of the learning task, determined by the number of interacting elements and learners' prior knowledge. Extraneous cognitive load refers to the additional mental effort imposed by instructional materials or presentation formats. Germane cognitive load encompasses the mental effort devoted to meaningful learning processes, such as organizing information and constructing knowledge. The cognitive theory of multimedia learning by Mayer (2005) provides principles to align instructional design with learners' cognitive architecture. Following these principles, optimal instructional design aligns the complexity of the task with the learner's prior knowledge levels and minimizes extraneous load (Mayer, 2005).

According to cognitive load theory, managing cognitive load is the primary interest of instructional designers and educators, expecting that through well-designed instructional materials extraneous cognitive load can be minimized (Mayer, 2005). Since not all learning materials are well-designed and learners have increased accessibility to other information in the digitalized world, taking an SRL perspective to understand how learners can self-manage their cognitive load adds new perspectives to theories of cognitive load and SRL (Mirza et al., 2020; Eitel et al., 2020; de Bruin & van Merriënboer, 2017). Mirza et al. (2020) found that lowering extraneous cognitive load is not only the task of the instructional designer or teacher but can also be dealt with by the learner, namely by self-managing the extraneous load of a task. By investing mental effort, learners can compensate for poor instructional design and therefore prevent high levels of extraneous load (Mirza et al., 2020). Eitel et al. (2020) in their literature review describe that learners' willingness to invest effort in cognitive processing of instructions depends on motivation and the results of metacognitive monitoring.

It is important to help learners monitor their levels of cognitive load and to help them distinguish between intrinsic and germane load (de Bruin & van Merriënboer, 2017). By understanding their cognitive load, learners can regulate their learning processes in ways that benefit

their learning outcomes (de Bruin & van Merriënboer, 2017). For example, if learners perceive high intrinsic load due to complex material, they might break down the complex information into smaller, manageable chunks helps reduce intrinsic load. Learners could also choose to put more cognitive resources into germane load, for example by self-explaining the complex material (Dunlosky., 2013).

Gen-AI chatbot as a SRL Intervention

When learning from texts, learners often use summarization as a study strategy (Dunlosky et al., 2013; Blasiman et al., 2017). The study by Bretzing and Kulhavy (1979) involved high school students reading prose passages, taking notes using different methods (summary, paraphrase, verbatim, letter search), and completing a retention and recall test. Results indicated that summarizing and paraphrasing notes led to better recall and comprehension. Research by King (1992) showed that for listening to a lecture, students who utilized summarization or self-questioning as a study strategy scored higher on a comprehension test than students who engaged in notetaking and reviewing. According to Chi (2009), reason behind the effectiveness of summarization is the comparison of constructive (generative, e.g. summarization, self-questioning) with active (notetaking, highlighting) strategies. Strategies that require learners to make sense of new information and go beyond the presented information are more effective than strategies that require only doing something physically (Chi, 2009).

Despite these findings of summarization as an effective study technique (Bretzing & Kulhavy, 1979; King, 1992), it falls in the middle when compared to other techniques, such as practice testing (Roediger & Karpicke, 2006b) and distributed practice (Budé et al., 2011). Dunlosky et al. (2013) in their literature review rate summarization as low utility. Summarization requires high levels of mental effort, as learners must remember all information they have read in their working memory. For learners with lower working memory, summarizing can be a difficult task, impacting the quality of their summaries. Dyer et al. (1979) in their study found that summarization was not benefitting learning performance because of the amount of information that was missing in the summaries.

While it is possible to improve summarization through extensive training, the need for time and resources makes training summarization skills less feasible (Dunlosky et al., 2013).

The study of Leopold et al. (2013) compared the effectiveness of using self-generated or pre-defined summaries for studying texts. Results showed that students who used pre-defined summaries performed better on the comprehension test (Leopold et al., 2013). Regarding cognitive processing, results of Leopold et al. (2013) indicate that pre-defined summaries allowed students to focus on the relevant content and to allocate their resources to germane cognitive processes (Mayer, 2009). In contrast, self-generating summaries might have increased students' levels of extraneous cognitive processing, keeping them from deep cognitive processing (Mayer, 2009).

By providing a high-quality summary of a complex text in a structured and simplified manner, Gen-AI support could potentially lower learners' extraneous cognitive load during reading tasks. However, according to the generative learning model (Wittrock, 1990), providing a Gen-AI summary might not be optimal for SRL since learners might not engage in active processing when they are presented with a predefined summary. To overcome this limitation of using pre-defined summaries, an effective approach to support learners' use of SRL strategies could be prompting. Research has shown that prompting is an effective approach to support learners' use of SRL strategies, assuming that learners do not engage in SRL strategies spontaneously (Bannert & Reimann, 2012). The study of Bannert et al. (2015) showed that students in the metacognitive prompts condition performed significantly better on knowledge tests than the control group. Results indicated that the use of prompts stimulates learners' engagement in SRL strategies and deeper information processing (Bannert et al. 2015). Specifically, when learners used prompts for reflecting on their own learning process, SRL strategies for deep understanding were fostered (Bannert et al., 2015). Zhang et al. (2015) in their study found that combining cognitive and metacognitive prompts especially helped learners with low metacognition in monitoring their learning. While providing learners with Gen-AI summaries potentially lowers learners' levels of extraneous load, adding prompts could stimulate their engagement in germane cognitive processes and their use of SRL strategies.

The Present Study

Given learners' increasing widespread use of Gen-AI for various purposes including generating summaries when learning from texts, there is an urgent need for research to understand the effects of using Gen-AI on learning outcomes and provide guidelines on using Gen-AI in education. The current study aims to investigate whether Gen-AI could potentially address the limitations of traditional summarization strategies and whether additional SRL support (e.g., prompts) is needed. The immediate summary condition of Anderson & Thiede (2008) was replicated and used as comparison to examine the effects of the Gen-AI interventions, resulting in four experimental conditions: 1) self-summary only, 2) self-summary with prompts, 3) Gen-AI summary only, 4) Gen-AI summary with prompts. The main research question is formulated as *What is the effect of Gen-AI support and prompting on perceived levels of cognitive load, and metacomprehension accuracy?*

Well-designed instructions are intended to minimize extraneous load and optimize germane load (Mayer, 2005). Receiving support from Gen-AI to create a summary could be less cognitively demanding than having learners create their own summary since they do not have to go through the effortful process of retrieving information from memory. Instead, they could use the summary to evaluate the information they acquired from the previously read text. Furthermore, the summary could help learners to identify information that they might have missed during reading. Therefore, we hypothesize that learners in the Gen-AI support conditions (both with and without prompts) will perceive lower levels of extraneous cognitive load than learners in the self-summarize conditions (Hypothesis 1A).

Under the assumption that learners in the Gen-AI support conditions (both with and without prompts) would perceive lower extraneous load, they will also have more resources to devote to germane processing (Mayer, 2005). Additionally, guidance in the form of prompting is expected to stimulate learners' cognitive processes (Bannert & Reimann, 2012). Therefore, we expect learners in the Gen-AI support with prompts condition to perceive the highest levels of germane load, followed

by the Gen-AI support only condition, then the self-summary with prompts condition, and finally the self-summary only condition (Hypothesis 1B)

Previous research emphasizes the positive effect of prompting on learners' use of metacognitive strategies (Zhang et al., 2015). Also, lowering learners' extraneous load may allow learners to allocate more of their cognitive capacity to germane processes which include also metacognitive activities (Mayer, 2005). Van Gog et al. (2023) found that in complex tasks, cognitive load and metacognitive monitoring processes are negatively related. Therefore, we hypothesize that learners who receive a Gen-AI summary with the addition of support in the form of prompts have highest metacomprehension accuracy, followed by the Gen-AI summary only condition, then the self-summary with prompts condition, and finally the self-summary only condition (Hypothesis 2).

Method

Research Design

The present study was part of a larger study, called (AI)ntervention. An experimental quantitative 2 (summary: self and Gen-AI) x 2 (reading prompts: with or without) between-group design was deployed, resulting in four conditions: (1) self-summary only, (2) self-summary with prompts, (3) Gen-AI summary only (4) Gen-AI summary with prompts. The dependent variables were cognitive load and metacomprehension accuracy. A priori power analysis was conducted using G*Power analysis tool (version 3.1.) to determine the minimum sample size (Faul et al., 2019). Results indicated that the required sample size to achieve 80% power for detecting a medium effect ($f = .25$) based on the effect sizes reported in Anderson and Thiede (2008) and Prinz et al. (2020) at a significance criterion of $\alpha = .05$, was $N = 180$. The ethics application for the study was approved by the university's faculty ethical review board (FERB) filed under number 24-0070.

Participants

There were 105 participants (53.3% male, 46.7% female) who completed the online experiment, despite having 268 respondents who started the experiment (39.18% response rate). No

rewards were given for participation in the study. The largest group of participants were between the ages of 25 and 44 (66.6%), the rest differed between 19-24 (8.6%), and 45-84 (24.8%). Most participants had a university education (bachelor or master level) (60%) and the rest differed among various levels (3.8% high school, 1.9% secondary vocational education, 23.8% higher professional education, 6.7% above university masters' level, 3.8% preferred not to say).

Instrumentation and Measurement

Learning Task and Performance Test

The texts and performance test were taken from Anderson and Thiede (2008). The texts were expository texts about the IQ and monetary policy that described complex causal relations (Appendix A). Each text was approximately 600 words long, with an average Flesch-Kincaid readability score of 11.8. Each text had corresponding multiple-choice test items designed to assess comprehension rather than memory of details in the text (see example in Appendix B). The final test consisted of 10 questions in total (5 questions per text). One point was awarded for each question that was answered correctly or 0 points if answered incorrectly. A maximum of 10 points could be obtained. To measure performance, sum scores were calculated for each participant. The reliability of the comprehension test was assessed using Cronbach's alpha, which was found to be 0.62, indicating a low reliability.

Generative AI summary

ChatGPT-3.5 was used for creating the summaries for the Gen-AI conditions. For both texts, the original content was first uploaded to ChatGPT-3.5. Then, a prompt was used to generate a few summary versions for each text. The prompts included the components "qualitative summary", "key aspects", and "to prepare for comprehension test". The different versions of summary were checked against the original comprehension text and test items to create a summary that was between 150 and 250 words long for each text.

Reflection Prompts

For the self-summary with prompts condition and the Gen-AI summary with prompts conditions, three reflection prompts were added to the task of reading either their own summary (condition 2) or the Gen-AI summary (condition 4). The reflection prompts were: 1) “How well would your own/this Gen-AI summary help others to gain an understanding of the original text you summarized?” 2) “What are the key concepts of the text you/Gen-AI summarized?”, 3) “What would you add to your own/this Gen-AI summary?”.

Metacomprehension Judgments

Metacomprehension judgments were measured by a one-item survey adapted from Glenberg and Epstein (1985) asking participants to judge their comprehension of each text by answering the question: “*How well do you think you understood each text?*”. Glenberg and Epstein’s (1985) approach has been widely utilized in metacomprehension research and has demonstrated reliability in previous studies. In current study, Cronbach’s alpha was .73, indicating acceptable reliability. The title of the text was displayed on the left side of the screen, next to which, on the right side, participants were asked to score their metacomprehension on a 7-point scale, ranging from 1 (very poorly) to 7 (very well). Metacomprehension was operationalized as the mean score of the comprehension rating across the two texts.

Metacomprehension Accuracy

For assessing the precision of participants’ metacomprehension judgments, metacomprehension accuracy was measured by the absolute accuracy. These absolute accuracy scores can be interpreted as the discrepancy between metacomprehension judgments and performance (Schraw, 2009). To calculate metacomprehension accuracy, the metacomprehension judgements scores were converted into a 10-point scale since the test score consisted of a 10-point scale. For example, if a participant judged his metacomprehension by rating it a 3.5 on the 7-point scale, this rating would be converted into a 5 on a 10-point scale. If the actual performance score of this participant was 7, the absolute metacomprehension accuracy score would be 2.

Cognitive Load

Cognitive load was operationalized by measuring the three types of cognitive load identified by Sweller (1994): intrinsic, extraneous, and germane cognitive load. An adaptation of the questionnaire validated by Krieglstein et al. (2023) was used for measuring participants' perceived levels of cognitive load (Appendix C). The adapted questionnaire included eleven items regarding cognitive load (see Table 1): Intrinsic cognitive load (3 items), extrinsic cognitive load (3 items), and germane cognitive load (5 items) were measured by statements on a 7-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). Reliability of each subscale was assessed by Cronbach's alpha. For intrinsic load, Cronbach's alpha was initially .59, indicating low reliability. Item analysis showed that deleting item 3 (ICL3) would improve the reliability. To improve reliability, item 3 was deleted, resulting in acceptable reliability for intrinsic load ($\alpha = .73$). Reliability was acceptable for extraneous load ($\alpha = .77$), and germane load ($\alpha = .78$). Mean scores for each subscale were calculated. For each subscale, mean scores were interpreted as ranging from low (1) to high (7) levels of perceived cognitive load.

Table 1

Cognitive Load Questionnaire

Cognitive Load Type	Item
<i>Intrinsic Cognitive Load</i>	
ICL1 (item 1)	The texts were difficult to understand.
ICL2 (item 2)	The texts were complex.
ICL3 (item 3)	Without prior knowledge, the texts were not understandable.
<i>Extrinsic Cognitive Load</i>	
ECL1 (item 4)	It was difficult to gain an overview of the structure of the texts.
ECL 2 (item 5)	The design of the summary activity made it difficult to find relevant information quickly.
ECL 3 (item 6)	Because of the design of the summary activity, I had the impression that I could not concentrate on the content.
<i>Germane Cognitive Load</i>	
GCL 1 (item 7)	I actively reflected upon the texts.

GCL 2 (item 8)	I made an effort to understand the texts.
GCL 3 (item 9)	I achieved a comprehensive understanding of the texts.
GCL 4 (item 10)	I was able to expand my prior knowledge with the texts.
GCL 5 (item 11)	I can apply the knowledge that I acquired through the texts quickly and accurately.

Note. ICL = Intrinsic Cognitive Load, ECL = Extraneous Cognitive Load, GCL = Germane Cognitive Load.

Pilot

All instructions and materials were provided in English. The full instrument was piloted with three researchers to identify errors and check for clarity of language and instruction. Based on their feedback, several adjustments were made. The font size was changed to a bigger font, consistently for the whole experiment. A short repetition of instructions about what learning tasks participants will perform was added before each task. The question “Did you think about your summary when you made your rating for how well you understood the texts?” was deleted, because of this was perceived as confusing for the participants and did not have a clear purpose. Terminology in the cognitive load questionnaire was adjusted from an inconsistent use of the words: ‘information,’ ‘learning content,’ or ‘learning materials’ into a consistent use of ‘texts’. The randomization of the comprehension test questions was modified from randomization of all ten questions to only randomization within the five questions per texts.

Procedure

Participants were invited to participate voluntarily and anonymously in an online experiment. The link of the online experiment was distributed by several initiatives. Flyers were shared with students and educators at the University. People with interests in education or Gen-AI were recruited via LinkedIn. Emails were sent to teachers and professors from three other University Colleges across the country in the Netherlands. After clicking on the anonymous link, participants received an information letter and consent form regarding the purpose of the study, the study procedure, duration, data storage, and protection. Participants who gave consent for participating on a

voluntary and anonymous basis proceeded with participating in the experiment. The intervention took place in an online survey platform, Qualtrics.

Participants were randomly assigned to one of the four conditions. Table 2 provides an overview of the procedure across the four conditions. For each text, all participants had 7 minutes to read. Depending on the condition they were assigned to, participants received associated instructions about the learning task (Appendix D). The learning task started with generating a summary, either by self-summarizing or by writing a prompt to let Gen-AI generate a summary. After generating a summary, participants reread their summary, with or without reflection prompts (depending on the conditions).

After completion of the learning tasks, all participants judged their metacomprehension for both texts and answered a survey about perceived levels of the three types of cognitive load. Then, participants performed a comprehension test, which contained five questions about each text. Finally, participants responded to one close-ended question in which they were asked if they have used generative AI before. If the answer was “yes”, participants received a question about how often and for which purposes they have used generative AI before. The experiment concluded with an option for participants to voluntarily fill in their e-mail addresses to receive the research paper after the research project has finished.

Table 2*Overview of the Procedure Across the Four Conditions*

Condition 1	Condition 2	Condition 3	Condition 4
Instruction of the topic and tasks			
Read text 1	Read text 1	Read text 1	Read text 1
Summarize text 1	Summarize text 1 + Reading prompts	Gen-AI summary text 1	Gen-AI summary text 1 + Reading prompts
Read text 2	Read text 2	Read text 2	Read text 2
Summarize text 2	Summarize text 2 + Reading prompts	Gen-AI summary text 2	Gen-AI summary text 2 + Reading prompts
Metacomprehension Judgments (2 items)			
Mental Effort (2 items)			
Comprehension Test (Performance score) (10 items)			
Gen-AI usage (1-3 items)			
Debriefing			

Note. Additional steps in the experiment not used in this research are left out.

Data Analysis

Data Preparation

All data was anonymized by detaching participants' email addresses (for receiving the research paper once this study is finished) and stored in a secure research data-management system hosted by the university, Yoda. Participants who did not consent were deleted from the dataset, as well as those who did not finish the whole experiment. The data was analysed using SPSS. Potential outliers in a cleaned dataset were identified by visual inspection of boxplots for each outcome variable. No outliers were identified.

Statistical Assumption Checks

Before conducting each analysis, the corresponding assumptions were checked. Normality was evaluated by a Kolmogorov Smirnov test and was indicated by a non-significant result. In case of violated normality assumptions, data transformations were considered. If normality was not attained through transformation, non-parametric tests like Mann-Whitney U or Kruskal-Wallis tests were used, ensuring robust analyses despite normality deviations. For each analysis, relevant assumptions were checked. Homogeneity of variances (homoscedasticity) was checked by using Levene's test. For the measurement of the correlation coefficients, linearity between the two variables was checked. If

assumptions were violated, a sensitivity analysis was performed to assess the robustness of the results by exploring the impact of assumption violation. Also, any violations were transparently reported. A significance level of $p \leq .05$ was applied for all statistical tests. In addition, significant differences between the experimental conditions post-hoc tests with $\eta^2 \geq .01$ were considered small effect size, $\eta^2 \geq .06$ medium effect size, and $\eta^2 \geq .14$ large effect size (Cohen, 1988). Cronbach's alpha was used for reliability analysis, with cutoff scores of 0.70 indicating acceptable reliability and 0.80 or higher indicating good reliability. Assumption checks for the respective statistical analysis were conducted, all assumptions were met. Results of the assumption checks are included in Appendix E.

To gain insights on the effect of conditions on cognitive load, we conducted a MANOVA which included all three types of cognitive load and the four experimental conditions. Mean-scores for each of the three cognitive load types were used as dependent variables, and the experimental conditions were fixed factors. Post-hoc analyses using Bonferroni were performed to examine individual differences across the conditions and three cognitive load scales. Pillai's Trace was used to test whether the experimental conditions significantly explained the variance on the cognitive load questionnaires.

To gain insights on the effect of the conditions on metacomprehension accuracy, we conducted one independent ANOVA, which evaluated the effect of condition on metacomprehension accuracy. Condition was used as fixed factor and metacomprehension accuracy was used as dependent variable to identify if the dependent variable differed across conditions. If there were significant differences, those differences were evaluated with a Bonferroni post-hoc analysis to identify which conditions differed from each other.

Results

A correlational analysis showed significant relationships between study variables. The reliability coefficients between study variables are shown in Table 3. Table 4 provides the means and standard deviations for performance, metacomprehension judgments, metacomprehension accuracy, and three types of cognitive load per condition.

Table 3*Reliability and Correlations of Study Variables*

Variable	1	2	3	4
1. Performance				
2. Metacomprehension judgements	.50**			
<i>Cognitive Load</i>				
3. Intrinsic Cognitive Load	-.30**	-.60**		
4. Extraneous Cognitive Load	-.32**	-.45**	.50**	
5. Germane Cognitive Load	.47**	.71**	-.38**	-.40**

Note. $N = 105$. * Correlation is significant at $p < .05$ (two-tailed test). ** Correlation is significant at $p < .01$ (two-tailed test). ^a Cronbach's Alpha (α) reliability coefficient.

Table 4*Means and Standard Deviations of Study Variables per Experimental Condition*

Variable	Self-summary only ($N = 25$)		Self-summary + prompts ($N = 34$)		Gen-AI summary only ($N = 25$)		Gen-AI summary + prompts ($N = 21$)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Performance	6.12	2.09	5.71	1.99	5.64	2.25	6.38	2.33
Metacomprehension judgements	6.80	2.00	7.16	1.91	7.14	1.52	7.65	1.62
Metacomprehension accuracy	1.81	1.19	1.95	1.21	2.06	1.62	1.79	1.52
<i>Cognitive Load</i>								
Intrinsic	4.32	1.51	3.65	1.18	4.34	0.98	3.81	1.45
Extraneous	4.09	1.19	3.57	1.32	3.99	1.18	3.56	1.50
Germane	4.78	0.96	4.70	1.08	4.84	0.75	5.07	1.11

Cognitive Load

For intrinsic cognitive load, means in the self-summary only condition and the Gen-AI summary only condition were above 4, while the conditions that were supported with reflection

prompts were below 4. For extraneous load, only the means in the self-summary only condition was above 4. Regarding germane load, means in all conditions were above 4, highest levels of germane load were reported in the Gen-AI summary with prompts condition ($M = 5.07$, $SD = 1.11$).

A MANOVA was used to examine the effect of conditions on participants' perception of intrinsic, extraneous, and germane cognitive load. While correlations between the three types of cognitive load are significant, they do not impact multicollinearity for the MANOVA. The MANOVA was statistically non-significant, $F(3) = 1.21$, $p = .291$, partial $\eta^2 = .04$, indicating the absence of any meaningful differences on perceived levels of cognitive load between conditions. Group means (and standard deviations) for each dependent variable are presented in Table 4. Based on these findings, hypotheses 1A and 1B are rejected.

Metacomprehension Accuracy

To evaluate the effectiveness of four conditions on metacomprehension accuracy, three one-way ANOVAs were conducted. The first ANOVA investigated the effect of conditions on learning performance. The F test was not significant, $F(3) = .65$, $p = .586$, $\eta^2 = .02$, which is characterized as a small effect. The second ANOVA investigated the effect of condition on metacomprehension judgments and was also not significant, $F(3) = .87$, $p = .460$, $\eta^2 = .03$, considered a small effect. The third ANOVA investigated the effect of condition on metacomprehension accuracy and was also not significant, $F(3) = .21$, $p = .891$, $\eta^2 = .01$, considered a small effect. The results suggest that conditions did not differ significantly regarding their levels of performance, metacomprehension judgments, and metacomprehension accuracy. Therefore, hypothesis 2 is rejected.

Discussion

Self-regulated learning (SRL) plays an important role in academic success, but many learners are poor at SRL and need to be supported to enhance their SRL (Dunlosky & Lipko, 2007; Thiede et al., 2009). With the emergence of Gen-AI tools, learners have the opportunity to use such tools to support their SRL, for example by supporting summarization (Leopold et al., 2013). The present study investigated the effects of SRL interventions with Gen-AI support and reflection prompts on

participants' cognitive load (intrinsic, extraneous, germane) and metacomprehension accuracy. The results revealed no significant effect of the Gen-AI support or reflection prompts to aid summarization when learning from texts on cognitive load or metacomprehension accuracy. The following discussion provides insights into the key findings (or the lack thereof), their implications and potential opportunities for future research.

Effects of SRL Interventions on Cognitive Load

Optimizing learning by reducing extraneous cognitive load remains an important aspect of instructional design (Mayer, 2005). Prior research based on Mayer's cognitive theory of multimedia learning suggested that well-designed interventions, such as using Gen-AI summaries, could potentially reduce extraneous cognitive load (Mayer, 2005). Similarly, past research showed that predefined summaries can be helpful in lowering extraneous cognitive load (Leopold et al., 2013). However, our results did not reflect any beneficial effect of providing Gen-AI summary on reducing extraneous cognitive load.

One of the reasons for the lack of effect on extraneous cognitive load could be that the task is not overly complex. Moreover, the learners have high educational background levels. Therefore, they could have rich experience in summarization. The mean scores of the extraneous load across all four conditions were moderate (around four out of a scale of seven), suggesting that the participants, on average, did not find it too challenging to gain an overview of the structure of the texts nor did the summary activity made it too difficult for them to find the relevant information or focus on the content. Future research could examine the influence of task complexity in the effect of Gen-AI interventions on extraneous cognitive load. Given that (Van Gog et al., 2023) found that monitoring decreases performance and increases cognitive load only in complex tasks, it could be that providing Gen-AI support and reflection prompts to reduce extraneous cognitive load is only beneficial when the task is too complex, for example when reading lengthy texts on difficult topics.

Additionally, the summarization task was given right after the reading task. Therefore, information from the texts that they have read are still active in their working memory and can be

more easily retrieved (Sweller et al., 1998). This could have resulted in a comparable ease being experienced when summarising immediately after reading and when studying a Gen-AI provided summary. The study employed a between-subjects design and the participants were not given an opportunity to experience both self-summarising and studying a Gen-AI summary. It could be interesting to examine whether experiencing the ease of Gen-AI summary would have an impact on perceived extraneous load when compared to self-summarising. However, there are concerns about learners relying too heavily on Gen-AI models (Kasneci et al., 2023). Educators and learners should be aware of the way they use Gen-AI and its potential risks: Gen-AI should function as a support tool to enhance learning, not as a replacement for critical thinking (Kasneci et al., 2023). Future research could explore how Gen-AI could support other study strategies than summarization, for example self-testing. Self-testing is considered as an effective study strategy (Dunlosky et al., 2013). Despite students' intentions to use self-testing for studying, they reported using practice testing less often than other study strategies that are less effective (Blasiman et al., 2017). Reasoning behind this is that students start studying later than planned, and then tend to go for easier-to-use, but less effective study strategies such as reading notes or rereading text (Blasiman et al., 2017). Future research could explore how Gen-AI could support students' use of effective study strategies like self-testing, for example by generating questions for self-testing.

For germane cognitive load, the results stand in contrast to previous research that found beneficial effects of prompting (Bannert et al., 2015; Zhang et al., 2015). Prompting could potentially enhance germane cognitive load by encouraging deeper processing (Bannert et al., 2015). Additionally, previous research indicated that lowering extraneous load can help learners allocate cognitive resources to germane cognitive load (Mayer, 2005). The lack of significant effect of prompting on germane cognitive load could be related to learners' motivation. Mirza et al. (2020) describe that learners can choose to allocate resources to germane processes, and therefore self-manage their levels of cognitive load. Similarly, Eitel et al. (2020) highlight that learners' invested effort in cognitive processing depends not only on the intervention design and the results of

metacognitive monitoring, but also on their motivation levels. These studies suggest that student motivation may play a critical role in whether students engage in germane processing. The current study is conducted as an online experiment where participation is voluntary with no rewards. Therefore, it could be that the participants are not motivated to put in the effort to learn the text. Furthermore, the topics of the texts might not be personally relevant to the participants. Motivational beliefs are drivers of self-regulated behaviors (Schunk & Zimmerman, 2012). Future research could include motivational aspects to understand their impact on self-management of cognitive load. For example, by integrating the study into a course, providing rewards, or using texts relevant to the participants.

Effects of SRL Interventions on Metacomprehension Accuracy

While we hypothesized that Gen-AI summaries and prompting would support learners' metacomprehension accuracy, this was not reflected in the results of this study. For comprehension judgments, participants generally overestimated their comprehension levels, this is in line with previous studies that found that learners often overestimate their comprehension and knowledge (Dunlosky & Lipko, 2007). Interestingly, the use of Gen-AI did not lead to increased overestimation of comprehension. This suggests that Gen-AI tools do not exacerbate learners' metacognitive misjudgments, which is a positive finding for their potential use in educational settings. However, the general overestimation of comprehension highlights the importance of developing interventions that support metacomprehension accuracy.

Prior research has indicated that learners' use of SRL strategies and monitoring processes can be supported by prompts (Bannert et al., 2015; Zhang et al., 2015). However, this study did not find significant improvements in metacomprehension accuracy due to the interventions. This insignificance may be due to the lack of learners' self-assessment skills, and therefore indicate a need for more explicit guidance or scaffolding to help learners utilize metacomprehension activities effectively. Researchers argue that scaffolding metacognitive skills is necessary to effectively support learners in developing accurate metacomprehension (Seufert, 2018). Previous research of van Loon

et al. (2017) showed that learners who received directly feedback were able to judge their learning more accurate than learners who had to observe their own learning process. A strategy to increase learners' self-assessment could be to incorporate the principles of providing explicit guidance or scaffolding by van Merriënboer and Sluijsmans (2009). For example, initially using Gen-AI as a supportive tool to provide direct feedback and gradually fading this support until learners are able to assess their own learning accurately. Another approach is to scaffold self-assessment skills by initially focusing on one indicator, such as the ability to self-explain concepts, until learners can effectively evaluate all aspects of their learning. Future research could for example investigate these scaffolding techniques to improve metacomprehension accuracy. Additionally, by conducting studies that use Gen-AI in real web-environment instead of a simulation, Gen-AI could also function as a feedback mechanism and provide adaptive assistance and feedback.

Limitations and Future Recommendations

There are some limitations that should be considered when interpreting the results. First, for satisfactory statistical power to be achieved, data of 180 participants were required. The sample size was too low to meet this criterium. Small sample sizes lead to a lower power (Jones et al., 2003), and can significantly affect the results of the analyses (Fitzner & Heckinger, 2010). The small sample size might have contributed to the insignificant results of this study. The participants consisted of a large group of Social Sciences students and teachers, which might have resulted in higher levels of prior knowledge for the text about IQ. This homogeneity in educational background impacts the generalizability of the findings of this study to other educational disciplines. Future studies should aim for larger and more diverse samples to enhance the robustness and generalizability of results.

Second, the study did not account for the type and quality of prompts for the summaries provided. We used standardized Gen-AI summaries, ensuring that the quality of the summaries was not influenced by the learners' prompts. This approach, however, did not allow us to assess the learners' abilities to generate high-quality prompts. Future research could explore learners' abilities to write qualitative prompts to better understand how prompt quality influences the effectiveness of

the interventions. Furthermore, Gen-AI applications can be conversational and interactive (Kasneci et al., 2023). Therefore, it could help learners process the text more if they could discuss the text with a Gen-AI chatbot rather than only using it to create a summary. Further research could explore other ways in which Gen-AI can be employed during SRL.

Finally, the use of only self-reports to measure cognitive load is a limitation of this study. When participants self-report their cognitive load levels, results are based on perceptions and thus can be biased and inaccurate. Also, participants might have responded in a way which they believe is socially acceptable or favourable. These factors may have influenced the results and affect the validity of findings. Future research could consider complementing self-report measures with objective methods such as eye movements or the index of cognitive activity to provide a more comprehensive assessment of cognitive load (Korbach et al., 2017).

Conclusion

Despite the limitations and lack of findings, this study contributes to research at a theoretical level by examining cognitive load when supporting SRL and at a practical level by investigating the use of Gen-AI summary as opposed to self-summarization when learning from texts. While studying a Gen-AI text did not reduce extraneous cognitive load nor increase germane load and metacomprehension accuracy, it did not seem to have a negative impact. Given that we only examine this in a one-off task, more research is needed to examine the long-term effects of using Gen-AI on SRL and learning outcomes. With the growing widespread use of Gen-AI among learners in higher education, there remains a need to explore how Gen-AI can be optimally used by learners to enhance their SRL and academic success while considering the ethical concerns and potential bias of Gen-AI outputs (Labadze et al., 2023). By highlighting the importance of self-management of cognitive load and scaffolding learners' metacomprehension, this study sets the stage for the design of future Gen-AI interventions to promote SRL.

References

- Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, *128*(1), 110–118.
<https://doi.org/10.1016/j.actpsy.2007.10.006>
- Azevedo, R., Johnson, A., Chauncey, A., Graesser, A., Zimmerman, B., & Schunk, D. (2011). Use of hypermedia to assess and convey self-regulated learning. In B. J. Zimmerman, D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance*, (pp. 102-121). Routledge.
- Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science*, *40*, 193-211. <https://doi.org/10.1007/s11251-011-9167-4>
- Bannert, M., Sonnenberg, C., Mengelkamp, C., & Pieger, E. (2015). Short-and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. *Computers in Human Behavior*, *52*, 293-306.
<https://doi.org/10.1016/j.chb.2015.05.038>
- 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-psych-113011-143823>
- Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2016). The what, how much, and when of study strategies: comparing intended versus actual study behaviour. *Memory*, *25*(6), 784–792.
<https://doi.org/10.1080/09658211.2016.1221974>
- Bretzing, B. H., & Kulhavy, R. W. (1979). Notetaking and depth of processing. *Contemporary Educational Psychology*, *4*(2), 145-153. [https://doi.org/10.1016/0361-476X\(79\)90069-9](https://doi.org/10.1016/0361-476X(79)90069-9)
- Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, *27*, 1-13. <https://doi.org/10.1016/j.iheduc.2015.04.007>

- Budé, L., Imbos, T., van de Wiel, M. W., & Berger, M. P. (2011). The effect of distributed practice on students' conceptual understanding of statistics. *Higher Education, 62*, 69-7.
<https://doi.org/10.1007/s10734-010-9366-y>
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281. <https://doi.org/103102/00346543065003245>
- Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73-105. <https://doi.org/10.1111/j.1756-8765.2008.01005.x>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- de Bruin, A. B. H., & van merriënboer, J. J. G. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction, 51*, 1-9.
<https://doi.org/10.1016/j.learninstruc.2017.06.001>
- De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. S. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294–310. <https://doi.org/10.1016/j.jecp.2011.02.005>
- Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, 28*, 425-474. <https://doi.org/10.1007/s10648-015-9320-8>
- Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232.
<https://doi.org/10.1111/j.1467-8721.2007.00509.x>

- Dunlosky J., & Thiede, K.W. (1998) What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica*, 98(1), 37-56. [https://doi.org/10.1016/s0001-6918\(97\)00051-6](https://doi.org/10.1016/s0001-6918(97)00051-6)
- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction*, 22(4), 271–280. <https://doi.org/10.1016/j.learninstruc.2011.08.003>
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4-58. <https://doi.org/10.1177/1529100612453266>
- Dyer, J. W., Riley, J., & Yekovich. (1979). An analysis of three study skills: notetaking, summarizing, and rereading. *Journal of Educational Research*, 73, 3–7. <https://doi.org/10.1080/00220671.1979.10885194>
- Eitel, A., Endres, T., & Renkl, A. (2020). Self-management as a bridge between cognitive load and self-regulated learning: The illustrative case of seductive details. *Educational Psychology Review*, 32(4), 1073-1087. <https://doi.org/10.1007/s10648-020-09559-5>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41. <https://doi.org/10.3758/BRM.41.4.1149>
- Fitzner, K., & Heckinger, E. (2010). Sample size calculation and power analysis: a quick review. *The Diabetes Educator*, 36(5), 701-707. <https://doi.org/10.1177/0145721710380791>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10). <https://doi.org/10.1037/0003-066X.34.10.906>

- Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 702–718.
<https://doi.org/10.1037/0278-7393.11.1-4.702>
- Jones, S., Carley, S., & Harrison, M. (2003). An introduction to power and sample size estimation. *Emergency Medicine Journal*, *20*(5), 453. <https://doi.org/10.1136/emj.20.5.453>
- Kasneji, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... & Kasneji, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, *29*(2), 303-323.
<https://doi.org/10.3102/00028312029002303>
- Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, 101237.
<https://doi.org/10.1016/j.cogpsych.2019.101237>
- Korbach, A., Brünken, R., & Park, B. (2017). Measurement of cognitive load in multimedia learning: a comparison of different objective measures. *Instructional Science*, *45*, 515-536.
<https://doi.org/10.1007/s11251-017-9413-5>
- Krieglstein, F., Beege, M., Rey, G. D., Sanchez-Stockhammer, C., & Schneider, S. (2023). Development and validation of a theory-based questionnaire to measure different types of cognitive load. *Educational Psychology Review*, *35*(1), 9. <https://doi.org/10.1007/s10648-023-09738-0>
- Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, *20*(56). <https://doi.org/10.1186/s41239-023-00426-1>

- Leopold, C., Sumfleth, E., & Leutner, D. (2013). Learning with summaries: Effects of representation mode and type of learning activity on comprehension and transfer. *Learning and Instruction*, 27, 40-49. <https://doi.org/10.1016/j.learninstruc.2013.02.003>
- Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.) *Metacognition in educational theory and practice*, (pp. 131-158). Routledge.
- 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 Educational Psychology*, 97(4), 723. <https://doi.org/10.1037/0022-0663.97.4.723>
- Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.) *The Cambridge handbook of multimedia learning*, (pp. 31-48). Cambridge University Press.
- Mayer, R. E. (2009). *Multimedia Learning (2nd ed)*. New York: Cambridge University Press.
- Mirza, F., Agostinho, S., Tindall-Ford, S., Paas, F., & Chandler, P. (2019). Self-management of cognitive load: 157 Potential and challenges. In S. Tindall-Ford, S. Agostinho, & J. Sweller (Eds.) *Advances in cognitive load theory* (pp. 157-167). Routledge.
- Nelson, T.O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings, In G.H. Bower (Ed.), *The psychology of learning and motivation*, (pp. 125-141). Academic Press, Inc.
- Panadero, E. (2017). A review of self-regulated learning: six models and four directions for research. *Frontiers in Psychology*, 8(422). <https://doi.org/10.3389/fpsyg.2017.00422>
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130(2), 261. <https://doi.org/10.1037/0033-2909.130.2.261>

- Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249-255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roth, A., Ogrin, S. & Schmitz, B (2016). Assessing self-regulated learning in higher education: a systematic literature review of self-report instruments. *Educational Assessment, Evaluation and Accountability*. 28, 225–250. <https://doi.org/10.1007/s11092-015-9229-2>
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, 4, 33-45. <https://doi.org/10.1007/s11409-008-9031-3>
- Schunk, D. H., & Zimmerman, B. J. (2012). *Motivation and self-regulated learning: Theory, research, and applications*. Routledge.
- Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educational Research Review*, 24, 116-12. <https://doi.org/10.1016/j.edurev.2018.03.004>
- Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin*, 137(3), 421–442. <https://doi.org/10.1037/a0022777>
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295-312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10, 251-296. <https://doi.org/10.1023/a:1022193728205>
- 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](https://doi.org/10.1016/S0361-476X(02)00011-5)

- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4), 1024. <https://doi.org/10.1037/0278-7393.25.4.1024>
- Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). Routledge.
- Van de Pol, J., De Bruin, A. B. H., Van Loon, M. H., & Van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology*, 56, 236-249. <https://doi.org/10.1016/j.cedpsych.2019.02.001>
- Van Gog, T., Kester, L., & Paas, F. (2011). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology*, 25(4), 584-587. <https://doi.org/10.1002/acp.1726>
- Van Merriënboer, J. J., & Sluismans, D. M. (2009). Toward a synthesis of cognitive load theory, four-component instructional design, and self-directed learning. *Educational Psychology Review*, 21, 55-66. <https://doi.org/10.1007/s10648-008-9092-5>
- Wang, T., Li, S., Tan, C., Zhang, J., & Lajoie, S. P. (2023). Cognitive load patterns affect temporal dynamics of self-regulated learning behaviors, metacognitive judgments, and learning achievements. *Computers & Education*, 207, 104924. <https://doi.org/10.1016/j.compedu.2023.104924>
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Routledge.
- Wittrock, M.C. (1989). Generative processes of comprehension. *Educational Psychologist*, 24(4), 345-376. https://doi.org/10.1207/s15326985ep2404_2

- Wong, J., Baars, M., de Koning, B. B., & Paas, F. (2021). Examining the use of prompts to facilitate self-regulated learning in Massive Open Online Courses. *Computers in Human Behavior, 115*, 106596. <https://doi.org/10.1016/j.chb.2020.106596>
- Yang, C., Zhao, W., Yuan, B., Luo, L., & Shanks, D. R. (2023). Mind the gap between comprehension and metacomprehension: Meta-analysis of metacomprehension accuracy and intervention effectiveness. *Review of Educational Research, 93*(2), 143-194. <https://doi.org/10.3102/00346543221094083>
- Zhang, W. X., Hsu, Y. S., Wang, C. Y., & Ho, Y. T. (2015). Exploring the impacts of cognitive and metacognitive prompting on students' scientific inquiry practices within an e-learning environment. *International Journal of Science Education, 37*(3), 529–553. <https://doi.org/10.1080/09500693.2014.996796>

APPENDIX A

Reading Texts

IQ

The idea that ethnic groups differ in intelligence lacks supporting evidence and is not logically plausible in light of certain facts. It is true that, on average, Asians score a couple of IQ points higher than Caucasians. In addition, Hispanics score several points lower than Caucasians, and African-Americans score several points lower than Hispanics. However, these IQ scores do not provide evidence of real differences in any kind of general mental ability.

The first problem is that there is no such thing as a single “Intelligence” that determines a person’s ability to think and make decisions in all situations. To behave intelligently means to adapt to one’s environment and situation and make appropriate choices. Different situations require different types of adaptations, and therefore different types of “Intelligence”. Athletes, artists, doctors, and scientists all use different mental skills to perform their tasks well. People who can solve complex math problems are not the same people with the most common sense or people who solve social problems. IQ tests reduce many different aspects of thinking down to a single score. These many different abilities to perform very different tasks cannot be treated as a single ability.

Standard intelligence tests require the test-taker to interpret the words and concepts used in the problems and examples on these tests. Different ethnic groups, even within the U.S., experience very different environments, live in different parts of the country, and have different languages, vocabulary, or grammar. Intelligence tests were constructed and are generally scored by Caucasians. It is no surprise that many non-Caucasians do not perform as well on such tests.

Even the differences on limited and biased IQ tests are not caused by the biological differences between races. Ethnic groups differ genetically in only very superficial ways when a single gene is involved, such as skin color, height, and other outward physical appearances. Solving complex mental problems involves many aspects of our thoughts, emotions, and behaviors that are the result of multitudes of genes interacting in complex ways. Also, even when two things are genetically identical they behave differently, because of their surrounding environment. For example, if you plant a handful of corn seeds in the fertile ground of Iowa and plant genetically identical seeds in the Arizona desert, the two groups of plants will come out very differently. Likewise, even identical twins differ in IQ, and siblings are often more different from each other than to people they are not related to.

The skills and knowledge that a child is exposed to in school play a large role in how the child performs on IQ tests. Quality of education depends upon circumstances, such as class size, teacher training, computers, and up-to-date books. In addition, parents and grandparents who obtained a better education will be better able to facilitate their child or grandchild’s education. The studies that show ethnic differences in IQ look at students who had already completed their primary education and were already affected by the quality of their education.

In addition to low education quality, all school-related performance is harmed by home-life stress and hardships of poverty. Children who are hungry, malnourished, and concerned about family conflict will lack

the focus and motivation needed to perform well on such tests. On average, most minority groups are lower in economic and social status compared to whites.

The gap between Caucasians and minorities on intellectual tests such as SAT and ACT has been getting smaller over the past decades. Also, the average IQ score for all people today is several points higher than it was 50 years ago. Obviously, there is virtually no biological difference between children of today and their grandparents. Changes in scores over time are the result of changes in the economic status of minorities.

A more historical problem is that the American culture has lead minority groups to believe they are inferior and not as intelligent, for many generations. Low self-confidence will impact any kind of test that attempts to evaluate and compare people. Regardless of current economic success or education, most African-Americans share this cultural history which subtly impacts their confidence in their own abilities.

MONETARY POLICY

The U.S. is the largest economy in the world. Therefore, the U.S. dollar is considered a stable value. Many factors affect the strength of the dollar relative other currencies, including the trade deficit or surplus, the size of the Government deficit, interest rates, and the strength of the U.S. economy. The strength of the dollar is also affected by the monetary policy imposed by the Federal Reserve System.

The Fed, as it is called, is the central bank of the U.S. The Fed's duties include conducting the nation's monetary policy by influencing money and credit conditions in the economy in pursuit of full employment, stable prices, and promoting the stability of the financial system. The Fed conducts monetary policy using three major tools. It buys and sells U.S. Treasury and federal agency securities in the open market; it sets the discount rate, which is the interest rate that banks pay the Fed to borrow money; and it sets reserve requirements, which is the amount of funds that banks must hold in reserve against deposits made by their customers. Monetary policy can affect short-term interest rates, foreign exchange rates, long-term interest rates, the amount of money and credit, and, ultimately, a range of economic variables, including employment, output, and prices of goods and services.

Monetary policy works by affecting the amount of money circulating in the economy. The Fed can change the amount of money that banks are holding in reserves by buying or selling existing U.S. Treasury bonds. The Fed sells bonds, which decreases banks' reserves and their ability to make loans. As banks lend less and the money supply decreases, interest rates increase. The Fed buys bonds, which increase banks' reserves and their ability to make loans. As banks lend more and the money supply increases, interest rates decrease.

Lower interest rates mean that consumers pay less when they charge purchases. They may be more willing to spend. They may even buy expensive goods, like cars and refrigerators, to take advantage of lower interest rates. As the demand for more goods increases, either businesses will increase production to satisfy the demand or prices of goods will increase.

Lower interest rates may encourage businesses to expand to meet the increasing consumer demand. They may run extra shifts or build new factories. This may create new jobs. As workers, who were

previously unemployed, return to the workforce, they will eventually spend their paychecks. This too will increase the demand for goods. Again, either businesses will increase production or prices of goods will increase.

Sometimes consumer spending is so great that production can't keep up with demand. The excessive demand for goods can lead to inflation. Inflation can also occur as a result of increasing the amount of money circulating in the economy. Inflation means dollars are worth less. The Fed will try to keep inflation in check.

Inflation may undermine the strength of the economy. Inflation increases the difficulty of forecasting prices and costs of doing business, so it discourages businesses from planning and investing. People also may be uncertain and reluctant to spend. Both of these factors could reduce the long-term level of economic growth. Inflation also increases the cost of carrying out transactions. Inflation in U.S. increases cost of U.S. goods; therefore, imports increase and exports decrease.

APPENDIX B

Comprehension-Test Items

IQ

1. Which of the following are true of scores on IQ tests?

A. Asians score higher than Hispanics

B. Hispanics score higher than Caucasians

C. African-Americans score higher than Hispanics

D. All groups score the same

2. Which of the following is implied by the text?

A. Grandparents will tend to have higher IQ scores than their grandchildren

B. Grandparents will tend to have lower IQ scores than their grandchildren

C. Grandparents will have the same IQ scores as their grandchildren

D. IQ tests have only existed for 20 years, so no one knows

3. Which would be true if African-Americans and Caucasians were compared in the first grade rather than in high school?

A. Their IQ scores would be more different

B. Their IQ scores would be more similar

C. African-Americans would have higher IQs than Caucasians

D. Both A and C

Which of the following is NOT true, according to the text?

A. There is no such thing as a single general intelligence

B. IQ tests are culturally biased

C. Only African-Americans who are poor will score lower than Caucasians

D. None of the above, they are all true according to the text

Why can't differences between ethnic groups in solving complex problems be caused by biological differences?

A. African-Americans have suffered historical injustices

B. People who live in poverty receive a worse education

C. IQ tests were created by Caucasians

D. Ethnic groups only differ in ways that involve a single gene

Monetary Policy

1. If interest rates are lowered, consumers are more likely to

A. buy more cars

B. buy more food

C. save more

D. travel less

2. Which of the following is NOT a likely result of lower interest rates?

A. prices of goods will decrease

- B. consumers are willing to spend more
- C. consumers will buy more expensive goods
- D. businesses will decrease production

3. Which is likely to occur when the Fed increases the reserve requirement?

- A. consumer spending will increase

B. interest rates will increase

- C. local banks will increase lending
- D. the economy will grow

4. What might the Fed do if it wants to affect the economy in a way that is similar to that of lowering income taxes?

- A. decrease loans to consumers and businesses

B. decrease the reserve requirement

- C. increase the discount rate
- D. decrease the money supply

5. Unemployment will tend to decrease when

A. interest rates decrease

- B. consumer demand decreases
- C. business investment decreases
- D. the money supply decreases

APPENDIX C

Cognitive Load Questionnaire by Krieglstein et al. (2023)

Cognitive load type	Item
<i>ICL</i>	
ICL1 (item 1)	The learning content was difficult to understand
ICL2 (item 2)	The explanations of the learning content were difficult to understand
ICL3 (item 3)	The learning contents were complex
ICL4 (item 4)	The learning content included much complex information
ICL5 (item 5)	Without prior knowledge, the information was not understandable
<i>ECL</i>	
ECL1 (item 6)	It was difficult to gain an overview of the structure of the learning material
ECL2 (item 7)	The design of the learning material made it difficult to recognise links between individual information units
ECL3 (item 8)	The design of the learning material was inconvenient
ECL4 (item 9)	The design of the learning material made it difficult to find relevant information quickly
ECL5 (item 10)	Because of the design of the learning material, I had the impression that I could not concentrate on the learning content
<i>GCL</i>	
GCL1 (item 11)	I actively reflected upon the learning content
GCL2 (item 12)	I made an effort to understand the learning content
GCL3 (item 13)	I achieved a comprehensive understanding of the learning content
GCL4 (item 14)	I was able to expand my prior knowledge with the learning content
GCL5 (item 15)	I can apply the knowledge that I acquired through the learning material quickly and accurately

Note. Items 2, 8, and 9 were deleted because they were not applicable. Item 3 was deleted because of its similarity with item 3.

APPENDIX D

Test Instructions for Each Condition

Condition	Condition-Specific Instructions
1: Self-summary only	<p>Please write a summary of the text about (title text) that you have just read. You have a maximum of 7 minutes to write the summary.</p> <p>Reread your summary to prepare yourself for the comprehension test. You have a maximum of 3 minutes.</p>
2: Self-summary + prompts	<p>Please write a summary of the text about (title text) that you have just read. You have a maximum of 7 minutes to write the summary.</p> <p>Reread your summary to prepare yourself for the comprehension test. Think about the following questions while reading: How well would your summary help others to gain an understanding of the original text? How well are the key concepts of the text summarized? What would you add to the summary? You have a maximum of 3 minutes.</p>
3: Gen-AI summary only	<p>You can ask open AI chatbots, like ChatGPT, to write a summary. The prompts you give to an AI chatbot determine the quality of the output. Please write a prompt for an AI chatbot to get a qualitative summary about the text on (title text). You have a maximum of 3 minutes.</p> <p>Read the summary to prepare yourself for the comprehension test. You have a maximum of 7 minutes.</p>
4: Gen-AI summary + prompts	<p>You can ask open AI chatbots, like ChatGPT, to write a summary. The prompts you give to an AI chatbot determine the quality of the output. Please write a prompt for an AI chatbot to get a qualitative summary about the text on (title text). You have a maximum of 3 minutes.</p> <p>Think about the following questions while reading: How well would this AI generated summary help others to gain an understanding of the original text? How well are the key concepts of the text summarized? What would you add to the summary? You have a maximum of 7 minutes.</p>

APPENDIX E

Assumptions Details

Analysis	Normality	Homoscedasticity	Linearity	
	Skewness + kurtosis	Kruskal- Wallis	Levene's test	
ANOVA 1	Normality supported	Not significant	.702	
ANOVA 2	Normality supported	Not significant	.373	
MANOVA	Normality supported	Not significant	For ICL: .137 For ECL: .476 For GCL: .523	Assumption was met for each dependent variable

Acknowledgement

During this thesis project, generative AI system ChatGPT-3.5 was used

(<https://chatgpt.com/auth/login>). To provide insight in the way generative AI was used, this

acknowledgement contains examples of prompts that were typically used.

Example Prompt #1: Brainstorming Different Perspectives

how to measure reading comprehension?



Measuring reading comprehension involves assessing an individual's ability to understand, interpret, and draw meaning from written text. Various methods and tools can be employed for this purpose. Here are some common approaches to measure reading comprehension:

1. Standardized Reading Tests:

- Use established standardized reading comprehension tests that are designed to assess a range of comprehension skills. These tests often include passages followed by questions that evaluate different aspects of comprehension, such as main idea identification, inference-making, and vocabulary understanding.

2. Cloze Tests:

- Cloze tests involve removing certain words from a passage and asking individuals to fill in the blanks with appropriate words. This assesses the reader's ability to use context clues and understand the overall meaning of the text.

3. Retelling or Summary Tasks:

- Ask individuals to retell or summarize a passage in their own words. This method assesses not only comprehension but also the ability to synthesize information and convey it coherently.



The output of such brainstorming prompts was not used to generate text to include in the thesis project, but merely to introduce myself with a topic. This output for example was used to get an overview of different ways to measure reading comprehension. After this, Google Scholar would typically be used to search for reliable in-depth literature about a specific topic.

Example Prompt #2: To Provide Basic Understanding of Concepts

how to perform a power analysis for an intervention study



Performing a power analysis is a crucial step in the design of an intervention study. Power analysis helps researchers determine the sample size needed to detect a statistically significant effect, given a certain level of power and significance level. Here's a general guide on how to perform a power analysis for an intervention study:

Steps for Power Analysis:

1. Define Study Parameters:

- **Effect Size (ES):** This is the magnitude of the difference you expect to observe. It's crucial to estimate this based on prior research or pilot studies.
- **Significance Level (α):** Typically set at 0.05, indicating a 5% chance of committing a Type I error (rejecting a true null hypothesis).
- **Power ($1 - \beta$):** Desired level of power, often set at 0.80 or 0.90, indicating the probability of correctly rejecting a false null hypothesis (avoiding a Type II error).

2. Choose Statistical Test:

- Depending on the nature of your intervention and the data type, choose an appropriate statistical test (e.g., t-test, ANOVA, regression).

When I struggled with the understanding of a topic or part of statistics, I sometimes consulted Gen-AI to provide a general description of the concept. Gen-AI was used for this purpose, because the advantage of it is that you can also continue to ask about specific parts you don't understand. The output of such prompts was used to deepen my understanding of concepts, so I could dive deeper into relevant scientific literature about those concepts, and eventually include this scientific literature in the thesis.

Example Prompt #3: To Provide a Plan or Instructions

I conducted an online experiment in qualtrics with 4 conditions. I downloaded the raw data file, i want to clean my data and prepare the data for analyzing the data in spss. How do i start?



Cleaning and preparing data for analysis in SPSS involves several steps. Here's a general guide to get you started:

1. Import the Data into SPSS:

- Open SPSS and create a new data file or open an existing one.
- Import your raw data file into SPSS. Go to File > Import Data > Excel (or appropriate file format) and follow the prompts to import your data.

2. Review the Data:

- Take a preliminary look at your data to understand its structure and variables.
- Check for any missing values, outliers, or errors in the data.

3. Variable Identification and Renaming:

- Rename variables if necessary to make them more descriptive and easier to understand.
- Identify the type of each variable (e.g., categorical, numerical) and note any transformations that may be needed.

This example prompt shows how I used Gen-AI to provide instructions or a plan to get started with certain tasks. Output of Gen-AI was used to support activities related to this thesis project. In this case, I wasn't quite sure how to begin the process of data cleaning. Gen-AI supported me by providing basic instructions and outlining the main steps of data cleaning. The output of these prompts was used to perform tasks relevant to the thesis.