# The Effects of Working Memory during a Search and Sensemaking Task

Bogeum Choi
School of Information and Library Science
University of North Carolina at Chapel Hill
North Carolina, USA
choiboge@unc.edu

Jaime Arguello
School of Information and Library Science
University of North Carolina at Chapel Hill
North Carolina, USA
jarguello@unc.edu

## Abstract

Working memory (WM) is involved in high-level cognitive tasks such as comprehension, reasoning, and learning. Search and sensemaking (SSM) is no exception—a wide range of (meta)cognitive activities are involved in the process of making sense of a complex topic by gathering information. Prior studies have found that WM can influence search behaviors, perceptions, and outcomes. However, little work has been done to gain insights into *how* WM might affect the SSM process. We report on a lab study ($N = 44$) in which participants were binned into low- and high-WM groups. During the study, participants were asked to learn about a complex and multifaceted topic by gathering information using a web search engine and taking notes. After the search session, participants were asked to produce a summary of everything they learned. The study investigated four research questions. RQ1 and RQ2 investigate the effects of WM on post-task perceptions and search behaviors. RQ3 investigates the effects of WM on the extent to which participants engaged in specific search, sensemaking, and cognitive activities. To address RQ3, the study used a think-aloud protocol. Search sessions (i.e., recorded actions and think-aloud comments) were then analyzed using qualitative techniques. Finally, RQ4 investigates the effects of WM on learning outcomes. Our RQ4 results found that high-WM participants had better learning outcomes. Our RQ2 and RQ3 results point to possible reasons why. Despite differences for RQ2-RQ4, there were no differences in post-task perceptions (RQ1).

## CCS Concepts

• **Information systems → Users and interactive retrieval**.

## Keywords

Working memory, search and sensemaking, search as learning

**ACM Reference Format:**
Bogeum Choi and Jaime Arguello. 2025. The Effects of Working Memory during a Search and Sensemaking Task. In *2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25), March 24–28, 2025, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3698204.3716449

## 1 Introduction

Our research in this paper lies at the intersection of two areas of research: (1) the effects of working memory and (2) search and sensemaking.

Working memory (WM) refers to an individual's ability to maintain and manipulate information in memory, when it is no longer perceptually present. Outside of interactive information retrieval (IIR), studies have found that WM plays an important role in many higher-level cognitive tasks, such as reading [9, 15, 19], comprehension [33, 51], logical reasoning [16], and attentional control [6]. IIR studies have also examined the effects of WM on search behaviors, and outcomes. For example, studies have found that high-WM searchers tend to exhibit greater effort [12, 23] and achieve better outcomes [35, 43], especially during complex search tasks.

Sensemaking involves making sense of experiences, situations, or topics. Different models have been proposed to characterize sensemaking processes [25–27, 40]. At a high level, sensemaking involves: (1) developing a representation of a topic; (2) encoding information into the representation; and (3) modifying the representation when information does not fit into it or when aspects of the representation become less useful. Zhang and Soergel [52] proposed a comprehensive model of sensemaking, drawing on literature on task-based information seeking and acknowledging the importance of information search activities in sensemaking processes. In our study, we use the term "search and sensemaking" (SSM) to highlight both aspects of sensemaking. This delineation can be particularly helpful when the primary task involves learning about a complex topic through searching for information online. To our knowledge, prior studies have not investigated the role of WM in the SSM process.

We report on a lab study that proceeded in two phases. During Phase 1, 60 participants completed a working memory task to measure their working memory capacity. Forty-four participants with the most extreme scores were invited to participate in Phase 2 and were binned into low- and high-WM groups (22 per group). During Phase 2, participants completed a complex SSM task by gathering information using a web search engine and taking notes. Specifically, participants were asked to learn about the "gut-brain connection", a complex topic that involves many facets and perspectives. To gain deeper insights into the role of WM during the SSM process, the study used a think-aloud protocol. Think-aloud comments and recorded screen activities (i.e., search, reading, and note-taking activities) were analyzed using qualitative techniques. After completing the SSM task, participants were asked to produce a written summary of everything they learned.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*CHIIR '25, March 24–28, 2025, Melbourne, VIC, Australia*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1290-6/25/03
https://doi.org/10.1145/3698204.3716449

The study investigated four research questions, which compared dependent variables between low- and high-WM groups:

- **RQ1:** What are the effects of working memory on participants' post-task perceptions?
- **RQ2:** What are the effects of working memory on participants' search behaviors?
- **RQ3:** What are the effects of working memory on the extent to which participants engaged in specific search, sensemaking, and other cognitive activities?
- **RQ4:** What are the effects of working memory on participants' learning outcomes?

Our results found several important trends. First, high-WM participants had greater knowledge gains. Second, high-WM participants exhibited different search behaviors—they had fewer abandoned queries, spent less time searching, and spent more time reading pages and taking notes. Third, low- and high-WM participants had different levels of engagement in specific SSM and cognitive activities. For example, high-WM participants were more active in evaluating the relevance of information based on their goals and monitoring their progress.

Our study enhances our understanding of the cognitive mechanisms involved in the SSM process through the lens of WM. Our findings can inform the design of search interfaces that support individuals with diverse cognitive profiles by encouraging and supporting specific SSM and cognitive activities, which in turn may improve learning outcomes.

# 2 Related Work

## 2.1 Working Memory

Working memory (WM) refers to an individual's ability to maintain and manipulate information in memory, even when the information is not perceptually present. WM involves holding relevant information in memory while processing or working with the *same* information (e.g., solving a math problem) or *different* information (e.g., recognizing connections between newly and previously encountered information). In this respect, WM is critical for making sense of anything that unfolds over time. WM is considered a core executive function (EF), and overlaps conceptually with other EFs (e.g., inhibition, updating, and shifting) in that they all relate to attentional control [10, 41]. That is, WM is involved in an individual's ability to direct their attention to relevant information while suppressing distractions, prioritize information most pertinent to the current goals, and redirect their attention as needed. Such cognitive processes are characterized as top-down, goal-driven, voluntary, and highly effortful [18].

A substantial body of literature has investigated the effects of WM on various cognitive tasks. Several studies have found correlations between WM capacity (measured using a reading span task) and measures related to reading comprehension, such as remembering facts, detecting semantic inconsistencies, and resolving pronouns [9, 15, 19]. Other studies have found that demands on WM can impact an individual's ability to integrate information from different parts of a text [33, 51]. Additionally, De Neys et al. [16] found that WM plays a central role in logical reasoning (e.g., by facilitating the retrieval of counterexamples from long-term memory). In terms of learning, Sanchez and Wiley [42] found that high-WM participants were less susceptible to the influence of seemingly interesting but irrelevant information. Banas and Sanchez [6] found that high-WM participants outperformed low-WM participants in recognizing and retaining relationships between interconnected concepts when learning in a wiki-like environment, where information about different concepts is dispersed across multiple documents. Overall, the underpinning role of WM in high-level cognitive tasks is by sustaining attention, resisting distractions, and integrating distinct pieces of information over time [24].

## 2.2 Working Memory & Search

Studies have investigated the effects of WM on search behaviors and outcomes. Sharit et al. [43] had participants aged 60-85 complete search tasks of varying complexity. WM was a significant predictor of task performance for complex tasks. Gwizdka [22] had participants complete simple tasks that required finding a fact and complex tasks that required finding a set items matching certain criteria. Results found an interesting interaction effect. During simple tasks, high-WM participants were more efficient—issued fewer queries, opened fewer documents, and took less time. However, during complex tasks, high-WM participants exhibited more effort, possibly because low-WM participants satisficed. Gwizdka [23] had participants complete search tasks in the medical domain and take notes. Based on eye-tracking data, high-WM participants spent more time reading pages, particularly toward the end of the search session. Additionally, high-WM covered a wider range of topics in their notes.

Choi et al. [12] conducted a study in which participants completed decision-making tasks of varying complexity. Tasks involved comparing a specific set of alternatives along a set of dimensions. Simple tasks involved two alternatives and two dimensions, and complex tasks involved four alternatives and four dimensions. During each task, participants were asked to choose an alternative and write a justification. High-WM participants exhibited greater effort—issued more queries, clicked on more results, interacted at a faster pace, and wrote longer justifications. The effect on the justification length was more pronounced for the complex task. WM did not influence participants' perceptions of workload or satisfaction.

Arguello and Choi [3] had participants interact with two aggregated search interfaces that combined results from different verticals (web, images, news, shopping, video). One interface had a fixed layout and organized results by vertical. The other was more dynamic and visually cluttered. Low-WM participants reported similar levels of mental demand with both interfaces. However, high-WM participants reported lower mental demand with the fixed-layout interface, possibly because they internalized the layout and directed their attention to relevant verticals.

Choi et al. [11] conducted a study in which participants interacted with two tools to save information. An experimental tool enabled participants to create "boxes" representing topics and to drag-and-drop passages from a page into specific boxes. A baseline tool only allowed participants to drag-and-drop passages into a list that could not be (re-)organized. In terms of search behaviors, regardless of the tool, high-WM participants issued queries targeting a broader range of topics. With the experimental tool, high-WM participants had *more* abandoned queries, suggesting that they were more deliberate and selective with the information they saved. In terms of perceptions, with the baseline tool, low-WM participants reported greater difficulty in deciding when they had enough information to complete the task, possibly because they could not organize the information by topic.

Studies have also considered the effects of WM on learning during search, with mixed results. Bhattacharya and Gwizdka [7] asked participants to learn about a medical subject. Participants were asked to list topically relevant terms before and after searching, and knowledge gains were measured based on the increase (and complexity) of terms listed. WM had no significant effects on knowledge gains. In contrast, Pardi et al. [35] asked participants to learn about a scientific subject. Learning was measured by counting the number of relevant concepts included in essays written after the search session. WM had a positive and significant correlation with the number of concepts included.

Research has also studied the search behaviors of dyslexic users. While dyslexia is a heterogeneous condition, it is associated with deficits in WM capacity [44, 46]. MacFarlane et al. [29] observed that non-dyslexic participants viewed more documents and judged a greater percentage as non-relevant, suggesting they were more selective. MacFarlane et al. [30] found that dyslexic participants were more likely to revisit previously viewed information. Morris et al. [32] studied the challenges of dyslexic searchers through interviews, a survey, and a lab study. Interview and survey responses found that dyslexic searchers prefer content with less clutter, shorter sentences, and more structural elements such as lists and tables. In the lab study, participants were asked to rate webpages along dimensions of understandability. Ratings from dyslexic participants were less positive. Fourney et al. [21] compared relevance judgments from dyslexic and non-dyslexic participants, and found that judgments from non-dyslexic participants were more bimodal (i.e., either highly relevant or highly non-relevant).

### 2.3 Search and Sensemaking

The current study builds on the sensemaking literature as developed in the field of human-computer interaction (HCI). Russell et al. [40] led the adoption of sensemaking theory from the fields of communication and library and information science [17] into the HCI field and developed a conceptual model of sensemaking. In their case study, they observed that participants engaged in an iterative process of generating and revising schemas using a knowledge representation tool as they tried to make sense of complex information. They identified three main phases in the sensemaking process: schema generation, data coverage, and representational shift. During the schema generation phase, people create representations to capture salient aspects of the task. During the data coverage phase, people search for task-relevant information and encode it into the existing representations. During the representational shift phase, people adjust the representations when they encounter information that does not fit or when they decide that parts of a representation are not useful.

Klein et al. developed the "data/frame theory" of sensemaking [25–27]. A frame refers to an individual's mental model of reality and data refers to observations that inform or challenge a frame. Frames determine which data are relevant and are modified when they are inadequate in explaining new data. Frames can take various forms, such as narratives or diagrams. Klein et al. identified six sensemaking activities: (1) seeking, (2) elaborating, (3) questioning, (4) preserving, (5) comparing, and (6) re-framing. Frame *seeking* involves searching for and selecting a relevant frame to generate a plausible explanation of a situation. *Elaborating* a frame involves gathering information to refine or expand the frame. *Questioning* occurs when the sensemaker identifies inconsistencies in data, and *preserving* occurs when the sensemaker reinterprets the data to maintain the integrity of the frame. Questioning a frame may result in *comparing* with other frames and *re-framing*—altering or replacing the frame to better fit the data. The concept of "frame" in their theory is comparable to the concept of "schema" in psychology.

Zhang and her colleagues sought to extend existing sensemaking models by integrating theories of cognition and learning. They drew on learning theories related to schema [2, 5, 38, 49] to better characterize the conceptual changes that occur during sensemaking. Additionally, they incorporated the dual-process model of reasoning [4, 45] to describe the cognitive mechanisms involved in sensemaking using terms like top-down (i.e., goal-driven) and bottom-up (i.e., data-driven) processes. Based on an extensive literature review, the authors developed a "comprehensive model of the cognitive process and mechanisms of individual sensemaking"[52]. This model was validated through a series of user studies [52–54]. Zhang's work is particularly relevant to our research for two reasons. First, their user studies were conducted in settings similar to ours. In their studies, participants in a lab setting completed complex information-seeking tasks and took notes. Search sessions were recorded to determine situations where sensemaking occurs. Second, their model describes search and sensemaking activities at a concrete level, facilitating the direct observation of these activities. In our work, we referred to their model [54] and the coding scheme developed in Zhang and Soergel [53] to deduce observable search and sensemaking activities in our data.

## 3 Methods

### 3.1 Study Overview

To investigate RQ1-RQ4, we conducted an in-person lab study that proceeded in two phases. Phase 1 involved 60 participants and Phase 2 involved 44 participants recruited from Phase 1. Phase 2 was the main phase of the study. Phase 1 helped us recruit participants for Phase 2 with different levels of WM capacity. Participants were recruited through an opt-in mailing list of employees at our university. The study was approved by our university's Institutional Review Board (IRB).

Similar to prior work [3, 7, 11, 12, 22, 23], participants in Phase 2 were binned into a low-WM group and a high-WM group (22 participants per group). All our analyses for RQ1-RQ4 focus on differences between groups. We wanted these two groups to have significantly different levels of WM capacity. To this end, during Phase 1, 60 participants completed a working memory task called the Operation Span (OSPAN) task (Section 3.2). The possibility of not being invited to the second study phase was outlined in our consent form and was verbally communicated by the study moderator. All participants agreed to this condition. The Phase 1 session took about 10 minutes and participants were paid US$10.

Forty-four participants from Phase 1 participated in Phase 2. Participants were recruited based on their OSPAN scores. Our initial plan was to recruit the highest and lowest scoring participants from Phase 1 as the high- and low-WM groups, respectively. However, due to a few dropouts, the 22 participants in the low-WM group originated from the lowest 23 scores and the 22 participants in the high-WM group originated from the highest 25 scores.

Phase 2 involved a diverse participant sample. Thirty-four participants identified as female, seven as male, and three as non-binary. Their ages ranged from 18 to 65 ($M$ = 23.98, $S.D.$ = 11.26). Twelve participants were under the age of 20, 23 were in their 20s, 1 in their 30s, 2 in their 40s, and 2 in their 60s. Participants included 32 student employees and 12 non-student employees. In terms of highest educational degree attained, 27 participants reported having completed a high school degree[1], 1 an associate degree, 6 a bachelor's degree, and 10 a graduate degree.

## 3.2 Working Memory Task

To measure participants' WM capacity, we used the Operation Span (OSPAN) task [13]. During the OSPAN task, participants complete a series of trials. During each trial, participants are presented with a visual sequence of 3-7 words one at a time. After each sequence, participants are asked to recall the words in the order they were displayed from a 3 × 3 grid of words. Additionally, before each word in a sequence, participants are presented with a simple math problem and proposed solution (e.g., (8 x 2) - 8 = 9?). Participants must indicate whether the proposed solution is correct or not. The purpose of these math problems is to prevent participants from easily rehearsing the word sequence by engaging them in a secondary information processing task. To ensure participants are sufficiently engaged with the math problems, only OSPAN scores with greater than 80% accuracy on the math problems are considered valid. All of our participants met this requirement. A participant's final score is equal to the sum of lengths for those sequences perfectly recalled.

We obtained a wide range of OSPAN scores from Phase 1. OSPAN scores for the 44 participants who participated in Phase 2 ranged from 12-51 ($M$ = 35.5). Participants in the low-WM group had scores 12-31 ($M$ = 24.5) and those in the high-WM group had scores 38-51 ($M$ = 43).

## 3.3 Phase 2 Study Protocol

Phase 2 of the study involved the following sequence of steps. First, participants completed a short demographics questionnaire. Then, they watched a video that provided an overview of the study. To investigate participants' search and sensemaking activities, the study used a think-aloud protocol. The video instructed participants to verbalize their thoughts as they worked on the task. Next, participants were given a description of the main learning-oriented search task (Section 3.4) and were asked to read it aloud. Following this, participants completed a pre-task questionnaire about their perceptions of the task (Section 3.5). Then, participants completed the main search task. Participants were free to use any web search engine to gather information and were provided with a Google Doc to take notes. The Google Doc also included the task description at the top. We used a Chrome Extension to log participants' search behaviors. The main search task (i.e., screen activities and think-aloud comments) was recorded. Participants were given 30 minutes to complete the main search task and were notified when they had 5 minutes remaining. During the main task, the moderator prompted participants to "please keep taking" if they were silent for too long. After the main search task, participants were asked to write an essay outlining "everything you learned during the task". Participants could not look at their notes while writing their essay and were not given a time limit. Participants took 3-15 minutes to write their essays. Finally, participants completed a post-task questionnaire about their experiences (Section 3.5). The Phase 2 session took about 60 minutes and participants were paid US$30.

## 3.4 Search Task

Our study investigates the role of WM during a *complex* search and sensemaking task. To this end, we designed a learning-oriented search task with the following characteristics, inspired by characteristics of exploratory search tasks as defined by Wildemuth and Freund [50]: (1) the goal is open-ended, emphasizing learning; (2) the topic is multifaceted; (3) the task is likely to require consulting multiple sources; (4) the task involves uncertainty; and (5) the task involves cognitive processes at the level of *analyze* or higher [1].

The task description was outlined as follows:

**Scenario**: You recently attended a guest lecture on the emerging field of the gut-brain connection. The speaker explained the intricate relationship between the gut microbiome and one's physical and mental health. After the lecture, you realize there is much more to learn about the connection between your digestive system and your overall well-being.

**Objective**: To the best of your ability, try to find out and learn about the topic of the gut microbiome and an individual's physical and mental health. Potential sub-topics you can explore include but are not limited to: What is the notion of "gut-brain connection"? Through what mechanisms do gut microbiota influence one's physical and mental health? What factors can influence gut microbiota? What are some science-backed ways to improve your gut health?

In addition to having the five characteristics outlined above, we chose the topic of the "gut-brain connection" because there is a substantial amount of information on this topic as it has gained popularity in the health and well-being sector, which is relevant to many individuals. And yet, it remains not too well-known to the general public and it involves topics of debate (i.e., uncertainty).

## 3.5 Questionnaires

Participants completed a pre- and post-task questionnaire before the search task and after summarizing what they learned. In both questionnaires, participants responded to agreement statements on a 7-point scale ranging from strongly disagree (1) to strongly agree (7). The full text of both questionnaires is available online.

**Pre-task Questionnaire:** The pre-task questionnaire asked about: (1) interest (2 items), (2) prior knowledge (2 items), (3) motivation (2 items), (4) expected difficulty (3 items), and (5) *a priori* determinability (5 items), which measures the extent to which aspects of the task (e.g., requirements, goals, strategies) are known

[1] Many of these participants were undergraduates employed part-time at our university and working toward their bachelor's degree.

in advance [8]. Except for the two motivation items, all groups of items had high internal consistency (Cronbach's $\alpha > .80$). Therefore, responses were averaged to form composite measures for interest, prior knowledge, difficulty, and determinability. The two motivation items had lower internal consistency ($\alpha = .60$) and were therefore not combined.

**Post-task Questionnaire:** The post-task questionnaire was organized in three parts. First, participants were asked about the level of satisfaction with their performance (4 items). Second, participants were asked about the level of cognitive load experienced. Cognitive Load Theory (CLT) [34] argues that mental effort during a learning task can be attributed to different sources. Intrinsic load is attributed to inherent characteristics of the material to be learned (e.g., the complexity of the topic); extraneous load is attributed to the learning environment (e.g., distracting ads); and germane load is attributed to the activities the learner engaged in during the learning process (e.g., synthesizing information). The second part of the post-task questionnaire asked about intrinsic load (5 items), extraneous load (3 items), and germane load (5 items). Finally, the last part of the post-task questionnaire asked about the extent to which the participant engaged in different cognitive activities: (1) planning (5 items), (2) monitoring their progress and comprehension (4 items), (3) organizing information (3 items), and (4) evaluating/adapting their approach to the task (5 items).

The groups of items for satisfaction, extraneous load, germane load, planning, and organizing had high internal consistency (Cronbach's $\alpha > 0.70$). Therefore, responses were averaged to form composite measures for these constructs. The items for intrinsic load, monitoring, and evaluating/adapting had lower internal consistency (Cronbach's $\alpha < 0.70$) and were therefore not combined.

## 3.6 Search Behaviors

To investigate RQ2, we used a Chrome extension to log search events and computed the following behavioral measures:

(1) Number of queries.
(2) Number of abandoned queries (i.e., no clicks on the SERP).
(3) Average number of words per query.
(4) Number of distinct URLs visited.
(5) Completion time (minutes).
(6) Time on the SERP (minutes).
(7) Time on pages and notes (minutes).
(8) Number of queries not issued by others.
(9) Number of query terms not used by others.
(10) Number of URLs not visited by others.

Measure #6 captures the amount of time participants spent searching. Measure #7 captures the amount of time participants spent reading and taking notes. Participants often positioned the landing page and their notes side-by-side and iterated between reading and taking notes. Thus, we decided to use one measure to capture time spent on both activities. Measures #8-10 capture the extent to which participants adopted search strategies that were different from other participants.

## 3.7 Qualitative Analysis of Search Sessions

To address RQ3, we conducted a qualitative analysis of search sessions using the recorded videos, which included screen activities and think-aloud comments. Qualitative codes were associated with three categories: (1) search activities, (2) sensemaking activities, (3) cognitive activities. Qualitative codes were developed by author A1 using a combination of deductive and inductive coding [20, 31]. The deductive aspect involved including codes developed by Zhang et al. [55] to characterize search and sensemaking activities. The inductive aspect involved introducing new codes based on observations. New codes were related to cognitive activities associated with working memory (e.g., participants making connections between information encountered at different points in time). Next, we describe the coding process and then describe our code definitions.

Our qualitative analysis of search sessions involved the following steps. First, author A1 converted each search session into a sequence of codable units. Each session was represented as a spreadsheet with three columns: (1) timestamp, (2) screen activity, and (3) think-aloud comment (if any). Each row represented a codable unit. Not every codable unit was assigned a code. For example, a codable unit might describe the participant spending a long time reading an article without making any comments. No code applied to this activity. Second, after A1 transcribed every session and developed an initial coding guide, authors A1 and A2 independently coded sessions from two participants. Coding involved watching the video of the search session and adding codes to codable units on the corresponding spreadsheet. After this, A1 and A2 met to discuss the coding guide, refine code definitions, and establish detailed rules for when to apply each code. Finally, A1 and A2 independently coded sessions from 5 new participants (11% of the data). Intercoder reliability was measured for each code independently. In terms of Cohen's $\kappa$, agreement was $\kappa \geq .885$ across all codes, which is considered "almost perfect" [28]. At this point, A1 re-coded the initial 2 participants and coded the remaining 37.

Tables 1-3 describe our codes associated with search, sensemaking, and cognitive activities, respectively. Several of our codes might benefit from additional clarification. In terms of search activities, we observed two types of querying behavior. Participants issued queries that were heavily influenced by topics in the task description (structure-driven) as well as topics encountered during the search session (data-driven). Prior work suggests that sensemakers engage in both top-down and bottom-up mechanisms [55]. Structure-driven querying can be considered as a top-down mechanism, while data-driven querying can be considered as a bottom-up mechanism.

In terms of sensemaking activities, instantiation and accretion both involve adding new information into the current knowledge structure in the participant's notes. However, instantiation involves elaborating on information that is *already* in the notes. Tuning and re-structuring both involve modifying the current knowledge structure in the notes. However, tuning is a gradual change (e.g., renaming a heading) and re-structuring is a more radical change (e.g., splitting a heading into two).

In terms of cognitive activities, active maintenance refers to instances where it is clear that the participant maintained information active in memory. This may be evidenced by the participant relating new information with previously found information or noticing that two sources corroborate or contradict each other. Our coding guide is also available online.

**Table 1: Search Activity Codes**

| Code | Definition | Example |
|---|---|---|
| Structure-driven Query Formulation | The query is guided by (sub)topics mentioned in the task description. | After reading the task description, the participant issues the query "the gut-brain connection". |
| Data-driven Query Formulation | The query is guided by topics encountered during the search session. | While reading an article, the participant issues the query "the blood-brain barrier". |
| Being Selective with Source | The participant judges the usefulness of a source based on factors such as credibility, reliability, or intended audience. Alternatively, the participant issues a query for a specific source (e.g., Wikipedia). | After clicking a search result, the participant says: "This is too technical. I need something for lay people." |
| Searching for Structure | The participant seeks information to get a high-level overview of the topic. | The participant scans the table of contents of a book on Amazon to understand how the topic is structured. |

**Table 2: Sensemaking Activity Codes**

| Code | Definition | Example |
|---|---|---|
| Gap Identification | The participant identifies a gap in their knowledge. | "I have no idea what the endocrine system is." |
| Building | The participant creates a structure in their notes. This can take the form of adding headings, making a list of topics, or starting a new paragraph. | The participant adds "how to improve gut health" as a heading in their notes. |
| Semantic Fit | The participant evaluates whether and how new information meets their goals and/or fits into the current structure in their notes. | While reading an article, the participant says: "Okay, here is information on how to improve your gut health." |
| Accretion | The participant adds relevant information into their notes. | Under the heading "Impacts on physical health", the participant adds a note "An abnormal gut microbiome can affect the immune system". |
| Instantiation | The participants adds more detailed information to their notes. This can take the form of adding lower-level bullet points, adding examples to existing ideas, or elaborating on previously recorded statements. | After noting that fiber-rich foods can improve gut health, the participant mentions that fruits and vegetable are examples of fiber-rich foods. |
| Tuning | The participant modifies the existing structure in their notes to reflect a more refined understanding of a topic. This can take the form of editing a heading or emphasizing the importance of something by bolding, underlining, resizing, or reordering. | The participant encounters the acronym GBA, which stands for the gut-brain axis. After adding this concept to a bulleted list of "mechanisms", she rearranged the list to position GBA above other subtopics (e.g., neurotransmitters), stating that "it makes more sense here." |
| Restructuring | The participants revises the existing structure in their notes. This can take the form of splitting, merging, or deleting (sub)headings, or splitting a paragraph due to a change in understanding. | The participant combines information under bullets "gut bacteria" and "neurotransmitters" under the newly created heading "mechanisms". |

**Table 3: Cognitive Activity Codes**

| Code | Definition | Example |
|---|---|---|
| Planning | The participant engages in planning activities before or during the search session. | [After taking notes]..."That's a lot about physical health. I'm going to look back at the article to see if I can find anything on anxiety and depression." |
| Monitoring | The participant tracks their progress toward specific goals, acknowledges their focus/direction, or checks how much time remains to finish the task. | "Okay, looking back to my notes to see what I left out... I haven't looked at all about mental health except for stress and anxiety, I guess." |
| Reflecting | The participant pauses their search and enters a reflective phase to review the gathered information and consolidate their learning. | The participant enters a reflective phase, scrolls up-and-down their notes, summarizes the topics covered, and closes tabs no longer wanted/needed. |
| Active Maintenance | The participant keeps information active in their mind and uses it to make connections with newly encountered information. | "I've seen this in other articles... that the [adult] gut microbiome weights about four pounds." |

## 3.8 Learning Assessment

After completing the search task, participants were given two minutes to review their notes and then were asked to write an essay describing everything they learned. Participants were not given a time limit and were not allowed to see their notes while writing the essay. To support the scoring of essays, author A1 examined all the essays and produced a hierarchy of correct statements included by participants. When necessary, the correctness of statements was determined by consulting the source from which the statement was extracted by the participant (determined using the recording of the search session). We refer to this hierarchy of correct statements as the "topic tree". Statements in the "topic tree" were organized into six general topics: (1) the gut microbiome (e.g., organs where gut microbiomes are found); (2) mechanisms through which the gut and brain communicate (e.g., neural pathways); (3) impacts of the gut microbiome on physical health and vice versa (e.g., gut health impacts metabolic health); (4) impacts of the gut microbiome on mental health and vice versa (e.g., poor gut health is linked to depressive symptoms); (5) factors influencing gut health (e.g., genetics); and (6) ways to improve gut health (e.g., sleep). Our "topic tree" of correct statements is also available online.

To measure learning, essays were *primarily* scored based on the number of correct statements included. Other measures are described below. To validate our scoring of essays, authors A1 and A2 independently scored essays produced by 10 participants (23% of the data). For each essay, A1 and A2 highlighted the correct statements in the "topic tree" included by the participant. We measured agreement by computing the Jaccard coefficient between the correct statements identified by A1 and A2. Across all 10 essays, the average Jaccard coefficient was 93.78%. Additionally, we measured the correlation between scores produced by A1's and A2's annotations. The Pearson correlation was 0.992 ($p < .001$) and the Kendall's tau correlation was 0.955 ($p < .001$). Given this high level of agreement, A1 assessed the essays produced by the remaining 34 participants. Ultimately, we measured learning based on A1's annotations.

In addition to scoring essays based on the number of correct statements included, we computed three other measures that leveraged the organization of correct statements in the "topic tree" into the six general topics described above. Our *breadth* measure considers the number of general topics (0-6) with at least two correct statements. Our *depth* measure considers the number of correct statements associated with the general topic with the most correct statements. Finally, to measure the balance between breadth and depth, we considered the harmonic mean of breadth and depth scores.

## 3.9 Statistical Analysis

Our four research questions (RQ1-RQ4) focus on differences between low- and high-WM groups. Most of our dependent variables (94%) were not normally distributed. Therefore, we decided to use non-parametric Mann-Whitney $U$ tests to check for statistically significant differences between groups. In addition to reporting $p$-values, we report $U$ statistic values. Our $U$ statistic values can be interpreted as the number of "pairwise wins" between participants in the high-WM group versus participants in the low-WM group. Given that we had 22 participants in each group, our $U$ statistic values range from 0 to 484 ($22 \times 22$). Values closer to 0 or 484 imply significant differences between groups. Values closer to 242 (the midpoint) imply no significant differences between groups. Given the exploratory nature of our study, all tests were run as two-tailed tests.

# 4 Results

In the following sections, we present our results for RQ1-RQ4. To conserve space, we only include figures for outcome measures with significant differences. All figures are box plots. In the text, we use $M_{\text{L}}$ and $M_{\text{H}}$ to denote the median for the low-WM and high-WM groups, respectively.

## 4.1 Effects on Pre-task Perceptions

Before presenting our results for RQ1-RQ4, we report on differences in pre-task perceptions. As with any between-subjects study, there is always a risk of confounding factors based on differences between participant groups (other than WM). For example, what if high-WM participants were more interested in the task topic by random chance? Based on responses to the pre-task questionnaire, there were no significant differences between low- and high-WM groups in terms of pre-task perceptions of: (1) interest in the task topic, (2) prior knowledge, (3) motivation, (4) expected difficulty, or (5) *a priori* determinability (all $p \geq .54$).

## 4.2 RQ1: Effects on Post-task Perceptions

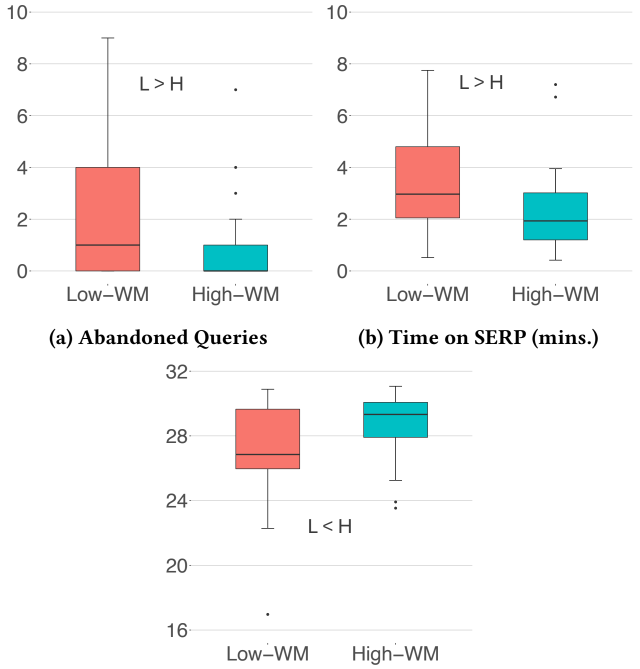
In RQ1, we investigate the effects of working memory on post-task perceptions. There were no significant differences between low- and high-WM groups. Specifically, there were no significant differences in terms of satisfaction or cognitive load (intrinsic, extraneous, and germane). Additionally, there were no significant differences in terms of the extent to which participants perceived to have engaged in: (1) planning, (2) monitoring their progress and understanding, (3) organizing information, or (4) evaluating/adapting their approach to the task.

## 4.3 RQ2: Effects on Search Behaviors

In RQ2, we investigate the effects of working memory on search behaviors. To this end, we considered the different behavioral measures described in Section 3.6. As shown in Figure 1, we found significant differences between low- and high-WM groups for three measures. On average, high-WM participants had fewer abandoned queries ($M_{\text{H}} = 0.00$ vs. $M_{\text{L}} = 1.00$, $U = 167$, $p = .025$); spent less time on the SERP ($M_{\text{H}} = 1.93$ vs. $M_{\text{L}} = 2.97$, $U = 165.5$, $p = .035$); and spent more time either on pages visited or their notes ($M_{\text{H}} = 29.33$ vs. $M_{\text{L}} = 26.85$, $U = 341.5$, $p = .032$).

## 4.4 RQ3: Effects on SSM and Cognitive Activities

In RQ3, we investigate the effects of working memory on the extent to which participants engaged in specific search and sensemaking (SSM) and cognitive activities. RQ3 is the research question that mostly distinguishes our work from prior work. Therefore, Table 4 shows all differences between groups. Before delving into differences that were statistically significant (*), it is worth noting



**(a) Abandoned Queries**

**(b) Time on SERP (mins.)**

**(c) Time on Pages and Notes (mins.)**

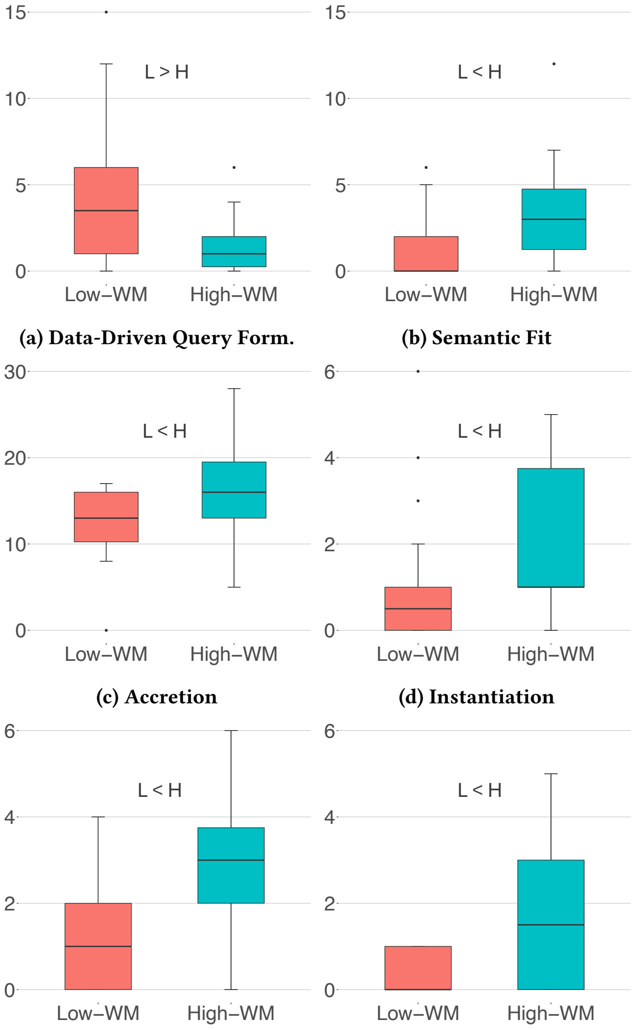
**Figure 1: Effects on Search Behaviors**

that some activities (i.e., searching for structure, tuning, restructuring) were rarely observed. As shown in Section 3.4, the task description given to participants included different subtopics that participants should consider. Perhaps this explains why participants rarely needed to search for structure. Additionally, we observed that participants primarily used these subtopics to structure their notes, which is a logical approach. Perhaps this explains why they rarely needed to tune or restructure the organization of their notes.

Our RQ3 results found significant differences for six SSM and cognitive activities. As shown in Figure 2, high-WM participants engaged in significantly *less* data-driven query formulation ($U = 152$, $p = .033$) and *more* semantic fit ($U = 384.5$, $p = .001$); accretion ($U = 339$, $p = .023$); instantiation ($U = 330$, $p = .032$); monitoring ($U = 356.5$, $p = .006$); and active maintenance ($U = 357.5$, $p = .004$). Median values are shown in Table 4.

### 4.5 RQ4: Effects on Learning Outcomes

In RQ4, we investigate the effects of working memory on learning outcomes. To this end, participants were asked to write an essay describing everything they learned during the search task. Essays were scored based on four measures described in Section 3.8. We found significant differences between groups for all four measures. As shown in Figure 3, high-WM participants included significantly more correct statements in their essays ($M_{\mathrm{H}} = 22.50$ vs. $M_{\mathrm{L}} = 14.50$, $U = 376$, $p = .002$); covered a broader range of topics ($M_{\mathrm{H}} = 4.50$ vs. $M_{\mathrm{L}} = 4.00$, $U = 324.5$, $p = .045$); covered at least one topic in greater depth ($M_{\mathrm{H}} = 8.50$ vs. $M_{\mathrm{L}} = 6.00$, $U = 360$, $p = .005$); and had higher harmonic means between breadth and depth ($M_{\mathrm{H}} = 5.77$ vs. $M_{\mathrm{L}} = 4.61$, $U = 382$, $p = .001$). Based on these measures, our results suggest that high-WM participants had better learning outcomes.



**(a) Data-Driven Query Form.**

**(b) Semantic Fit**

**(c) Accretion**

**(d) Instantiation**

**(e) Monitoring**

**(f) Active Maintenance**

**Figure 2: Effects on SSM and Cognitive Activities**

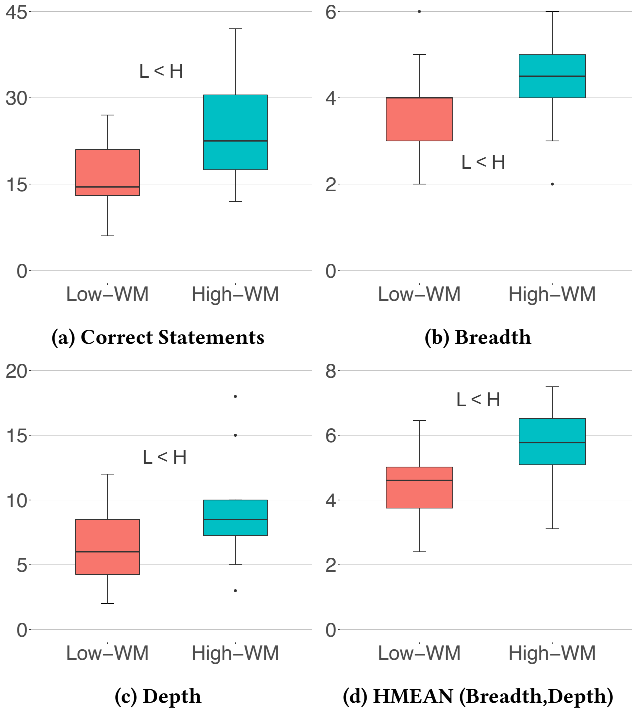
## 5 Discussion

In this section, we summarize our results, compare them to results from prior work, discuss their implications, and propose directions for future research.

**RQ1: Post-task Perceptions:** In terms of post-task perceptions, we did not find any significant differences between groups. Low- and high-WM participants reported similar levels of satisfaction with their performance; cognitive load (intrinsic, extraneous, and germane); and the extent to which they engaged in certain cognitive activities during the task (planning, monitoring, organizing, and evaluating/adapting).

Table 4: Differences in SSM & cognitive activities between groups. * indicates activities with significant differences ($p < .05$)

| | | low-WM | | | high-WM | |
|---|---|---|---|---|---|---|
| | min | $M_L$ (Q1,Q3) | max | min | $M_H$ (Q1,Q3) | max |
| structure driven | 1 | 3 (2,5) | 7 | 1 | 2.5 (2,4) | 9 |
| data driven* | 0 | 3.5 (1,6) | 18 | 0 | 1 (0.25,2) | 6 |
| selective with source | 0 | 0.5 (0,2) | 5 | 0 | 2 (0,4) | 6 |
| search for structure | 0 | 0 (0,0) | 0 | 0 | 0 (0,0) | 1 |
| gap identification | 0 | 1 (0,2) | 6 | 0 | 0 (0,2) | 6 |
| building | 0 | 1 (0.25,3) | 6 | 0 | 2 (1,3) | 6 |
| semantic fit* | 0 | 0 (0,2) | 6 | 0 | 3 (1.25,4.75) | 12 |
| accretion* | 0 | 13 (10.25,16) | 17 | 5 | 16 (13,19.5) | 34 |
| instantiation* | 0 | 0.5 (0,1) | 6 | 0 | 1 (1,3.75) | 9 |
| tuning | 0 | 0 (0,0) | 1 | 0 | 0 (0,1) | 2 |
| restructuring | 0 | 0 (0,0) | 3 | 0 | 0 (0,0) | 3 |
| planning | 0 | 0 (0,1) | 4 | 0 | 1 (0,2) | 3 |
| monitoring* | 0 | 1 (0,2) | 4 | 0 | 3 (2,3.75) | 7 |
| reflecting | 0 | 0 (0,0.75) | 2 | 0 | 0.5 (0,1) | 2 |
| active maintenance* | 0 | 0 (0,1) | 1 | 0 | 1.5 (0,3) | 5 |



(a) Correct Statements

(b) Breadth

(c) Depth

(d) HMEAN (Breadth,Depth)

Figure 3: Effects on Learning Outcomes

Based on our results and those from prior work, the impact of working memory on post-task perceptions does not seem straightforward. Prior studies have found conflicting results. Choi et al. [12] also found that, while high-WM participants exerted more effort, both groups reported similar levels of workload and satisfaction. In contrast, Choi et al. [11] conducted a study using two tools to help participants save information. One tool enabled participants to organize the saved information by topic and the other (baseline) tool did not. High-WM participants reported similar perceptions with both tools. Conversely, when using the baseline tool, low-WM participants reported greater difficulty in deciding when they had enough information.

Results from the studies above suggest that post-task perceptions may be more greatly affected among low-WM participants when they are exposed to challenging situations with a clear reference for comparison. Similar to Choi et al. [12], our study did not involve a within-subjects system manipulation, and participants used familiar tools (i.e., Google Search and a Google Doc) to complete the task. This might have contributed to the lack of group differences.

Interestingly, our RQ1 results are *incongruent* with our RQ3 and RQ4 results. In terms of RQ3, high-WM participants exerted more effort (e.g., more accretion & instantiation) but did *not* report higher levels of germane cognitive load. Similarly, high-WM participants engaged in more monitoring activities but did *not* report to have engaged in higher levels of monitoring. In terms of RQ4, high-WM participants had better learning outcomes but did *not* report higher levels of satisfaction. It may be the low- and high-WM participants had different *standards* for measuring effort and the quality of the task outcome (i.e., knowledge gains).

**RQ2: Search Behaviors:** Low-WM participants had more abandoned queries, spent more time on the SERP, and spent less time on pages and their notes. There are two possible explanations for this trend.

One explanation is that low-WM participants were less efficient at finding the information they sought on the SERP, for several reasons. First, results from Fourney et al. [21] suggest that dyslexic searchers make less extreme relevance judgments whereas non-dyslexic searchers are more likely to judge results as either highly relevant or highly non-relevant. While a heterogeneous condition, dyslexia is associated with deficits in working memory [44, 46]. Therefore, low-WM participants may have taken longer to evaluate results on the SERP before deciding what to click on. Second, our RQ3 results found that low-WM participants issued more data-driven queries—queries influenced by information encountered during the session instead of queries guided by topics in the task description. Therefore, perhaps low-WM participants spent more time seeking answers to highly specific questions rather than exploring more general topics (e.g., factors that influence gut health).

A second explanation is that low-WM participants were more likely to engage with information directly on the SERP (e.g., "featured snippets", "people also ask" cards, "things to know" cards, etc.). Anecdotally, we observed quite a bit of this behavior in our video analysis of sessions for RQ3. However, we did not record exact frequencies of this occurring to compare between low- and high-WM groups.

Our RQ2 results have both similarities and differences with those from prior work, which may be due to the types of tasks assigned to participants across studies. In terms of similarities, Gwizdka [23] also found that high-WM participants spent more time reading pages. Our task was very similar to those used by Gwizdka [23]—learning about multifaceted health-related topics. In terms of differences, prior studies found that high-WM participants issued more queries [12, 22]. We did *not* find significant differences in the number of queries issued between high- and low-WM participants. The tasks used in Gwizdka [22] and Choi et al. [12] involved comparing a set of alternatives, which may have required more querying. Therefore, our RQ2 results suggest that the impact of working memory on search behaviors may be task-dependent.

**RQ3: SSM and Cognitive Activities:** Prior work has not considered the effects of working memory on SSM and cognitive activities. Therefore, here, we elaborate on our RQ3 results.

Low-WM participants issued more data-driven queries. This suggests that low-WM participants were more likely to engage in bottom-up processes. Bottom-up processes are guided by unanticipated, highly specific needs that emerge during the SSM process. Conversely, top-down processes are guided by preset goals or gaps identified within the current knowledge structure. In terms of data-driven querying, we observed the following three behaviors. First, low-WM participants conducted look-up searches more frequently. The need to define unknown terms seemed immediate and they often interrupted their reading to initiate a search. Second, low-WM participants were more likely to develop unanticipated information needs based on information encountered and their personal interests. These needs were not always pertinent to the overall goal of the task. Finally, resolving these bottom-up needs often required several query reformulations, increasing the number of data-driven queries and explaining the greater number of abandoned queries (RQ2) from low-WM participants.

High-WM participants engaged in more accretion. This suggests that high-WM participants exerted more effort in extracting relevant information from pages and recording it in their notes. Prior studies have also found that high-WM participants exert more effort [12, 22, 23]. The greater level of accretion from high-WM participants may also suggest that they were better able to switch between searching, reading, and note-taking. Cognitive science research has shown that working memory impacts task-switching ability more than many other cognitive abilities and personality traits [36].

High-WM participants engaged in more instantiation. Instantiation captures instances where participants integrated new information into their existing notes to reinforce, elaborate on, or expand on previously recorded concepts and ideas. Our results suggest that high-WM participants were better able to connect newly encountered information with previously recorded notes. This is further supported by high-WM participants engaging in more active maintenance (discussed later).

High- and low-WM participants engaged in similar levels of building, tuning, and restructuring. In terms of building, 10 out of 44 participants did not engage in any observable organizing activities. Instead, they took notes chronologically based on the order in which they encountered information. These 10 participants were evenly split between high- and low-WM groups. Participants who engaged in building activities the most were also evenly split between high- and low-WM groups. It may be that working memory does not systematically impact an individual's tendency to organize their notes by topic. As shown in Table 4, high- and low-WM participants rarely engaged in tuning and restructuring. There are two possible explanations for this trend. First, the task description included topics that participants were asked to consider. This may have served as a form of scaffolding, making tuning and restructuring unnecessary. Second, tuning and restructuring typically occur after some time [38, 39]. We may have observed more tuning and restructuring if participants had been given much longer than 30 minutes to work on the task. Additionally, a lab-based setting might not be the most conducive for studying activities like tuning and restructuring, which requires the repeated use of a structure or schema. Longitudinal studies involving learning over weeks or months outside of a lab could provide a better setting for studying these activities in depth.

High-WM participants engaged in more semantic fit. This suggests that high-WM participants more actively considered how information was relevant to their current goals, topics in the task description, and the structure in their notes. Actively assessing the semantic fit between encountered information and the current or overall goals of the task can help prevent one from being sidetracked. This reflects an individual's ability and effort to maintain a big-picture perspective on how the encountered information fits into the broader understanding of the topic, which can be considered a top-down strategy.

High-WM participants engaged in more active maintenance. That is, high-WM participants more frequently demonstrated their ability to keep information active in their memory. This was observed in four scenarios: (1) when participants recognized that information was relevant to their current goal; (2) when participants recognized that information was relevant to other subtopics; (3) when participants noted that information corroborated or contradicted previously encountered information; and (4) when participants identified relations between different pieces of information. Active maintenance is a key aspect of working memory. The scenarios above illustrate how working memory can manifest in search and sensemaking.

Finally, high-WM participants engaged in more monitoring. This was observed in three scenarios: (1) when participants revisited the task description to monitor their progress across subtopics; (2) when they reviewed their notes to identify gaps; and (3) when they checked with the study moderator about the remaining time. While high-WM participants engaged in more monitoring, perceptions of monitoring were not significantly different between groups (RQ1). This result suggests that *perceptions* of engagement in a specific cognitive activity may not always align with *actual* engagement.

**RQ4: Learning Outcomes:** High-WM participants learned more. This was evidenced by the number of correct statements in their knowledge summaries produced after the task. Additionally, their correct statements had greater breadth and depth. Pardi et al. [35] found a similar result—high-WM participants included a greater number of relevant concepts in their knowledge summaries after searching. Our study extends this prior work in several ways. First, we used a more complex task and took a different approach to measure learning. Second, and more importantly, our RQ3 results help explain *why* high-WM participants might have achieved better learning outcomes. They engaged in more top-down and fewer bottom-up processes; they actively evaluated information based on their goals; they maintained information in active memory; and they actively monitored their progress.

**Opportunities for Future Work:** Our findings underscore the need for search tools to support neurodiverse users. Specifically, we envision four different types of tools. Features of these tools could be developed using state-of-the-art Generative AI (GenAI) technologies.

First, our results highlight the need to support searchers with both top-down (goal-driven) and bottom-up (data-driven) processes. While both processes are essential, data-driven processes (e.g., querying for definitions while reading a page) can sometimes disrupt the task flow. Low-WM individuals may struggle to re-focus after pursuing a data-driven need. Experimental tools could enable searchers to query a system and see results directly from a document. Such tools could help prevent users from getting sidetracked and losing sight of their goals.

Second, several prior studies experimented with a tool that allowed participants to explicitly write subgoals associated with the task, take notes with respect to specific subgoals, and mark subgoals as completed [47, 48]. In both studies, access to the tool resulted in better learning outcomes. One study found that access to the tool resulted in participants engaging in higher levels of monitoring [48]. Specifically, participants with access to the goal-setting tool were more active in: (1) tracking their progress toward each subgoal; (2) judging the relevance of information with respect to their subgoals; and (3) evaluating the time allocated to specific subgoals. In our study, high-WM participants engaged in higher levels of monitoring and semantic fit. Therefore, goal-setting tools could be particularly beneficial for low-WM individuals. Pop-up and notifications could also remind searchers about the current subgoal or neglected subgoals. Additionally, GenAI technology could be used to automatically highlight when information is relevant to an explicitly written subgoal.

Beyond goal-setting, prior studies have also experimented with tools to support searchers during complex, multifaceted search tasks. Examples include: (1) tools to annotate documents and see a summary of annotations [37]; (2) visualizations about the coverage of subtopics during the search session [14]; (3) and tools to save and organize information by subtopic [11]. These tools have one thing in common—they help searchers track their progress. Low-WM individuals may benefit from tools that provide a visual reminder of the different subtopics being pursued and their progress across subtopics. Additionally, enabling searchers to *explicitly* represent the different subtopics being pursued could enable systems to predict when information is relevant to a subtopic, even when the subtopic is not the one being currently pursued.

Finally, in our study, high-WM participants demonstrated higher levels of active maintenance. This manifested in several ways. For example, high-WM participants noticed when new information corroborated/contradicted previously encountered information and drew connections between different pieces of information. GenAI tools could be used to highlight when new textual passages relate to previously read passages.

## 6 Conclusion

We reported on a lab study that investigated the role of working memory (WM) during a search and sensemaking (SSM) task. We investigated the effects of WM on: (RQ1) post-task perceptions; (RQ2) search behaviors; (RQ3) the extent to which participants engaged in specific search, sensemaking, and cognitive activities; and (RQ4) learning outcomes. Our results showed the following trends. We did not observe significant differences in post-task perceptions between low- and high-WM groups. Interestingly, however, we observed significant differences for RQ2-RQ4. In terms of RQ2, high-WM participants had fewer abandoned queries, spent less time on the search interface, and spent more time reading pages and taking notes. We observed several significant differences for RQ3. In terms of search activities, low-WM participants issued more data-driven queries (i.e., motivated by information encountered during the session versus topics that were part of the task description). In terms of sensemaking activities, high-WM participants were more active in: (1) evaluating information based on their goals and the structure in their notes (semantic fit); (2) adding information to their notes (accretion); and (3) elaborating on information in their notes (instantiation). In terms of cognitive activities, high-WM participants were more active in monitoring their progress and demonstrated more active maintenance—keeping information in memory to notice and make connections. Finally, in terms of RQ4, high-WM participants had better learning outcomes. Our RQ2 & RQ3 results provide insights into *why* high-WM participants may have had better learning outcomes. We have discussed possible tools to support neurodiverse users during complex SSM tasks.

## References

[1] Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, complete edition.*

[2] Richard C Anderson. 1984. Some reflections on the acquisition of knowledge. *Educational researcher* 13, 9 (1984), 5–10.

[3] Jaime Arguello and Bogeum Choi. 2019. The effects of working memory, perceptual speed, and inhibition in aggregated search. *ACM Transactions of Informatin Systems* 37, 3 (2019), 1–34.

[4] W Brian Arthur. 1994. Inductive reasoning and bounded rationality. *The American economic review* 84, 2 (1994), 406–411.

[5] David Paul Ausubel, Joseph Donald Novak, Helen Hanesian, et al. 1968. *Educational psychology: A cognitive view.* Vol. 6. Holt, Rinehart and Winston New York.

[6] Steven Banas and Christopher A. Sanchez. 2012. Working memory capacity and learning underlying conceptual relationships across multiple documents. *Applied Cognitive Psychology* 26, 4 (2012), 594–600.

[7] Nilavra Bhattacharya and Jacek Gwizdka. 2018. Relating eye-tracking measures with changes in knowledge on search tasks. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, Article 62.

[8] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information Processing & Management* 31, 2 (1995), 191–213.

[9] Judy Cantor, Randall W Engle, and George Hamilton. 1991. Short-term memory, working memory, and verbal abilities: How do they relate? *Intelligence* 15, 2 (1991), 229–246.

[10] Stephanie M Carlson. 2005. Developmentally sensitive measures of executive function in preschool children. *Developmental neuropsychology* 28, 2 (2005), 595–616.

[11] Bogeum Choi, Jaime Arguello, Robert Capra, and Austin R. Ward. 2023. The influences of a knowledge representation tool on searchers with varying cognitive abilities. *ACM Transactions on Information Systems* 41, 1, Article 18 (2023), 35 pages.

[12] Bogeum Choi, Robert Capra, and Jaime Arguello. 2019. The effects of working memory during search tasks of varying complexity. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 261–265.

[13] Andrew Conway, Michael Kane, Michael Bunting, Zach Hambrick, Oliver Wilhelm, and Randall Engle. 2005. Working memory span task: A methodological review and user's guide. *Psychonomic bulletin & review* 12 (2005), 769–786.

[14] Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. 2021. Searching to learn with instructional scaffolding. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 209–218.

[15] Meredyth Daneman and Patricia A Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior* 19, 4 (1980), 450–466.

[16] Wim De Neys, Walter Schaeken, and Géry d'Ydewalle. 2005. Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking & Reasoning* 11, 4 (2005), 349–381.

[17] Brenda Dervin. 1998. Sense-making theory and practice: An overview of user interests in knowledge seeking and use. *Journal of knowledge management* 2, 2 (1998), 36–46.

[18] Adele Diamond. 2013. Executive functions. *Annual review of psychology* 64 (2013), 135–168.

[19] Randall W Engle, John K Nations, and Judy Cantor. 1990. Is "working memory capacity" just another name for word knowledge? *Journal of Educational Psychology* 82, 4 (1990), 799–804.

[20] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods* 5, 1 (2006), 80–92.

[21] Adam Fourney, Meredith Ringel Morris, Abdullah Ali, and Laura Vonessen. 2018. Assessing the readability of web search results for searchers with dyslexia. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, New York, NY, USA, 1069–1072.

[22] Jacek Gwizdka. 2013. Effects of working memory capacity on users' search effort. In *Proceedings of the International Conference on Multimedia, Interaction, Design and Innovation*. ACM.

[23] Jacek Gwizdka. 2017. I can and so I search more: effects of memory span on search behavior. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*. New York, NY, USA, 341–344.

[24] M. J. Kane, M. K. Bleckley, A. R. A. Conway, and R. W. Engle. 2001. A controlled-attention view of working-memory capacity. *Journal of experimental psychology: General* 130, 2 (2001), 169–183.

[25] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems* 21, 4 (2006), 70–73.

[26] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems* 21, 5 (2006), 88–92.

[27] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data–frame theory of sensemaking. In *Expertise out of context*. Psychology Press, 118–160.

[28] J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.

[29] A. MacFarlane, A. Albrair, C. R. Marshall, and G. Buchanan. 2012. Phonological working memory impacts on information searching: An investigation of dyslexia. In *Proceedings of the 4th Information Interaction in Context Symposium*. ACM, New York, NY, USA, 27–34.

[30] Andrew MacFarlane, George Buchanan, Areej Al-Wabil, Gennady Andrienko, and Natalia Andrienko. 2017. Visual analysis of dyslexia on search. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 285–288.

[31] Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. Sage Publication, Inc.

[32] Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. 2018. Understanding the needs of searchers with dyslexia. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12.

[33] Jane Oakhill, Joanne Hartt, and Deborah Samols. 2005. Levels of Comprehension Monitoring and Working Memory in Good and Poor Comprehenders. *Reading and Writing* 18, 7 (2005), 657–686.

[34] Fred Paas, Tamara Van Gog, and John Sweller. 2010. Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational psychology review* 22 (2010), 115–121.

[35] Georg Pardi, Johannes von Hoyer, Peter Holtz, and Yvonne Kammerer. 2020. The Role of Cognitive Abilities and Time Spent on Texts and Videos in a Multimodal Searching as Learning Task. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. ACM, 378–382.

[36] Thomas S. Redick. 2016. On the relation of working memory and multitasking: Memory span and synthetic work performance. *Journal of Applied Research in Memory and Cognition* 5, 4 (2016), 401–409.

[37] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 229–238.

[38] David E Rumelhart and Donald A Norman. 1976. *Accretion, tuning and restructuring: Three modes of learning*. Technical Report. California Univ San Diego La Jolla Center for human information processing.

[39] David E Rumelhart and Donald A Norman. 1981. Analogical Processes in Learning. In *Cognitive Skills and Their Acquisition*. Psychology Press, 335–359.

[40] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, New York, NY, USA, 269–276.

[41] Timothy A Salthouse, Thomas M Atkinson, and Diane E Berish. 2003. Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of experimental psychology: General* 132, 4 (2003), 566–594.

[42] C. A. Sanchez and J. Wiley. 2006. An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition* 34, 2 (2006), 344–355.

[43] Joseph Sharit, Mario A. Hernández, Sara J. Czaja, and Peter Pirolli. 2008. Investigating the Roles of Knowledge and Cognitive Abilities in Older Adult Information Seeking on the Web. *ACM Transactions on Computer-Human Interaction* 15, 1 (2008).

[44] J. H. Smith-Spark and J. E. Fisk. 2007. Working memory functioning in developmental dyslexia. *Memory* 15, 1 (2007), 34–56.

[45] Stephen Toulmin, Richard D Rieke, and Allan Janik. 1979. An introduction to reasoning. (1979).

[46] S. Turker, A. Seither-Preisler, S. M. Reiterer, and P. Schneider. 2019. Cognitive and behavioural weaknesses in children with reading disorder and AD(H)D. *Scientific Reports* 9, 1 (2019), 1–11.

[47] Kelsey Urgo and Jaime Arguello. 2023. Goal-setting in support of learning during search: An exploration of learning outcomes and searcher perceptions. *Information Processing & Management* 60, 2 (2023), 103158.

[48] Kelsey Urgo and Jaime Arguello. 2024. The effects of goal-setting on learning outcomes and self-regulated learning processes. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 278–290.

[49] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

[50] Barbara M Wildemuth and Luanne Freund. 2012. Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the symposium on human-computer interaction and information retrieval*. 1–10.

[51] Nicola Yuill, Jane Oakhill, and Alan Parkin. 1989. Working memory, comprehension ability and the resolution of text anomaly. *British journal of psychology* 80, 3 (1989), 351–361.

[52] Pengyi Zhang and Dagobert Soergel. 2014. Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. *JASIST* 65, 9 (2014), 1733–1756.

[53] Pengyi Zhang and Dagobert Soergel. 2016. Process patterns and conceptual changes in knowledge representations during information seeking and sensemaking. *Journal of Information Science* 42, 1 (2016), 59–78.

[54] Pengyi Zhang and Dagobert Soergel. 2020. Cognitive mechanisms in sensemaking: A qualitative user study. *Journal of the Association for Information Science and Technology* 71, 2 (2020), 158–171.

[55] Pengyi Zhang, Dagobert Soergel, Judith L Klavans, and Douglas W Oard. 2008. Extending sense-making models with ideas from cognition and learning theories. *Association for Information Science and Technology* 45, 1 (2008), 23–23.