Foundations and Trends® in Information Retrieval

Search as Learning

Suggested Citation: Kelsey Urgo and Jaime Arguello (2025), "Search as Learning", Foundations and Trends® in Information Retrieval: Vol. 19, No. 4, pp 365–556. DOI: 10.1561/1500000084.

Kelsey Urgo
University of San Francisco
kurgo@usfca.edu

Jaime Arguello
University of North Carolina at Chapel Hill
jarguello@unc.edu

now
the essence of knowledge
Boston — Delft

Contents

Search as Learning

Kelsey Urgo[1] and Jaime Arguello[2]

[1]*Department of Computer Science, University of San Francisco, USA; kurgo@usfca.edu*

[2]*School of Information and Library Science, University of North Carolina at Chapel Hill, USA; jarguello@unc.edu*

ABSTRACT

Search systems are often designed to support simple look-up tasks, such as fact-finding and navigation tasks. However, people increasingly use search engines to complete tasks that require deeper learning. In recent years, the search as learning (SAL) research community has argued that search systems should also be designed to support information-seeking tasks that involve complex learning as an important outcome. This monograph aims to provide a comprehensive review of prior research in search as learning and related areas.

Searching to learn can be characterized by specific learning objectives, strategies, and context. Therefore, we begin by reviewing research in education that has aimed at characterizing learning objectives, strategies, and context. Then, we review methods used in prior studies to measure learning during a search session. Here, we discuss two important recommendations for future work: (1) measuring learning retention and (2) measuring a learner's ability to transfer their new knowledge to a novel scenario. Following this, we discuss studies that have focused on understanding factors that influence learning during search and search behaviors that are

---

Kelsey Urgo and Jaime Arguello (2025), "Search as Learning", Foundations and Trends® in Information Retrieval: Vol. 19, No. 4, pp 365–556. DOI: 10.1561/1500000084.

predictive of learning. Next, we survey tools that have been developed to support learning during search. Searching for the purpose of learning is often a solitary activity. Research in self-regulated learning (SRL) aims to understand how people monitor and control their own learning. Therefore, we review existing models of SRL, methods to measure engagement with specific SRL processes, and tools to support effective SRL. We conclude by discussing potential areas for future research.

1

Introduction

For over a decade, researchers in the field of Search as Learning (SAL) have recognized that users frequently turn to search systems not only for simple fact-finding but to engage in complex learning tasks. This recognition has led to a growing body of work investigating how search systems can better support users in achieving complex learning outcomes. Over the years, researchers have explored many dimensions of SAL, including how learning objectives are defined in search contexts, the strategies learners use during search, and the factors that influence learning during information-seeking processes.

This monograph aims to highlight the significant progress made in SAL research, synthesizing key contributions while also framing future directions for this evolving field. Recent advancements, such as the integration of generative AI with search systems, underscore the need to revisit foundational theories and methodologies in light of new technologies. By reflecting on what has been accomplished and identifying gaps and opportunities, this monograph seeks to inspire future research and innovation in SAL. Addressing these research gaps will help to ensure that search systems are equipped to meet the ever-evolving demands of individuals by supporting their learning needs in a thoughtful, human-centered manner.

Learning theory is a vast and multidisciplinary field that includes a wide range of perspectives and approaches. This review focuses on specific sub-areas of learning theory, particularly those that have been most influential in shaping research and practice in SAL. While we highlight frameworks such as self-regulated learning (SRL) and tools like MetaTutor as exemplars, we also draw on foundational theories, including the Anderson & Krathwohl Taxonomy of Learning, constructivism, and the Zone of Proximal Development (ZPD), to frame our discussion. These frameworks and theories represent critical dimensions of how learning processes can be supported through search systems. However, we acknowledge that these perspectives are not exhaustive. Researchers engaging in SAL work are encouraged to explore relevant areas of the learning sciences that align with their study's theoretical lens and build on the foundational perspectives outlined in this monograph.

Additionally, while this monograph provides foundational perspectives to guide SAL research, it is important to acknowledge that it does not comprehensively connect all existing SAL research to the broader theories, frameworks, and empirical research from the learning sciences or other related fields. Given the breadth and complexity of these domains, this work emphasizes perspectives and connections most directly relevant to advancing SAL. This approach highlights opportunities for future researchers to explore novel connections between SAL and the wider landscape of learning sciences. Such efforts can enrich the field and inform the design of human-centered search systems that better support complex learning tasks.

In this section, we provide an overview of SAL. In particular, we discuss the foundations of SAL research and its primary objectives as established by researchers in the field. Next, we discuss concepts from developmental psychology and the learning sciences in which SAL is rooted, including constructivism, Vygotsky's Zone of Proximal Development (ZPD), and scaffolding. Then, we discuss the adjacent field of Intelligent Tutoring Systems (ITS) as this work is rooted in the same theory as SAL and shares similar objectives. Finally, we discuss exploratory search as it is a framework that centers learning and creating as important outcomes of information seeking.

1.1 Overview of Search as Learning (SAL)

Search systems are often designed, implemented, and evaluated as tools to help people find information. However, more than ever before, people use search systems to learn about a topic. For the most part, SAL research is concerned with scenarios in which a person interacts with a search system to fulfill a specific *learning objective*.

Key Takeaway

Search as Learning (SAL) explores how people interact with search systems to achieve their learning objectives.

A natural question is: What is a learning objective? Learning objectives have been characterized from different perspectives. One common characterization views learning objectives as having two main parts. First, a learning objective has a specific topic or domain. This can be called the *knowledge type* of the objective. Knowledge types can range from factual, to conceptual, to procedural knowledge. For example, imagine a searcher who wants to find the depth of the deepest part of the ocean. This searcher is aiming to gain factual knowledge. Imagine a searcher who wants to learn about the biological process of osmosis. This searcher is aiming to gain conceptual knowledge. Finally, imagine a searcher who wants to learn how to compute the area of a circle. This searcher is aiming to gain procedural knowledge.

Second, a learning objective has a specific *cognitive process*. The cognitive process of the objective defines the types of mental processes the learner wants to be able to engage in with the acquired knowledge. Cognitive processes vary by complexity. Perhaps a searcher simply wants to be able to recall the formula for computing the area of a circle. This is a simple objective that only requires rote memorization. Conversely, perhaps a searcher wants to understand why the area of a sphere is four times the area of a circle with the same radius. This is a more complex objective that requires understanding the relation between two procedures. In Section 2, we provide details on this characterization of learning objectives using the Anderson & Krathwohl Taxonomy of Learning (Anderson *et al.*, 2000).

Searching to fulfill a particular learning objective is an iterative process (Urgo and Arguello, 2022c) and can involve multiple sessions. During the search as learning process, searchers often interact with multiple sources, take notes, break the learning objective into smaller learning-oriented subgoals, and revisit topics to build on and check their own understanding. SAL research argues that searching for information not only involves finding answers but also acquiring new knowledge and understanding.

SAL research is multidimensional and considers a wide range of research questions. Some research might focus on understanding the real-world contexts in which people search for the purpose of learning. Other research might focus on better understanding the SAL process. That is, what do people do when they search for the purpose of learning? Other research might focus on developing tools to encourage and support learning during search. Research might also focus on discovering search behaviors that predict learning during search. Finally, SAL research might have a more methodological aim. For example, how might we analyze an artifact like an essay produced after the search session in order to measure learning?

1.2 Early Calls for SAL Research

Learning has been a subject of research in information retrieval (IR) for many years. Three meetings were central to the establishment of the SAL research community: The Second Strategic Workshop on Information Retrieval in Lorne (SWIRL) (Allan *et al.*, 2012), Dagstuhl Seminar 13441 (Agosti *et al.*, 2014), and Dagstuhl Seminar 17092 (Collins-Thompson *et al.*, 2017).

In 2012, the three-day SWIRL workshop emphasized the importance of supporting searching and learning as one of many emerging topics. In 2013, Dagstuhl Seminar 13441 included a working group that focused on "From Searching to Learning." Topics discussed included behaviors that are correlated with learning during search and ways to measure learning during search. Subsequently in 2017, Dagstuhl Seminar 17092 was entirely dedicated to SAL. Discussions from the seminar established four main areas for future research: (1) examining search as a learning process; (2) measuring learning performance and outcomes during search; (3) investigating the contexts in which people search to learn; and (4) developing search tools and interventions to promote learning.

In addition to these workshops, two conference workshops (Freund *et al.*, 2014; Gwizdka *et al.*, 2016), an ASIST panel (Rieh *et al.*, 2014), and two special journal issues focused on SAL (Hansen and Rieh, 2016; Eickhoff *et al.*, 2017).

1.3 Related Topics and Fields

SAL research aims to develop search environments that encourage and support learning. To this end, we must grapple with a few fundamental questions. How do people learn? What is an individual capable of learning at a given point in time? What is the best way for a system to encourage and support learning? SAL researchers are not the first to think about these questions. The SAL research community has pulled from a variety of theories and frameworks established in psychology and education. In this section, we provide an overview of three important concepts: constructivism, the zone of proximal development (ZPD), and scaffolding.

Constructivism is a theory of how people learn. Learning requires an individual to integrate new information into their existing knowledge structures. In this respect, learning requires an individual to be an *active* participant in their own learning process. SAL research is concerned with scenarios in which individuals learn by actively interacting with information using a search system. Therefore, a constructivist perspective on learning is central to SAL research.

The concepts of ZPD and scaffolding go hand in hand. Helping searchers learn begs the question: What can someone learn completely unaided and what can someone learn with some guidance? The ZPD is defined as the range of things an individual might be able to learn with some guidance from a more knowledgeable peer or system. Scaffolding is defined as instructional interventions that support learning while still letting the learner "figure it out on their own." Systems that provide scaffolding adopt a constructivist perspective on learning (i.e., supporting learners in actively constructing their own understanding rather than passively receiving information).

In this section, we also discuss two related research areas: intelligent tutoring systems (ITS) and exploratory search. Research in ITS aims to develop non-search, computer-based systems that help people learn. Exploratory search considers search tasks that involve learning and creativity as important outcomes.

1.3.1 Constructivism

How do people learn? Introduced by Jean Piaget (Piaget and Cook, 1952), the theory of constructivism argues that individuals learn through experiences and social interaction, and by integrating new information with their existing knowledge. That is, individuals are not empty vessels that acquire knowledge only through absorption during direct instruction. Instead, learning requires an individual to engage with new material and integrate it into their existing knowledge. In this respect, constructivism indicates that learners must be active participants in their own learning process. For example, someone is likely to learn about a procedure more deeply by using the procedure to solve a problem rather than simply memorizing and reciting the steps.

Key Takeaway

Constructivism asserts that meaningful learning occurs when learners actively engage in experiences, enabling them to integrate new knowledge into their existing understanding.

The theory of constructivism argues that people learn through the processes of assimilation and accommodation (Piaget and Cook, 1952; Hanfstingl *et al.*, 2021). Assimilation is the process of taking new information and fitting it into an existing schema. Sometimes, the new information does not fit neatly into an existing schema. Therefore, accommodation is the process of using newly acquired information to revise or redevelop the existing schema, resulting in a more accurate and/or complete schema. Constructivism argues that learning is not a passive activity. People cannot learn by simply "taking in information." They must reflect on it, link it to what they already know, and create new knowledge structures when necessary. Therefore, people learn more when they are active participants in their own learning. Learners that participate in the active construction of their own knowledge gain a deeper understanding, are more able to generalize beyond the learning context, and have higher levels of motivation (Sawyer, 2014).

For decades, much research in information retrieval has adopted a constructivist approach. Talja *et al.* (2005, p. 83) describe the constructivist perspective of the user in information science: "An information user is not a passive information processing system but actively makes sense of the surrounding reality and attaches personal meanings to information."

Within SAL, Eickhoff *et al.* (2017, p. 399) underscored the important role of constructivism in advancing future search system design: "knowledge is derived from personal experience and ideas rather than an aggregation of loose facts and formulas." They also emphasize that: "Despite the wide acceptance and demonstrated success of constructivist methods in pedagogy, common retrieval models do not explicitly manifest any notion of *contextual learning*" (Eickhoff *et al.*, 2017, p. 399).

Constructivism emphasizes that learning occurs when people actively construct knowledge by integrating new information with their existing understanding. However, current search systems are not designed to support these fundamental learning processes of assimilation (i.e., fitting new information into an existing knowledge schema) and accommodation (i.e., adapting or revising a knowledge schema to fit new information). While search engines excel at retrieving relevant information, they do not help learners connect new discoveries with their prior knowledge, nor do they encourage the active engagement necessary for meaningful learning. Search results are typically presented as isolated pieces of information rather than as building blocks that can be integrated into a learner's existing knowledge structure. To support learning, search systems might be designed instead to facilitate active knowledge construction by helping learners connect information to their existing understanding.

1.3.2 Vygostky's Zone of Proximal Development (ZPD)

Vygotsky and Kozulin (1962) introduced the notion of *social constructivism*, which emphasized the importance of social learning through models such as parents or peers. Shown in Figure 1.1, Vygotsky's *Zone of Proximal Development* (ZPD) is a model that positions learning stages into three categories: (1) that which the student can learn on their own; (2) that which the student can learn given assistance from a more knowledgeable peer or mentor; (3) and that which the student is not yet able to learn even with help (Vygotsky, 1980).

The ZPD represents the optimal space for learning, between what learners cannot yet understand and what they are already able to understand on their own. Current search systems, however, present information without consideration for where it falls within a learner's ZPD. However, prior work in SAL has aimed to explore ways in which we can improve upon these existing systems. Smith *et al.* (2022) demonstrate how knowledge graph coverage could be used to infer a learner's ZPD, allowing systems to identify content that is neither too basic nor too advanced for individual learners. Such a system could potentially enhance learning outcomes by ensuring that search results align with a learner's current capabilities and their potential for growth.

In information seeking research, Kuhlthau (1994) models the zone of intervention on the ZPD. The zone of intervention underscores the need for a more knowledgeable peer or instructor to select the appropriate intervention for the individual at the appropriate moment during the information-seeking process. Mechanisms put in place during such an intervention are often known as *scaffolding*.

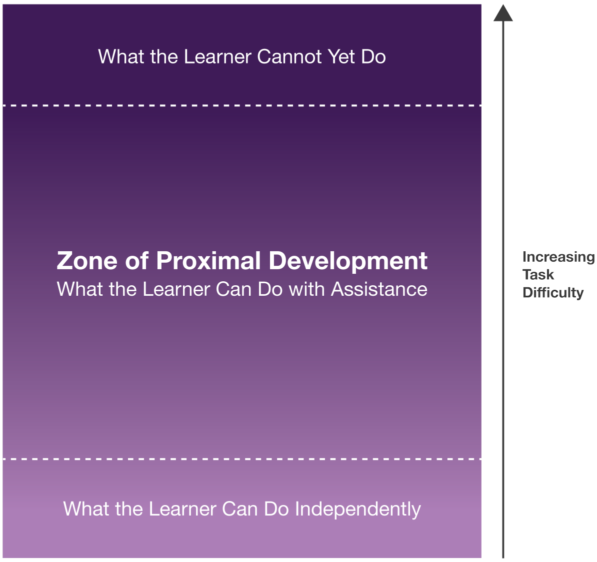

Figure 1.1: Vygotsky's Zone of Proximal Development (adapted from Vygotsky, 1980).

1.3.3 Scaffolding

Scaffolding, as the metaphorical term implies, are mechanisms of support provided by an instructor which are gradually removed or faded as higher levels of cognitive understanding are achieved. Scaffolding is help that is tailored to the learner's needs in order to achieve their goals (Sawyer, 2014).

Key Takeaway

Scaffolding are supports provided by an instructor to facilitate learning that are gradually removed or faded as understanding is achieved.

While simply giving a learner an answer will help them achieve their goal quickly, scaffolding is applied for effective long-term learning. Scaffolding takes a constructivist approach to learning. Good scaffolding provides hints and prompts that help the learner figure things out on their own (Sawyer, 2014). That is, good scaffolding keeps the learner as an active participant in their own learning.

As shown in Table 1.1, Mariani (1997) characterizes effective scaffolding through the dimensions of challenge and support. Ideally, scaffolding is both *high* challenge and *high* support. The other three combinations are likely to lead to suboptimal outcomes. If both challenge and support are low, the learner may become bored and unmotivated. If challenge is high but support is low, the learner may become frustrated and anxious. Finally, if challenge is low but support is high, the learner might feel that they are doing "busy work" and getting little out of the exercise. The best combination, high challenge and high support, is most likely to result in greater engagement, improvements in self-confidence, and better learning outcomes.

Table 1.1: Benefits of High Challenge and High Support Scaffolding (adapted from Mariani, 1997).

	Challenge	
Support	**Low**	**High**
Low	• Low motivation • Boredom • Apathy	• Low self-confidence • High anxiety • Frustration • Failure likely
High	• Low learning • Comfortable • Busy work • Dumbing down	• High learning • High engagement • High self-confidence • Extension of capability

Scaffolding has traditionally referred to the support provided by a teacher or more knowledgeable peer. More recently, a large body of work has broadened scaffolding to include support that is provided by tools, resources, and environments (Sharma and Hannafin, 2007). These tools and resources demonstrate relevant aspects of a task or provide strategies in achieving a learning objective. In particular, such scaffolding has been instantiated in computer-based or technology-enhanced learning environments. These environments implement scaffolding in different ways, for example by: (1) helping the learner understand the landscape of a complex task or domain; (2) visualizing and modeling complex scientific phenomena; and (3) providing interactive guidance and support (Puntambekar and Hübscher, 2005).

While there are clear benefits of using computer-based scaffolding for learning (Belland *et al.*, 2017), Puntambekar and Hübscher (2005) argue that much of the prior work in this area has used a broad application of the term scaffolding that has led to certain shortcomings. Technology-enhanced or computer-based learning tools typically provide passive support. This means that learners do not benefit from the dynamic or adaptive scaffolding that can be provided from a one-on-one teacher. Most often these tools employ blanket supports that are the same for all learners.

Fading (i.e., gradually decreasing scaffolding (McNeill *et al.*, 2006)) is an important and mostly overlooked component of scaffolding by computer-based tools. Typically, support is ongoing and unchanging. Without fading support, learners do not benefit from intermittent self-evaluation of distinguishing what they can and cannot do without support. Current search systems often provide static interfaces and functionality regardless of a user's evolving capabilities and needs. This non-contextualized approach fails to provide the adaptive scaffolding necessary for meaningful learning.

1.3.4 Intelligent Tutoring Systems (ITS)

Individual tutoring is an important method for teaching and learning that researchers have attempted to emulate from the earliest years of computing (Smith and Sherwood, 1976). Intelligent tutoring systems (ITS) have existed for decades. Corbett *et al.* (1997) recognize the first intelligent tutoring program to be SCHOLAR (Carbonell, 1970) from 1970.

There are two categories of ITS: step-based and substep-based. Step-based systems (Kim *et al.*, 1989; Woo *et al.*, 2006) allow learners to enter the steps of their problem-solving process without a tutoring intervention. Substep-based systems (Evens *et al.*, 1997) provide scaffolding and feedback at a finer level of detail than the learners' problem-solving steps. The primary difference between step-based and substep-based systems is that substep-based systems engage learners in a dialogue in order to better understand their reasoning (e.g., ask a learner *why* they made a particular decision) and potentially correct errors at a deeper level of understanding.

The origin of cognitive tutors is rooted in work by Anderson *et al.* (1985), who designed an ITS aimed at supporting the acquisition of cognitive skills, which they define as units of goal-related knowledge. Aleven *et al.* (2006, p. 102) introduced the term Cognitive Tutor as a type of ITS that "is designed to support *learning by doing* and features a cognitive model of the targeted skills, expressed as production rules." The cognitive models integrated into a cognitive tutor represent a learner's thinking in a particular domain and include early learner strategies and misconceptions common to the path from novice to expert. Built on top of these cognitive models are rich graphical problem-solving environments, the combination of which are designed to support individual learning.

1.3.5 ITS Integration of Constructivism, ZPD, and Scaffolding

MetaTutor was developed by Azevedo *et al.* (2009) and differs from other cognitive tutors because it is both an ITS *and* hypermedia learning environment. While cognitive tutors were designed specifically for learning procedural knowledge (using production rules or proof logic formalism), MetaTutor is focused on teaching conceptual knowledge, specifically complex biological processes (e.g., knowledge about circulatory, digestive, and nervous systems) (Azevedo *et al.*, 2009; Azevedo *et al.*, 2012).

In Section 7, we focus on MetaTutor as an example to demonstrate how the foundations of constructivism, Vygotsky's ZPD, and scaffolding have been successfully integrated into learning systems outside of SAL. MetaTutor is also rooted in self-regulated learning (SRL) theory and aims to support SRL processes, which is the focus of Section 7.

1.3.6 Exploratory Search

In the early 2000's, IR researchers recognized that people use search systems for more than simple lookup tasks. However, search systems were inadequate when faced with these types of user demands that included analysis, decision making, and learning about a new topic. Recognizing these emerging needs and expectations of users, Marchionini introduced exploratory search (Marchionini, 2006).

Marchionini identified three large categories of search processes: lookup, learn, and investigate. Lookup processes include fact-finding and verifying—gathering information about who, when, and where. In contrast, exploratory search answers questions related to what, how, and why. Exploratory search involves processes such as learning (e.g., knowledge acquisition, comparison, and integration) and investigating (e.g., analysis, evaluation, and synthesis).

Particularly relevant to the roots of SAL, Marchionini describes *learning searches* as part of exploratory search. Learning searches involve multiple search iterations, sifting through various types of media, complex cognitive processing, and comparing and judging information. Learning searches are rooted not only in traditional academics, but also in general lifelong and professional learning.

Key Takeaway

Learning searches involve multiple search iterations, multiple types of media, and complex cognitive processing like comparing and judging information

Investigation searches involve multiple iterations over an extended period. Investigative searchers are more critical of information before it is integrated into their existing knowledge structures. Like learning searches, investigation searches are also learning-oriented. However, they involve cognitive processes that are highly complex (e.g., analysis, evaluation, and synthesis). Investigative searchers aim to discover gaps in knowledge, create future plans, and transform existing information into a new framework or form.

1.4 Related Surveys

There have been several existing surveys that aim to better position and unify the SAL research agenda.

Rieh *et al.* (2016) position SAL research across three main categories: (1) studies that explore search behavior in learning environments; (2) studies aimed at improving the search skills of students; and (3) studies aimed at developing search environments that improve learning outcomes and experiences. Most SAL studies are motivated (directly or indirectly) by the vision of search systems that better support learning.

The International Workshop on Investigating Learning during Web Search (IWILDS) has been held for several years (Hoppe *et al.*, 2020; Hoppe *et al.*, 2021; Hoppe *et al.*, 2022). Topics presented at the workshop have included search algorithms to improve learning, as well as methods for capturing self-regulated learning (SRL) processes during a SAL study.

von Hoyer *et al.* (2022b) propose the so-called "spaceship" model of SAL. The end goal was to provide a vision of SAL that brings together ideas from information retrieval, education, and psychology. In particular, the model contains several key components: (1) the learner's context; (2) the learner; (3) the interface; (4) the IR backend; and (5) the collection of online resources. The model emphasizes the importance of self-regulated learning (SRL) in SAL. von Hoyer *et al.* (2022b) highlight the need for search systems to better support metacognitive monitoring and metacognitive control while learning during search.

Smith *et al.* (2022) envisioned a multi-component search environment to help students learn in the context of a school assignment. The framework involves modules that model the topical domain, the assignment, the learner's existing state of knowledge, the learner's past search behaviors and learning strategies, and the document corpus. These modules dynamically update each other when new evidence becomes available and influence the retrieval model so that the learner can engage with information that is relevant to the assignment, novel, and at the right level of complexity given their existing knowledge state.

Both proposed frameworks from von Hoyer *et al.* (2022b) and Smith *et al.* (2022) highlight the importance of SRL, the learner's context, and the dynamic adaptation of the search environment based on a learner's goals and progress.

1.5 Outline

In the sections that follow, we survey prior work relevant to SAL and propose directions for future research.

Section 2: SAL research is concerned with scenarios in which a user interacts with a search system to achieve a specific learning objective. Therefore, an important question is: How do learning objectives vary? In Section 2, we explore how learning objectives have been characterized in prior work. Much of this work originates from the field of education. Education researchers have proposed different taxonomies to define learning objectives. These taxonomies were developed to help teachers more clearly define learning objectives for students and to ensure that instructional activities and assessment methods align with the objectives. For example, if a teacher wants their students to be able to do XYZ, then the instructional activities should align with this goal. Similarly, to determine whether the instructional activities were successful, the learning assessment should test the students' ability to do XYZ. SAL researchers have leveraged these taxonomies of learning to systematically manipulate learning-oriented search tasks and to study the effects of those manipulations on different types of outcomes (e.g., behaviors, perceptions, challenges, etc.).

Section 3: SAL studies rely on measuring how much someone learned during a search session. As it turns out, there are many ways to do this. In Section 3, we review the different learning assessment methods that have been used in prior work. Importantly, we discuss the benefits and drawbacks of each method. To illustrate, multiple-choice assessments are easy to grade but may not capture everything someone learned. On the other hand, open-ended assessments—asking participants to describe what they learned—have a broader scope but can be difficult to grade. Additionally, we detail how past work has accounted for prior knowledge in order to measure knowledge *gains* during a search session. Finally, we propose directions for future work. For example, we argue that future work should consider knowledge retention (i.e., being able to use what was learned in the long term) and transfer of learning (i.e., being able to apply what was learned in a new context).

Section 4: SAL studies have explored how different factors may impact learning during search. In Section 4, we survey prior work that has investigated factors related to: (1) the search task or learning objective and (2) the individual searcher. With respect to the search task, most work has focused on the complexity of the task. With respect to the individual searcher, studies have focused on prior knowledge and specific cognitive abilities, such as working memory capacity, perceptual speed, and an individual's tendency to become distracted while working on a task.

Section 5: SAL researchers are interested in developing search environments that encourage and support learning. Therefore, an important question is: How can we *automatically* determine whether an existing system is helping users learn? In Section 5, we survey studies that have investigated whether and how specific search behaviors can predict learning during search. The idea is to predict learning using measures that can be easily logged by a search system.

Section 6: In Section 6, we survey SAL studies that have explored how different system features and tools can support learning during search. For example, studies have considered features of the search interface (e.g., visualizing the coverage of subtopics throughout the search session), as well as peripheral tools for annotating documents or taking notes.

Section 7: When someone searches to learn, they are in control of their learning process. That is, there is no human tutor instructing the searcher on what to do, when, and how. In education and the learning sciences, self-regulated learning (SRL) is a field of study that examines how people learn on their own. It examines the types of mental processes that lead to successfully achieving learning goals. SRL processes include setting goals, enacting effective strategies to achieve the goals, monitoring progress toward the goals, and making adjustments when necessary. In Section 7, we introduce SRL, describe different SRL models that have been proposed, and delve deeply into the Winne and Hadwin model of SRL (Winne and Hadwin, 1998). Goal-setting is a critical phase of the SRL process. Therefore, we also review prior work on the effects of goal-setting on learning and on the characteristics of goals that improve performance. Finally, we describe methods for capturing SRL processes during search. We argue that SAL researchers should more carefully study SRL processes during search and think about ways to support effective SRL toward meaningful learning.

Section 8: Finally, in Section 8, we propose future directions for SAL research. We discuss eight general areas for future work to consider: (1) transfer of learning, (2) designing context-aware SAL environments, (3) investigating long-term SAL processes through longitudinal research, (4) self-determined learning, (5) learning within highly debated topics, (6) scaffolding to encourage and support self-regulated learning (SRL) processes, (7) leveraging generative AI technologies to develop new features to help searchers learn, and (8) studying how groups of individuals learn together.

In this monograph, the tone shifts from descriptive to persuasive in certain sections to align with their distinct purposes. In Section 7 on SRL, the persuasive tone is grounded in two key considerations. First, a large body of empirical evidence from the learning sciences demonstrates that effective SRL significantly improves learning outcomes. Second, despite these established findings, SRL has not been adequately integrated into the theoretical frameworks or methodologies used in SAL research, representing a critical area of opportunity.

Similarly, the tonal shift in Section 8 on future directions reflects our intent not only to synthesize and lay a foundation of what has been done, but also to advocate for and highlight pressing gaps and research needs. While this monograph does not claim to cover all possible directions, it emphasizes those the collective research community has identified as impactful through existing work, aiming to inform and inspire future research in the field.

1.6 Target Audience and Reading Tips

Who is this monograph intended for? There are several audiences that may benefit. Certainly, we intended this monograph to be useful for information retrieval (IR) researchers who are *new* to SAL research. For example, a graduate student looking for a research topic related to SAL should benefit from learning about what has been done and what open questions remain.

Additionally, researchers *already* conducting SAL research should also benefit. For example, several sections may benefit a researcher planning a SAL study. Section 2 may provide ideas on how to systematically manipulate learning-oriented search tasks assigned to participants. Section 3 may provide ideas about how to measure learning during search. Section 6 and Section 7 may provide ideas about novel tools to support learning during search. Specifically, Section 7 may provide ideas about tools to both encourage and support self-regulated learning (SRL) processes that have been empirically shown to improve learning.

Finally, we also intended the monograph to be useful and interesting for researchers *outside* of IR. Researchers in education and cognitive science may find it interesting to see how IR researchers have investigated learning during search. SAL research is inherently multidisciplinary. We hope for this monograph to grab the attention of non-IR researchers. Multiple voices and perspectives may help SAL researchers avoid "reinventing the wheel", employ the best methods, and pursue the most impactful research directions.

Another important question is: What is the best way to read this monograph? We intentionally wrote each section to be self-contained. For example, the same study may be referenced in different sections for different reasons. Section 2 may discuss how the study manipulated learning objectives, Section 3 may discuss how the study measured learning, and Section 6 may describe the novel tools that were used to support learning during search. Therefore, we encourage readers to focus on those sections most interesting to them.

Finally, some readers may find some sections to be written in greater detail than others. For example, Section 7 describes prior research in SRL in great detail. This was done intentionally, as we believe that supporting effective SRL is an exciting area for future SAL research to consider.

2

Characterizing Learning

SAL research is concerned with scenarios in which someone is searching for information in order to learn about a topic. This often involves a specific learning objective: “I need to find information so that I am able to do <learning objective>.” Learning objectives vary greatly in their complexity, level of abstraction, topical domain, and more. Because of this variation, it is important to understand how characteristics of the learning objective impact search behaviors, challenges, and learning outcomes. To investigate these effects, we must systematically characterize learning objectives. In SAL, different frameworks have been used to inform the characterization (and assessment) of learning during search. In this section, we provide an overview of the frameworks that have been used in SAL to characterize learning objectives.[1] Then, we discuss research in SAL that has investigated the impact of learning objectives on search behaviors and learning processes during search.

[1]Later, in Section 7, we discuss self-regulated learning, which aims to characterize the cognitive and metacognitive processes that people engage in while pursuing a learning objective.

2.1 Bloom's Taxonomy

Bloom's taxonomy was introduced in 1956 (Bloom, 1956). It was introduced to help educators more clearly define learning objectives for students and to define important educational terminology (Conklin, 2005). At the time, there was no consensus on what it means for a person to "know." The verb "know" was used by educators to mean very different things. "One person might use 'know' to mean remembering some facts [and] another educator might mean that a person must really 'know' an entire discipline in all its complexity, modes of inquiry, scope, and sequence" (Conklin, 2005, p. 156). Bloom's taxonomy offered educators a common language, enabling them to be more specific when discussing learning objectives.

As shown in Figure 2.1, Bloom's taxonomy consists of six levels. Each level can be understood in terms of what a student should be able to do with their newly acquired knowledge:

1. *Knowledge*: At this level, the student should be able to recall specific facts, terminology, concepts, and methods (e.g., recalling molecular biology terms or historical dates and events).

2. *Comprehension*: At this level, the student should be able to summarize information in their own words, interpret information, and make inferences (e.g., inferring that an increase in sales may be due to the holiday season).

3. *Application*: At this level, the student should be able to use knowledge to complete a task or solve a problem (e.g., writing exam questions that exemplify each level of Bloom's taxonomy).

4. *Analysis*: At this level, the student should be able to relate different elements, which may include understanding their similarities, differences, and relationships (e.g., assessing the similarities and differences between the art movements of surrealism and dadaism).

5. *Synthesis*: At this level, the student should be able to assemble elements into a new whole (e.g., recognizing new connections between different theories).

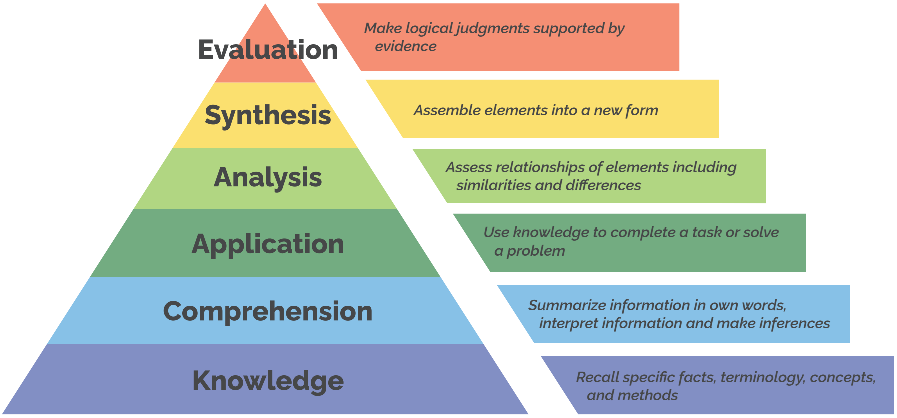

Figure 2.1: Bloom's taxonomy.

6. *Evaluation*: At this level, the student should be able to make judgments using evidence (e.g., judging which method is most effective for reducing energy consumption).

The levels in Bloom's taxonomy are organized hierarchically. This means that meeting an objective at a particular level of the taxonomy is likely to require processes from the lower levels. For example, suppose we want a student to be able to explain the similarities and differences between diffusion and osmosis. This objective is at the level of *analysis* in Bloom's taxonomy. In this respect, it is likely to require processes at the level of *knowledge*, *comprehension*, and *application*. In other words, to understand the similarities and differences between diffusion and osmosis, a student should be able to: (1) recall important terminology (*knowledge*), (2) explain the concepts in their own words (*comprehension*), and (3) use the concepts to explain different real-world phenomena (*application*). Being able to engage in these lower-level processes is a prerequisite to being able to engage in analysis-level processes.

Bloom's taxonomy can help us determine which pre-requisite processes are needed to achieve an objective. Additionally, it can also help us determine which pre-requisite processes are *most* important. In the example above, being able to engage in application-level processes (one level below analysis) is arguably the most important. That is, being able to apply the concepts of diffusion and osmosis to explain different real-world phenomena is especially helpful in recognizing their similarities and differences. For example, diffusion explains the smell of perfume dissipating in a room. Conversely, osmosis explains the rehydration of lettuce when submerged in water. This highlights the fact that osmosis is a special type of diffusion in which a solvent (water) passes through a semipermeable membrane to diffuse a solute on the other side of the membrane.

Bloom's taxonomy highlighted the importance of teaching specific, targeted objectives in order for educational professionals to discuss issues around curriculum and assessment with more clarity. Additionally, the taxonomy emphasized the need for an educator to identify an objective's *relative* ranking as compared with other objectives. This enabled educators to better understand which *prerequisite* objectives are likely to be useful in pursuing a *target* objective. Bloom's taxonomy was highly influential in that it introduced a practical method for specifying objectives and ensuring that a range of hierarchically lower- *and* higher-level skills were also being taught and assessed.

Key Takeaway

Bloom's taxonomy provided a framework for educators to teach specific, targeted objectives that were spread across a range of learning skills.

According to Krathwohl (2002), Bloom believed that the taxonomy could be useful to: (1) provide a common language for educators to communicate across grade levels and domains; (2) determine the meaning of broad, high-level educational goals; (3) assess the congruence between learning objectives, instructional activities, and assessment materials; and (4) determine the range of breadth and depth of a particular course.

Despite its important contributions, Bloom's taxonomy was practically narrow. Bloom (1956) only applied the taxonomy to learning assessment. That is, all the examples provided for each level of the taxonomy in Bloom (1956) are in the form of exam questions. Over time, it became apparent that a revision of the taxonomy was necessary to address both the advances in our understanding of human learning and to better support educators who were interested in implementing the taxonomy to develop learning objectives, instructional activities, along with assessment.

2.2 Anderson and Krathwohl Taxonomy

In the subsequent decades following Bloom's taxonomy, research and understanding about human learning evolved. There was a re-framing of learning around *how* learners learn (i.e., active cognitive processes) and how learners think about their own cognition (i.e., metacognitive processes) (Conklin, 2005). This evolution led education researchers and practitioners to call for an updated version of Bloom's taxonomy. The resulting taxonomy, called the Anderson & Krathwohl (A&K) taxonomy, was written by Lorin Anderson and David Krathwohl (an author of the original Bloom's taxonomy).

Like Bloom's taxonomy, the A&K taxonomy was designed to help educators more precisely define learning objectives (Anderson *et al.*, 2000). Different from Bloom's taxonomy, the A&K taxonomy provided thorough examples in the context of teaching to help educators use the taxonomy. Anderson & Krathwohl provided sample "vignettes" to support educators in developing assessments, but also specific activities and curricula to support the achievement of particular learning objectives. In this way, the A&K taxonomy supported the alignment of learning objectives with instructional activities and methods of assessment.

The A&K taxonomy modifies Bloom's taxonomy in three ways. First, cognitive processes are articulated as verbs rather than nouns (e.g., analyze versus analysis). Second, the *synthesis* cognitive process is renamed as *create* and is situated as the most complex process. Third, and most importantly, a second dimension is added to the taxonomy to categorize the knowledge type of the objective. While the cognitive process dimension relates to *how deeply* something is learned, the knowledge type dimension relates to *what* is being learned.

The A&K taxonomy, shown in Table 2.1, situates learning objectives at the intersection of two orthogonal dimensions: the knowledge type dimension and the cognitive process dimension. Each learning objective (and learning activity and assessment) is positioned in a single cell of the table at the intersection of a particular knowledge type and cognitive process.

Table 2.1: The Anderson and Krathwohl taxonomy of learning.

Knowledge Type	Cognitive Process					
	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual						
Conceptual						
Procedural						
Metacognitive						

Key Takeaway

The A&K taxonomy built upon Bloom's taxonomy, creating a two-dimensional taxonomy that situates learning objectives within both cognitive process and knowledge type dimensions

The *knowledge type* dimension defines the type of knowledge involved in the learning objective. This dimension is made up of four types of knowledge: (1) factual knowledge (i.e. isolated, declarative bits of information); (2) conceptual knowledge (i.e., concepts, principles, models, and schemas); (3) procedural knowledge (i.e., steps about how to do something); and (4) metacognitive knowledge (i.e., knowledge about one's own cognition or cognition in general).

The *cognitive process* dimension defines the cognitive process involved in the learning objective. In all, there are six cognitive processes ranging from simple to complex: (1) remember (e.g., recall and repeat verbatim); (2) understand (e.g., summarize, exemplify); (3) apply (e.g., use); (4) analyze (e.g., differentiate, break apart); (5) evaluate (e.g., judge, critique, prioritize); and (6) create (e.g., generate a novel representation).

Both dimensions of the A&K taxonomy run along different continuums. It is easy to see that the cognitive process dimension runs from simple to complex. The knowledge type dimension is slightly more nuanced. Anderson *et al.* (2000) argued that the knowledge type dimension runs along a continuum from concrete to abstract, with factual knowledge being the most concrete and metacognitive knowledge being the most abstract. However, this continuum is not perfectly linear. Facts are definitely more concrete than concepts. Facts are often about things that can be perceived by the senses or are linked to past events. As noted by A&K, "for the most part, factual knowledge exists at a relatively low level of abstraction" (Anderson *et al.*, 2000, p. 42). Conversely, concepts (e.g., totalitarianism, surrealism) are often used to organize bodies of knowledge and may have hazy boundaries. Compared to facts and concepts, procedures can be either concrete or abstract. Physical procedures (e.g., cooking pasta) tend to be concrete. However, mental procedures (e.g., computing the derivative of a function) tend to be more abstract. Metacognitive knowledge is knowledge about cognition and is therefore highly abstract.

Prior work in SAL has also argued that the knowledge type dimension runs along continuums from isolated to interrelated and objective to subjective (Urgo *et al.*, 2020). Compared to factual knowledge, conceptual and procedural knowledge are more internally interrelated. As noted by A&K, "factual knowledge can be isolated as elements or bits of information that are believed to have value in and of themselves" (Anderson *et al.*, 2000, p. 42). Conversely, concepts usually exist in a broader framework. In this respect, mastering a concept may require understanding how it relates to other concepts. For example, understanding one artistic style might require understanding how it relates to other artistic styles. Similarly, understanding a procedure may require understanding procedures that are nested within or alternative procedures for performing the same task. For example, understanding how to cook lasagna requires understanding how to boil pasta.

In terms of subjectivity, facts are objective bits of information. Facts are presumed to be true regardless of anyone's opinion. Conversely, concepts and procedures may involve more subjectivity. For example, determining whether a work of art exemplifies an artistic style or determining whether a procedure can be used to solve a problem may involve subjective judgement.

Categorizing a learning objective with the A&K taxonomy involves situating the objective at the intersection of these two orthogonal dimensions. The cognitive process is represented by the "verb" of the objective and the knowledge type is represented by the "noun" of the objective. For example, consider the following learning objective: Recall the height of the Empire State Building. Here, *recall* is the verb and *the height of the Empire State Building* is the noun. This learning objective is categorized as *remember/factual* because to recall is a remember cognitive process and the height of a building is factual knowledge. Consider a second example: Differentiate between Baroque music and Early Romantic music. In this example, *differentiate* is the verb and *Baroque music and Early Romantic music* are the nouns. This learning objective is categorized as *analyze/conceptual* because to differentiate is an analyze cognitive process and the Baroque and Early Romantic musical eras are examples of conceptual knowledge. Finally, consider the example: Invent a new method for finding the mathematical center of a circle. In this example, *invent* is the verb and *a method for finding the mathematical center of a circle* is the noun. This learning objective is categorized as *create/procedural* because to invent is a create cognitive process and a method for finding the center of a circle is procedural knowledge. In prior SAL work, Urgo *et al.* (2019) developed 24 learning-oriented search tasks across all cells from A&K's taxonomy (see Table 1 in Urgo *et al.*, 2019).

The A&K taxonomy can also be used to develop assessment materials. Specifically, it can be used to develop test questions that measure someone's ability to engage in specific cognitive processes, ranging from simple to complex. Consider a task that prompts participants to: "Learn everything you can about the concepts of diffusion and osmosis." Given that this is a conceptual learning task, test questions could ask participants to recall concepts (remember), describe examples (understand), categorize new examples (apply), compare and contrast concepts (analyze), judge which concept best explains a phenomenon (evaluate), and create a new exam question that tests whether a concept is fully understood (create).

In prior work, researchers have also leveraged the A&K taxonomy to develop new categories. Rieh *et al.* (2016) (building upon Lee *et al.*, 2015) developed the categories of *receptive*, *critical*, and *creative* learning during search. Here, the cognitive processes from A&K's taxonomy are combined into three groups. Receptive learning involves remembering and understanding; critical learning involves applying, analyzing, and evaluating; and creative learning involves creating. Rieh *et al.* (2016) argued that existing search systems support receptive learning, but that novel tools are needed to support critical and creative learning. Liu *et al.* (2019) investigated when and how learning happens during receptive versus critical learning tasks.[2]

Anderson & Krathwohl developed their taxonomy to help educators aim for *meaningful learning*. Anderson & Krathwohl define meaningful learning as: (1) the student possesses relevant knowledge; (2) the student can use their knowledge in contexts *similar* to those encountered during the learning process; and (3) the student can use their knowledge in contexts *different* from those encountered during the learning process. Prior SAL studies have leveraged the A&K taxonomy to measure whether participants achieve the first two criteria associated with meaningful learning. However, little work has explored whether participants achieve the third criterion. In Section 8, we advocate that future SAL studies should investigate transfer of learning—the learner's ability to use their new knowledge in novel contexts. Research has shown that students often fail to recognize that they possess knowledge that is relevant to a new situation. Additionally, research has shown that depth of understanding is a prerequisite for successful knowledge transfer. That is, successful transfer is strong evidence that something is deeply understood.

[2]We discuss results from this study in Section 2.3.1.

Key Takeaway

Meaningful learning indicates that the learner possesses and effectively uses relevant knowledge in familiar and novel contexts.

2.3 Using the A&K Taxonomy to Manipulate Search Tasks

As previously mentioned, the A&K taxonomy was developed to help educators define learning objectives for students and to help align learning assessment materials with learning objectives. In an analogous way, SAL studies have used the A&K taxonomy for two main purposes: (1) to manipulate learning-oriented search tasks and (2) to assess learning during search. In the next two sections, we review studies that have used the A&K taxonomy to manipulate search tasks and study the effects of such manipulations on search behaviors and perceptions. All of Section 3 is devoted to learning assessment methods. In that section, we discuss how different types of learning assessments can be used to measure a learner's ability to engage in specific cognitive processes before and after searching.

While the A&K taxonomy has two orthogonal dimensions (i.e., cognitive process and knowledge type), most studies have only leveraged the cognitive process dimension to manipulate search tasks. In Section 2.3.1, we survey studies that have only manipulated search tasks along the cognitive process dimension. In Section 2.3.2, we survey studies that have manipulated search tasks along the knowledge type dimension or both dimensions of the A&K taxonomy.

It is important to note that most of the studies in the next two sections did *not* measure learning during search. Instead, they simply manipulated search tasks to study the effects on search behaviors and perceptions. We decided to include these studies because they leveraged a taxonomy of learning to manipulate search tasks.

2.3.1 Manipulating the Cognitive Process

Several studies have manipulated search tasks along the cognitive process dimension of A&K's taxonomy. A&K's taxonomy has six cognitive processes, ranging from simplest (i.e., remember) to most complex (i.e., create). Therefore, manipulating search tasks along the cognitive process dimension has often been framed as manipulating the *cognitive complexity* (or simply the complexity) of search tasks.

To our knowledge, Jansen *et al.* (2009) was the first study to manipulate search tasks using A&K's cognitive processes dimension. Search tasks varied along all six cognitive processes. Interestingly, results found an inverted-U shaped trend between task complexity and search activity. Mid-complexity tasks required more search activity than the simplest and most complex tasks. The authors argued that participants may have relied more heavily on their prior knowledge (versus searching for information) during the most complex tasks.

Arguello *et al.* (2012) manipulated tasks along three cognitive processes: remember, understand, and analyze. The study considered two aggregated search interfaces that provided access to results from different back-end systems or verticals: web, images, video, news, blogs, community Q&A, and shopping. The non-blended interface provided access to different verticals through different tabs and the blended interface also showcased the top results from each vertical on the main SERP. Complex tasks required more search activity. Additionally, complex tasks had greater use of non-web results, but only with the blended interface. That is, the same trend was not observed when the top vertical results were *not* showcased on the main SERP.

Brennan *et al.* (2014) manipulated tasks along three cognitive processes: remember, analyze, and create. Complex tasks were associated with higher levels of workload and required more search activity (e.g., more queries, more clicks, and longer search sessions).

Kelly *et al.* (2015) manipulated search tasks along five cognitive processes: remember, understand, analyze, evaluate, create. Results found several trends. First, complex tasks were perceived to be more difficult but also more interesting and engaging. Second, complex tasks required more search activity (e.g., more queries, more clicks, and longer sessions). Third, complex tasks had more divergent search behaviors. Querying and clicking behaviors were compared across participants who completed the same task. During complex tasks, participants issued more queries not issued by other participants; used more query terms not used by other participants; and clicked on more URLs not clicked by other participants. In other words, participants used similar search strategies during simple tasks and different search strategies during complex tasks.

Capra *et al.* (2015) manipulated search tasks along four cognitive processes: remember, understand, analyze, and evaluate. The study investigated the effects of task complexity on participants' use of an auxiliary tool that displayed the search paths followed by other searchers for the same task. Participants were more likely to interact with the tool during complex tasks. Additionally, task complexity impacted participants' motivations for engaging with the tool. During simple tasks, participants engaged with the tool to verify information found on their own. During complex tasks, participants engaged with the tool to find new sources of information or new search strategies.

Choi *et al.* (2019b) manipulated search tasks along four cognitive processes: remember, understand, analyze, and create. The study investigated the effects of task complexity on participants' use of an auxiliary search tool that displayed facts, concepts, opinions, and insights related to the search task. Insights were defined as task-related tips and advice from people who completed the same task. During complex tasks, participants reported less demand for facts and greater demand for concepts, opinions, and insights. Additionally, task complexity impacted participants' motivation for engaging with the auxiliary tool. Similar to Capra *et al.* (2015), during simple tasks, participants engaged with the tool to verify information found on their own. Conversely, during complex tasks, participants engaged with the tool to gain an overview of the task topic.

Choi and Arguello (2020) investigated the functional role of information during tasks that varied along four cognitive processes: remember, understand, analyze, and create. The authors considered three functional roles of information. Problem information helps the searcher understand the task requirements; problem-solving information helps the searcher strategize on how to perform the task; and domain information can be directly applied to the task solution. Results found that all tasks required domain information. However, complex tasks required more problem and problem-solving information than simple tasks.

Hu and Kando (2017) studied task complexity in the context of music information retrieval and manipulated tasks along three cognitive processes: remember, understand, and analyze. Complex tasks were associated with higher perceptions of difficulty and lower satisfaction. Additionally, complex tasks were associated with lower levels of success (e.g., fewer songs found).

Chi (2019) had participants complete health-related information-seeking tasks that varied across three cognitive processes: understand, analyze, and evaluate. During complex tasks (particularly evaluate), participants visited more sources and visited a more diverse range of sources.

Kalyani and Gadiraju (2019) manipulated search tasks along all six cognitive processes from A&K's taxonomy. Learning assessments were developed with questions to match the cognitive complexity of the task. Knowledge gains occurred during all tasks. However, knowledge gains were greater for apply versus analyze tasks, possibly because applying is a less complex process than analyzing. Additionally, complex tasks required more search activity (e.g., more queries, longer queries, more unique query terms, and more pages visited).

Ghosh *et al.* (2018) conducted a diary study that tracked participants engaged in four learning-oriented search tasks over a two-week period. Tasks varied along four complexity levels: remember/understand (treated as one category), apply, analyze, and evaluate. Results found several trends. First, across all tasks, participants reported higher topical knowledge after the task. Second, as in Kelly *et al.* (2015), complex tasks were perceived to be more difficult but also more interesting. Finally, participants were asked to select verbs describing the cognitive processes that they engaged in during each task. Participants selected verbs such as "list" for the remember/understand task; "demonstrate" for the apply task; and "relate" for the analyze and evaluate tasks.

Liu *et al.* (2019) used mind maps to better understand how task complexity impacts learning during the search process. Participants completed a *receptive* task that involved gaining a deeper understanding of a topic (i.e., understand) and a *critical* task that involved evaluating ideas from different perspectives (i.e., evaluate). Participants were instructed to generate mind maps before searching and to modify their mind maps during their search. During receptive tasks, participants continuously updated their mind maps throughout the search session. In contrast, during critical tasks, they tended to modify their mind maps toward the end of the search session. This result suggests that task complexity impacts how knowledge shifts during the search session. Knowledge shifts are more evenly distributed for simpler tasks and are more likely to occur toward the end of the session for complex tasks.

Key Takeaway

Studies have leveraged the cognitive process dimension from A&K's taxonomy to manipulate search tasks. Studies have mostly found that complex tasks are perceived as more difficult and require more search activity.

2.3.2 Manipulating Knowledge Type

So far, we have touched upon studies that leveraged the cognitive process dimension from A&K's taxonomy to manipulate learning-oriented search tasks. Fewer studies have manipulated search tasks along the knowledge type dimension.

Urgo *et al.* (2020) manipulated tasks along both dimensions of A&K's taxonomy. Participants completed search tasks that varied along three cognitive processes (i.e., apply, evaluate, create) and three knowledge types (i.e., factual, conceptual, procedural). Interestingly, the task's knowledge type had much stronger effects on participants' perceptions and behaviors than the task's cognitive process. First, as expected, participants anticipated needing facts for factual tasks, concepts for conceptual tasks, and procedures for procedural tasks. Second, and more interestingly, the task's knowledge type affected the *cognitive processes* that participants engaged in during the task. During conceptual tasks, participants reported on engaging in more understanding (e.g., understanding concepts) and analyzing (e.g., distinguishing between related concepts). One possibility is that concepts typically exist in a broader framework, requiring not only understanding but also analyzing (e.g., differentiating). For example, understanding "parliamentary democracy" may require distinguishing it from "presidential democracy." During procedural tasks, participants reported on engaging in more applying, evaluating, and creating. It is quite reasonable for procedural tasks to involve applying—using a procedure to solve a problem. It is less obvious why procedural tasks may also involve evaluating and creating. The authors hypothesized that procedural tasks involve choosing between different approaches to a problem (i.e., evaluating) and modifying an approach based on personal preferences or constraints (i.e., creating). Finally, conceptual tasks were perceived to be the most difficult and required the most search activity (e.g., more queries, more abandoned queries, and longer sessions).

In a follow-up paper, Urgo and Arguello (2022c) reported on a qualitative analysis of search sessions during the same study. The goal was to investigate the effects of the task's learning objective (i.e., cognitive process and knowledge type) on the types of cognitive processes that people engage in while searching to learn. Specifically, the authors analyzed common transitions between cognitive processes during the search session. Results found that some transitions were more common than others depending on the learning objective's knowledge type (i.e., factual vs. conceptual vs. procedural). During factual objectives, participants tended to start with remember and "downshift" to remember from more complex processes. Conversely, during non-factual tasks, participants tended to start with understand and "downshift" to understand from more complex processes. During conceptual objectives, participants were more likely to transition from analyze to evaluate and vice-versa. Finally, during procedural objectives, participants were more likely to transition from simple to complex processes. Additionally, procedural objectives had more create-to-create transitions. These create-to-create transitions were observed when participants iteratively modified a procedure according to specific task constraints or personal preferences.

Pardi *et al.* (2023) conducted two studies in which participants completed four different task types: (1) causal conceptual, (2) relational conceptual, (3) sensorimotor procedural, and (4) cognitive procedural. A *causal conceptual* task involves concepts that share a cause-and-effect relationship. A *relational conceptual* task involves concepts that are related but do not share a cause-and-effect relationship. A *sensorimotor procedural* task involves physical objects and actions. A *cognitive procedural* task involves mental step-by-step processes.

Both studies investigated the effects of the task type on the modalities preferred by participants. Study 1 investigated four modalities (videos, images, text with images, and text only). Study 2 investigated only the two modalities that were preferred by participants during Study 1 (videos and text with images). Both studies found that the knowledge type of the task (i.e., conceptual versus procedural) is not enough to determine which modalities searchers prefer. In terms of conceptual tasks, videos were preferred for causal tasks but not relational tasks, suggesting that videos are particularly helpful in explaining cause-and-effect relationships between concepts. In terms of procedural tasks, videos were preferred for sensorimotor tasks but not cognitive tasks, suggesting that videos are particularly helpful during procedural tasks that require learning about spatiotemporal changes during the task. For example, participants preferred videos for procedural tasks such as "how to tie a figure-eight knot" and text with images for tasks such as "how to compute the electrical resistance in a parallel circuit."

To our knowledge, SAL studies have not investigated search tasks that involve acquiring new metacognitive knowledge as the target learning objective. As previously noted, metacognitive knowledge involves knowledge about one's own cognition or about cognition in general. Urgo *et al.* (2019) provide examples of metacognitive tasks across different cognitive complexity levels. To illustrate, a *remember/metacognitive* task could be: Recall the steps of a strategy for memorizing information. A *apply/metacognitive* task could be: Use memorization technique XYZ to memorize the first 15 U.S. presidents. An *evaluate/metacognitive* task could be: Compare different memorization techniques and determine which one works best for you and why.

Key Takeaway

Studies suggest that factual, conceptual, and procedural learning involves different processes. Additionally, differences within each knowledge type (e.g., mental procedural vs. physical procedural) may impact the types of information that are most useful.

2.4 Summary

Learning objectives are critical to SAL research. During all SAL studies, participants are given one or more learning objectives to complete. Many studies have systematically manipulated learning objectives. These studies have focused on understanding how characteristics of the learning objective may impact different outcomes (e.g., perceptions and behaviors). Other studies have given participants objectives of the same type (perhaps across different topical domains). Whether studies have manipulated learning objectives or controlled them (i.e., aimed to keep them consistent), it becomes necessary to characterize objectives in a systematic manner.

In this section, we have reviewed two taxonomies that can be used to characterize learning objectives: Bloom's taxonomy and the Anderson & Krathwohl (A&K) taxonomy. The latter is a revision of the previous. These taxonomies were developed by researchers in education to help educators satisfy three aims: (1) more clearly define learning objectives for students, (2) ensure that instructional activities align with the objectives, and (3) ensure that measurements of learning align with the objectives. The first aim is to reduce the ambiguity of learning objectives; the second aim is to create instructional activities that actually align with their intended outcomes; and third aim involves improving the evaluation of the instructional activities.

SAL studies have mostly leveraged the A&K taxonomy to manipulate learning objectives (or learning-oriented search tasks) and to develop assessment materials that measure the extent to which an objective was met. The A&K taxonomy situates learning objectives at the intersection of two orthogonal dimensions: knowledge type and cognitive process. The taxonomy defines four knowledge types: factual, conceptual, procedural, and metacognitive knowledge. The taxonomy defines six cognitive processes that range from simplest to most complex: remember, understand, apply, analyze, evaluate, and create.

In this section, we reviewed studies that have manipulated search tasks using the A&K taxonomy. Results from these studies suggest the following trends. First, complex tasks are usually perceived to be more difficult (although sometimes also more engaging and interesting) and require more search activity. Second, complexity seems to impact how knowledge changes during the search session. Results from one study suggest that simple tasks involve knowledge shifts throughout the session and that complex tasks involve more knowledge shifts toward the end of the session. Third, conceptual learning seems to be more difficult than factual and procedural learning. Fourth, the knowledge type of the objective may influence the types of content preferred by searchers. Additionally, categories such as conceptual versus procedural knowledge may be too coarse to determine which type of content is preferred. For example, one study found that videos are preferred for procedural tasks that are physical, but text with images are preferred for procedural tasks that are largely cerebral. Finally, the knowledge type of the objective seems to impact the cognitive processes that searchers engage in during the search session. For example, conceptual learning tasks seem to involve more understanding and analyzing, while procedural learning tasks seem to involve more applying, evaluating, and creating.

3

Learning Assessment

Learning assessment is a key component in SAL studies. Assessment materials are used to both gauge prior knowledge and measure learning during one or more search sessions. In this section, we provide a systematic review of different types of assessments used in SAL studies to date.

First, we review different types of assessments that have been used in SAL studies and discuss their pros and cons. Then, we provide recommendations for future research. Importantly, we argue that future studies should clearly define learning objectives for participants and use assessment materials that reliably capture the type of learning that is intended. For example, assessment materials should test a participant's ability to engage with specific cognitive processes, which may range from simple (e.g., recall information) to complex (e.g., use new knowledge to make a decision). Additionally, we argue that future studies should consider two dimensions of learning that are understudied in SAL: long-term retention (i.e., being able to use what was learned in the long term) and transfer of learning (i.e., being able to use what was learned in a novel context).

3.1 Self-Report

One way to measure learning is to ask participants directly or indirectly. There are three main types of self-report assessments that have been implemented in SAL research: questionnaires, learning diaries, and interviews.

3.1.1 Questionnaires

Studies have used questionnaire items to ask participants how much they learned during a search session or to rate their topic familiarity before and after searching. Capra *et al.* (2018) asked participants to rate their prior knowledge before each search task and their level of knowledge increase after each task. Collins-Thompson *et al.* (2016) asked participants to rate how much they learned on a scale of 0 to 100 using two questionnaire items. Ghosh *et al.* (2018) asked participants to rate their topic familiarity before each search task and their overall knowledge gains after each task. Liu *et al.* (2013) asked participants to complete a series of subtasks associated with the same general task. After each subtask, participants were asked to rate their familiarity with the topic of the subtask and the general task.

Studies have also used self-assessments of learning in conjunction with other learning assessments (Freund *et al.*, 2016; Heilman *et al.*, 2010; Kammerer *et al.*, 2009; Liu *et al.*, 2019; Wilson and Wilson, 2013; Zhang and Liu, 2020). O'Brien *et al.* (2020) measured perceptions of pre-task prior knowledge and post-task knowledge gains in addition to asking participants to create written summaries before and after each search task. Similarly, Collins-Thompson *et al.* (2016) and Abualsaud (2017) also included a self-reported learning score in addition to short-answer and open-ended learning assessments.

There are several benefits and drawbacks to self-report assessments. In terms of benefits, self-report assessments do not require grading since participants rate their own level of prior knowledge and/or learning. Second, they provide insights into subjective knowledge gains (Collins-Thompson *et al.*, 2016; Abualsaud, 2017). Third, questionnaire items that capture perceptions of prior knowledge and/or learning are easy to develop. Some studies have even used a single questionnaire item (Capra *et al.*, 2018; Ghosh *et al.*, 2018).

In terms of drawbacks, self-assessment may obscure the *types* of learning that took place. For example, questionnaire items may not distinguish between *breadth* and *depth* of learning or a participant's ability to engage in complex cognitive processes using the knowledge gained during the search task. Second, perceptions of learning may not align with actual learning. Some studies have found a correlation between perceptions of learning and actual learning (Collins-Thompson *et al.*, 2016; Abualsaud, 2017). However, other studies have found the opposite (Persky *et al.*, 2020; Pennycook *et al.*, 2017). Studies have found that less knowledgeable individuals tend to overestimate their knowledge (Pennycook *et al.*, 2017) and that men tend to overestimate their knowledge gains more than women (González-Betancor *et al.*, 2019).

3.1.2 Learning Diary and Interviews

Prior work in SAL has also used learning diaries and interviews to explore a searcher's perception of their own learning. Learning objectives (e.g., learning a new skill) can easily span multiple search sessions. A learning diary is an instrument that asks participants to reflect on each learning-oriented search session.

Cole (2022) asked participants to complete a learning diary entry after each search session within a 5-day period. Each entry required participants to answer a series of closed- and open-ended questions about the search session. One closed-ended question asked: "How much do you think you learned during this session?" An open-ended question asked: "Explain why you did or did not learn during your search session." Diary entries were later used to interview participants about specific search sessions.

Learning diaries have two main benefits. First, they can be used to capture participants' perceptions of their learning progress across multiple search sessions. Second, they can be used during an exit interview to jog participants' memories about their goals, outcomes, and feelings during a specific search session. During a longitudinal study, asking such questions immediately after each session runs the risk of altering behaviors during subsequent sessions.

Learning diaries have two main drawbacks. First, similar to questionnaire items, learning diaries capture *perceptions* of learning, which may deviate from how much was actually learned during a search session. Second, asking participants to reflect on their learning (e.g., "Explain why you learn or did not learn.") may unintentionally serve as a form of scaffolding. That is, participants may reflect on strategies that did or did not produce the expected results, which may influence behaviors during future search sessions.

Key Takeaway

Self-report data is easy to collect. However, perceptions of learning may not accurately reflect objective learning.

3.2 Implicit Measures

Implicit measures aim to detect learning using behavioral measures captured during the search session. Chi *et al.* (2016) experimented with two implicit measures: query complexity and click complexity. Both measures capture the extent to which participants issued queries and clicked on search results that were rarely issued/clicked by other participants during the same task.

Query complexity (QC) is defined as:

$$\mathbf{QC}(q) = \log \frac{N}{N_q},$$

where N denotes the total number of study participants who completed the task and N_q denotes number of participants who issued query q. Similarly, click complexity (CC) is defined as:

$$\mathbf{CC}(d) = \log \frac{N}{N_d},$$

where N_d denotes the number of participants who clicked on document d. Both measures assume that rare queries and clicks suggest greater knowledge. Chi *et al.* (2016) observed increases in query and click complexity at later stages during the search process, which was interpreted as evidence of learning.

Implicit measures of learning have two main benefits. First, implicit measures are generated from behaviors captured by the system and are therefore easy to compute. Second, implicit measures can be computed "on the fly" and can therefore be shown to searchers as a form of feedback about their learning during the search session.

Implicit measures of learning have two main drawbacks. First, implicit measures lack specificity. It is unclear what type of learning is being measured by implicit measures such as query and click complexity. Second, to some extent, implicit measures lack validity. Additional research is needed to compare implicit measures of learning with other measures of learning.

3.3 Closed-ended Assessments

Closed-ended assessments ask questions with pre-defined correct answers. Prior SAL studies have used two types of close-ended assessments: multiple-choice and short-answer.

3.3.1 Multiple-Choice

Multiple-choice assessments involve questions with a closed set of correct and incorrect options.

Freund *et al.* (2016) asked participants to read three articles and measured reading comprehension using two types of multiple-choice questions. First, participants were asked whether a provided statement accurately summarized the theme or position of a specific article. Second, participants were provided with six statements and were asked to select which three accurately conveyed themes present in all three articles.

Several studies have used multiple-choice items that asked participants to select "true" or "false" (or possibly "I don't know") in response to a statement (Gadiraju *et al.*, 2018; Yu *et al.*, 2018; Xu *et al.*, 2020; Qiu *et al.*, 2020; Kalyani and Gadiraju, 2019; Nelson *et al.*, 2009; Salmerón *et al.*, 2020). Studies have also used questions with a set of correct and incorrect options (Kalyani and Gadiraju, 2019; Hoyer *et al.*, 2019; Syed and Collins-Thompson, 2017a; Weingart and Eickhoff, 2016; Davies *et al.*, 2013; Urgo and Arguello, 2023; Urgo and Arguello, 2024). Kalyani and Gadiraju (2019) included multiple-choice questions that asked participants to categorize a set of items and order a series of events in their correct sequence based on their causal relations. To measure vocabulary learning, studies have also used fill-in-the-blank sentences with a set of vocabulary words as choices (Heilman and Eskenazi, 2006; Heilman *et al.*, 2010)

Multiple-choice assessments have three main benefits. First, multiple-choice tests have predefined correct answers and are therefore easy to grade. Second, multiple-choice tests are quick to administer because participants are only required to select the correct answers and not write their own responses. Finally, because they can be graded automatically, multiple-choice items can be used to provide feedback to searchers about their learning during the search session.

Multiple-choice assessments, however, also have several drawbacks. First, because the answer options are provided, participants may guess correctly, which may overestimate their knowledge gains. Second, multiple-choice tests have limited coverage and may not capture everything that someone learned about a topic. Finally, depending on their design, multiple-choice assessments may not test whether the learner is able to engage in complex cognitive processes. Some multiple-choice tests (e.g., true-or-false tests) may be more susceptible to this than others (e.g., "order the following events in their natural sequence"). Finally, multiple-choice tests are difficult to design. They should include questions and incorrect answer options that are grounded in common misconceptions within the domain of the search task. Ideally, they should be designed (or at least validated) by domain experts. To mitigate this drawback, one alternative is to use multiple-choice tests that has already been validated and to design a learning-oriented search task around the domain of the test. Urgo and Arguello (2023) and Urgo and Arguello (2024) adopted this approach and used a validated multiple-choice test known as the Osmosis and Diffusion Conceptual

Assessment (ODCA) (Fisher *et al.*, 2011). In terms of the search task, participants were asked to "learn everything you can about osmosis and diffusion."

Key Takeaway

Multiple-choice tests are easy to grade. However, they should be developed by domain experts that understand common misconceptions about the task topic.

3.3.2 Short-Answer

Short-answer assessments involve asking questions that are open-ended but have an objectively correct answer that is relatively short. Short-answer questions do not include a set of options for participants to choose from.

Many SAL studies have measured learning using short-answer questions (Hersh *et al.*, 1995; Hornbæk and Frøkjær, 2003; Moraes *et al.*, 2018; Roy *et al.*, 2020; Câmara *et al.*, 2021; Roy *et al.*, 2021; Davies *et al.*, 2013; Collins-Thompson *et al.*, 2016). Short-answer questions can have different formats. For example, several studies have asked participants to define concepts in their own words (Moraes *et al.*, 2018; Roy *et al.*, 2020; Câmara *et al.*, 2021; Roy *et al.*, 2021). Davies *et al.* (2013) had participants learn about plate tectonics. The short-answer questions consisted of diagrams with blank spaces for participants to label with the correct components or processes depicted in the picture.

Short-answer assessments have two main benefits. First, short-answer questions have a predefined correct answer and are therefore easy to grade. Second, participants are not provided with options to choose from. Therefore, compared to multiple-choice assessments, short-answer assessments are less susceptible to guesswork.

Short-answer assessments have three main drawbacks. First, like multiple-choice assessments, they have limited coverage. That is, they may not capture everything that a participant learned about a topic. Second, like multiple-choice assessments, they are difficult to design. The questions should target common misconceptions and should ideally be developed (or at least validated) by domain experts. Third, depending on their design, short-answer questions may not assess a participant's ability to engage in complex cognitive processes. For example, consider a short-answer question that ask a participant to define a concept in their own words (Moraes *et al.*, 2018; Roy *et al.*, 2020; Câmara *et al.*, 2021; Roy *et al.*, 2021). Such questions measure a participant's understanding of a concept. However, they do not measure a participant's ability to engage in more complex processes, such as differentiating between multiple concepts (analyze) or judging the value of a concept to explain a phenomenon (evaluate). Finally, compared to multiple-choice, certain short-answer questions (e.g., definitions) may require manual grading. Responses may need to be scored along a continuum versus being correct or incorrect.

Key Takeaway

Like multiple-choice tests, short answer tests are easy to grade, but have the added benefit of being less susceptible to guesswork.

3.4 Open-Ended Assessments

In contrast with closed-ended assessments, open-ended assessments consist of questions that do *not* have a pre-determined correct answer, but rather allow participants to develop their own responses.

3.4.1 Sentence Generation in Vocabulary Learning

Sentence generation is a technique that can be used to measure vocabulary learning. Sentence generation questions ask participants to generate a sentence using a specific vocabulary term. Prior studies have graded responses on a scale, depending on whether the generated sentence is grammatically correct, semantically correct, and *unambiguously* demonstrates that the meaning of a term is fully understood (Heilman and Eskenazi, 2006; Heilman *et al.*, 2010). These same studies also used sentence generation to measure *transfer of learning*. For a given term, participants were asked to generate sentences in a domain different from the one encountered during the vocabulary learning process. For example, if someone learned the meaning of “spectrum” in the context of color, can they generate a sentence in the context of sound?

Sentence generation assessments have two benefits. First, they are easy to develop. Second, participants must generate their own responses, which minimizes guessing.

Sentence generation assessments have three drawbacks. First, they require manual grading. Additionally, in some cases, it may be difficult to determine whether a sentence demonstrates that the meaning of a term is fully understood. Second, they do not measure whether participants learned terms beyond those targeted by the assessment. Third, they do not measure whether participants can engage in cognitive activities more complex than understand. For example, they can measure whether someone understands the meaning of “abandon”, but not whether they can explain the subtle differences between “abandon” and “relinquish.”

3.4.2 Free Recall

Free recall assessments involve asking participants to list as many important terms, phrases, or facts related to the topic of the search task. Studies have asked participants to list as many domain-relevant keywords or phrases (Kammerer *et al.*, 2009) or as many facts related to the items in the collection (Wilson *et al.*, 2008).

Free-recall responses can be scored in different ways. Some studies have scored free-recall responses by simply counting the number of items provided (Bhattacharya and Gwizdka, 2019; Wilson *et al.*, 2008). Other studies have manually assessed and counted the number of "reasonable" items provided (Kammerer *et al.*, 2009) or compared the items provided with a gold-standard list of items using measures such as precision and recall (Bhattacharya and Gwizdka, 2019). Precision measures the percentage of provided items that are in the gold-standard list, and recall measures the percentage of gold-standard items that are in the provided list.

Free recall assessments have three benefits. First, they are easy to develop. Participants must simply be instructed to enumerate keywords, phrases, or facts. Second, depending on the scoring strategy (e.g., simply counting items), they can be easy to grade. Third, participants must generate their own answers, which minimizes guessing.

Free recall assessments have two main drawbacks. First, they do not reliably measure whether participants can engage with cognitive processes more complex than remember. They do not even measure whether participants understand a term, concept, or fact. Second, depending on the scoring strategy, grading may require manual effort. For example, manual activities may include coding the relevance of items provided by participants or generating a gold-standard list of items and then manually assessing the correspondence between items provided and those in the gold-standard list.

3.4.3 Mind Map

A mind map is a visual representation of a topic, process, or domain. Mind maps typically involve nodes and edges. Additionally, they typically focus on one specific concept or idea and take the form of a hierarchy. Major concepts/ideas branch out from a central concept/idea and more specific concepts/ideas branch out from those major concepts/ideas, etc. Mind maps are typically viewed as a subjective versus objective representation of a domain. In this respect, SAL studies have not used mind maps to objectively measure learning. Instead, they have used them to understand knowledge shifts during the learning process (Liu *et al.*, 2019) or to study the effects of prior knowledge on search behaviors (Zhang and Liu, 2020).

Mind maps have three benefits. First, they are very open-ended and therefore have high coverage. Participants can convey any idea or relation deemed meaningful. Second, due to their structured nature, they can be used to understand knowledge shifts during the search process. For example, researchers can characterize mind map modifications based on their type (e.g., node additions, deletions, modifications) and their location (e.g., distance to the central concept/idea). Third, they are easy to administer. Mind map construction tools are readily available.

Mind maps have two main drawbacks. Mind maps only convey relationships between elements. Therefore, they may not assess whether participants can engage with cognitive processes more complex than analyze (i.e., compare, contrast, differentiate). Second, participants may not be familiar with mind maps. In fact, Liu *et al.* (2019) required mind-mapping experience while enrolling participants in their study.

Key Takeaway

Mind maps provide a structured way of measuring knowledge shifts during the search session.

3.4.4 Argumentative Essay

Argumentative essays ask participants to enumerate arguments for and against a specific stance or proposition. Demaree *et al.* (2020) asked participants to enumerate arguments for and against the proposition that nuclear power can solve the climate crisis. Responses were graded based on the number of correct pro and con arguments provided. Two independent coders graded a subset of essays and achieved high agreement in labeling arguments as correct or incorrect.

Argumentative essays have two main benefits. First, they are easy to develop. Participants simply need clear instructions about generating a list of correct pro and con arguments. Second, guesswork is minimized because participants must generate their own answers.

Argumentative essays have two main drawbacks. First, they require manual grading (e.g., labeling arguments as correct or incorrect). Second, they may not reliably measure whether participants can engage in cognitive processes more complex than remember. Participants may simply memorize arguments without being able to summarize them in their own words (understand), describe their differences (analyze), or judge their validity or importance (evaluate).

Key Takeaway

Open-ended responses can be scored in different ways. However, the scoring criteria are linked to the types of cognitive processes being tested. For example, counting pro and con arguments is only reliably testing a participant's ability to recall information.

3.4.5 Summary and Open-Ended

Summary and open-ended assessments ask participants to either summarize what they know about a topic or answer an open-ended question. In contrast to short-answer questions, there is no single correct answer. Therefore, responses are typically scored using qualitative coding techniques.

Collins-Thompson *et al.* (2016) and Abualsaud (2017) used three open-ended questions that asked participants to: (1) write an outline for a hypothetical paper on the topic of the search task, (2) describe what they learned about the topic, and (3) enumerate unanswered questions about the task topic. Responses were graded on a 7-point scale using a qualitative coding scheme. Grading criteria checked for the inclusion of facts, themes, issues, concepts, and relationships between concepts.

Kalyani and Gadiraju (2019) measured learning during search tasks associated with different cognitive processes from A&K's taxonomy. Each search task had its own learning assessment. Assessments for high-complexity tasks (evaluate, create) used open-ended questions. The evaluate question asked participants to consider different alternatives, choose the best option, and provide a justification. The create question asked participants to design a plan. The grading criteria for these two open-ended questions were not described in detail. The authors simply noted that responses were "manually checked and marked [...] as valid upon encountering complete and comprehensive submissions" (Kalyani and Gadiraju, 2019, p. 128).

Lei *et al.* (2015) asked participants to complete an open-ended worksheet after searching for videos on the topic of animal courtship. Worksheets were graded on a 10-point scale based on the specification of animal names and their respective behaviors/actions. Assessments were graded by three independent annotators. Interannotator agreement was measured using Kendall's τ, a rank correlation metric. Rank correlation metrics do not *directly* compare scores from different assessors. Instead, they measure the extent to which scores from different assessors yield similar rankings.

Wilson and Wilson (2013) proposed a novel qualitative coding scheme to evaluate open-ended summaries in which participants describe what they learned. The proposed coding scheme evaluates summaries along three dimensions. Each dimension was inspired by a different cognitive process from A&K's taxonomy. First, the D-Qual dimension represents the understand cognitive process and measures the quality of facts included in the summary. Second, the D-Intrp dimension represents the analyze cognitive process and measures the extent to which the summary draws *connections* between facts. Finally, the D-Crit dimension represents the evaluate cognitive process and measures the extent to which the summary includes evaluative statements (i.e., evidence of critical thinking).

O'Brien *et al.* (2020) measured learning by asking participants to summarize their knowledge before and after searching. Pre- and post-task summaries were scored using the D-Qual and D-Intrp dimensions from Wilson and Wilson (2013). In term of D-Qual, summaries were scored based on the number of accurate facts included. In terms D-Intrp, summaries were scored based on the number of explicit associations between facts. Knowledge gains were measured by comparing pre- and post-task scores along these two dimensions independently.

Palani *et al.* (2021) measured learning by students enrolled in a project-based design course. Participants searched for 30 minutes on a topic related to a course project. To measure learning, participants were asked to write pre- and post-task summaries about the topic. Additionally, after searching, participants were asked to write a problem statement describing their specific plan for the course project. Pre- and post-task summaries were compared along four dimensions. First, the authors compared the number of facts included in the summaries. Additionally, the authors scored summaries using the D-Qual, D-Intrp, and D-Crit dimensions from Wilson and Wilson (2013). Finally, project plans were scored on a 5-point scale. A score of 1 indicated that the plan was very ill-defined and a score of 5 indicated that the plan was specific, well-informed, and well-reasoned. The authors reported moderate-to-high levels of interannotator agreement across measures.

Roy *et al.* (2021) asked participants to summarize what they learned after each search task. Summaries were scored along two dimensions. F-Fact scores were computed by counting the number of facts included in the summary. T-Depth scores were computed by measuring the extent to which specific subtopics were covered in depth. Each search task was associated with a predefined set of subtopics. Each essay was scored on a 3-point scale for each subtopic, and a final T-Depth score was computed by averaging across subtopics. Covering a subtopic *in depth* involved including supporting evidence and examples. To validate this grading rubric, three assessors manually evaluated a common subset of essays. F-Fact and T-Depth scores were found to be highly correlated between assessors.

Kammerer *et al.* (2009) asked participants to gather information and create open-ended responses to different summarization tasks. For example, one of the search tasks asked participants to discuss three important trends regarding the future of architecture. Summaries were graded using task-specific criteria. For example, for the task above, summaries were graded based on the number of architectural trends discussed (4-point scale) and the overall quality of the descriptions (3-point scale).

Liu and Song (2018) investigated learning during search tasks that involved comparing and contrasting alternatives. Participants were asked to create an open response to the task. Responses were scored based on the number and relevance of facts provided, the number of facets considered, the presence of pro and con arguments, and the presence of opinionated statements. Two assessors coded all summaries and achieved moderate-to-high levels of agreement across all qualitative codes. A third assessor resolved disagreements.

Pardi *et al.* (2020) asked participants to gather information in order to explain how thunderstorms and lightning form. To measure learning, participants wrote open-ended explanations before and after searching. To score summaries, the authors identified 20 concepts related to the formation of thunderstorms and lightning. Summaries were scored based on the number of concepts mentioned. Interestingly, concepts were only counted if the summary also specified relations to other concepts.

Hornbæk and Frøkjær (2003) evaluated different interfaces for reading documents. Participants completed two types of tasks: (1) a question-answering task that asked participants to seek answers to specific questions and (2) a document-understanding task that asked participants to determine the main theses and ideas in the article. To measure learning during the question-answering task, participants answered open-ended questions. These responses were graded according to how many aspects of the question were covered in the response. For each question, the different aspects were determined in advance. To measure learning during the document-understanding task, participants were asked to write an essay describing the main theses and ideas in the article. These responses were graded based on the number of main theses and ideas included in the response (also determined in advance). In all cases, responses were graded on a 4-point scale by only one of the authors.

Salmerón *et al.* (2020) evaluated a system intervention to improve reading comprehension. Participants were asked to learn about a specific topic (i.e., climate change or genetically modified food) by reading documents displayed on a static SERP. To measure learning, participants were asked to write an open-ended essay on the given topic before and after each task. To evaluate their quality, essays were first divided into "idea units", defined as units that describe a specific event, activity, or state. Next, idea units were coded along two dimensions. First, idea units were coded based on whether the participant referenced the primary or secondary source of the idea. Second, idea units were assigned to three different categories based on the level of synthesis conveyed: (1) paraphrasing a single idea from a document, (2) combining two or more ideas from the same document that were not explicitly connected, and (3) combining two or more ideas from different documents. Along both dimensions, essays were analyzed based on the number of idea units belonging to each category. To validate this coding scheme, two annotators coded about 10% of the data and intercoder agreement was measured using Cohen's κ.

Davies *et al.* (2013) administered two types of open-ended assessments to measure conceptual learning during searches supported by different note-taking tools. Participants completed both assessments before and after searching on the subject of plate tectonics. One type of assessment asked participants to explain the tectonic processes depicted in a given diagram. These responses were scored on a 3-point scale based on the depth of the explanation provided. A second type of assessment asked participants to explain the relationship between pairs of concepts. These responses were assessed based on accuracy (2-point scale) and the depth of explanation provided (3-point scale).

Willoughby *et al.* (2009) investigated the effects of four factors on the quality of essays written by participants on a given topic (e.g., "How does human metabolism work?" or "What are major urban environmental issues?"). The four factors were: (1) prior knowledge, (2) search skills training, (3) searching before writing the essay, and (4) planning before writing the essay. Essays were scored based on the number of correct facts. The authors reported high levels of interannotator agreement.

Demaree *et al.* (2020) used summaries to measure prior knowledge. To score summaries, two independent coders counted the number of relevant concepts included in the summary. The authors reported an intercoder agreement of 73%.

Urgo and Arguello (2024) asked participants to write summaries describing everything they learned during the search task. Summaries were scored based on the percentage of correct statements included. This processes required splitting sentences into statements that are either entirely true or entirely false. Analyzing statements (versus sentences) was done in order to give participants partial credit for sentences containing both true statements and false statements. Statements were judged as correct or incorrect by two domain experts that worked together and resolved disagreements through discussion.

Summary and open-ended assessments offer four important benefits. First, summary and open-ended assessments have high coverage. These assessments give participants the ability to describe everything they learned during a search session. The open-endedness of the assessment allows researchers to gain insights about participants' breadth and depth of learning. Second, the assessment minimizes guessing because responses are fully generated by participants. Third, depending on the question, open-ended assessments can target varying levels of cognitive complexity. In other words, open-ended questions can be specifically designed to measure a participant's ability to effectively recall, understand, apply, analyze, evaluate, and create. Finally, depending on what participants are asked to produce, the assessment materials may be easy to develop. Participants can simply be asked to summarize what they learned during the search task.

Summary and open-ended assessments have two main drawbacks. First, grading is time-consuming. Grading requires generating a qualitative coding guide. This process involves defining grading criteria and measuring intercoder agreement to ensure that the coding guide is reliable. Second, the quality of responses may be difficult to compare across participants. This type of assessment imposes very few constraints on participants' responses. This may cause some participants to satisfice and not convey everything they learned during the task. Additionally, writing skills may vary across participants. Some participants may not be able to effectively communicate everything they learned.

Key Takeaway

Knowledge summaries can be scored based on the cognitive processes reflected in the response.

3.4.6 Mental Model Assessment

Mental models are internal representations of external systems or phenomena (Jones *et al.*, 2011). Prior studies in education and psychology have used mental models to measure learning. The key assumption is that learners with greater domain knowledge are able to generate more accurate and complete mental models (Nersessian, 2002). Communicating mental models often involves drawing pictures with words and symbols (Jones *et al.*, 2011). Additionally, mental model assessments can illuminate gaps in an individual's understanding of a system or phenomenon.

Chi *et al.* (2001) used mental models to measure learning during a tutoring session about the human body's circulatory system. Before and after the tutoring session, participants were given a sheet of paper with an outline of the human body and were asked to explain the path of blood through the circulatory system. To evaluate participants' mental models, Chi *et al.* (2001) developed seven different mental models. Six of these models had different degrees of errors. All seven models were ranked from the most accurate and complete "Double Loop-2" model to the most naïve "No Loop" model. The effectiveness of the tutoring session was evaluated in two ways. First, the authors counted how many students had the most accurate "Double Loop-2" model before and after the session. Second, the authors leveraged their ranking of mental models to compute the average number of mental models shifts before and after the tutoring session. For example, students who drew the most naïve "No Loop" model before the session and the most complex "Double Loop-2" model after the session received a score of six.

Mental model assessments have several strengths. This type of assessment aligns well with tasks that require participants to develop a deeper understanding of a complex system, process, or phenomenon. Second, responses are fully generated by participants, which reduces guesswork. Third, mental models can be evaluated based on their accuracy and completeness, which considers both depth and breadth of learning.

Mental model assessments have two drawbacks. First, scoring mental models requires domain expertise and possibly even pedagogical expertise. Chi *et al.* (2001) scored mental models by comparing them to seven different mental models of the human body's circulatory system that students typically form. Second, mental model assessments may not capture a searchers ability to engage in critical thinking (evaluate) or creative thinking (create).

3.5 Controlling for Prior Knowledge

Up to this point, we have discussed different types of assessments that can be used to measure prior knowledge and/or learning during search. Another important question is: If the goal is to measure learning during search, how do we account for differences in prior knowledge across participants?

Prior work has used different approaches to account for prior knowledge. The simplest approach is to ignore it. Some studies have only administered a post-task assessment. This can be risky in a between-subjects study. For example, consider a study in which participants are assigned to different interface conditions. What if participants assigned to one condition have higher levels of prior knowledge by pure chance? This risk can be mitigated in several ways. One way is to recruit participants that are expected to have similar levels of prior knowledge. For example, Urgo and Arguello (2023) used a task involving the concepts of diffusion and osmosis and participants were not allowed to be students majoring in biology nor chemistry. Another risk mitigation approach is to capture and compare participants' pre-task perceptions of the task. Urgo and Arguello (2023) captured participants' pre-task perceptions of interest, prior knowledge, expected difficulty, and *a pri-*

ori determinability—the extent to which different aspects of the task (e.g., requirements, outcomes) are known in advance. Participants assigned to different interface conditions reported similar perceptions—no differences between groups were statistically significant.

Education research has shown that prior knowledge has an effect on learning outcomes and has argued that methodological controls such as targeting novice participants are risky (Shapiro, 2004). Therefore, a better approach is to have participants complete the same assessment before and after the search task. Then, the question is: How do we compare pre- and post-task scores to measure learning during the search session?

There are several ways to combine pre- and post-task scores to form a dependent variable. The simplest approach is to compute the *gain*:

$$\text{Gain} = \text{PostScore} - \text{PreScore}. \tag{3.1}$$

However, this approach does not account for the fact that participants with lower prior knowledge have more to gain than participants with higher prior knowledge. This is problematic when averaging gains across participants. Therefore, another approach is to compute a measure known as *normalized gain:*

$$\text{Normalized Gain} = \frac{(\text{PostScore} - \text{PreScore})}{(MaxScorePossible - \text{PreScore})}. \tag{3.2}$$

This type of normalization is common in education (Hake, 2002) and SAL studies (Gadiraju *et al.*, 2018; Yu *et al.*, 2018; Xu *et al.*, 2020; Urgo and Arguello, 2024). Normalized gain accounts for participants' prior knowledge based on their pre-task scores. It essentially answers the question: Of the percentage a participant could have gained, what percentage did they actually gain?

It should be noted that, on rare occasions, gain and normalized gain can be negative (i.e., PreScore > PostScore). This can happen due to participants: (1) not learning, (2) guessing correctly on the pre-test, and (3) guessing incorrectly on the post-test. Therefore, studies have also converted negative gains and normalized gains to zero (Câmara *et al.*, 2021).

Studies have also considered other ways to combine pre-task and post-task assessment scores. For example, Syed and Collins-Thompson (2018) considered the following measures. First, they counted the number of items that were incorrect in the pre-test and correct in the post-test. Second, they weighted items by difficulty based on the number of participants who answered them incorrectly in the pre-test. Additionally, Syed and Collins-Thompson (2018) considered knowledge retention. Participants completed the same test several months after the study session. Pre-task, post-task, and retention assessment scores were combined in different ways, including: (1) the number items that were incorrect in the pre-task assessment and correct in the retention assessment; (2) the number of items that were incorrect in the pre-task assessment, correct in the post-task assessment, and still correct in the retention assessment; and (3) the number of items that were correct in the post-task assessment and still correct in the retention assessment.

Key Takeaway

Measuring prior knowledge is a good idea, especially during a between-subjects study.

3.6 Summary and Recommendations

In this section, we have seen that there are many ways to measure prior knowledge and learning during a search task. All methods have pros and cons. First, some methods (e.g., multiple-choice tests) are more susceptible to guesswork than other methods (e.g., open-ended). Second, some methods (e.g., short-answer) target specific topics and may not capture *everything* that someone knows before the search task or learns during the search task. Other methods (e.g., knowledge summaries) allow participants to convey everything they know or learned. Third, some methods are easier to grade than others. For example, multiple-choice tests can be graded automatically; short-answer tests must be graded manually but have objectively correct answers that can be easily verified; and open-ended assessments may involve greater effort. Scoring open-ended responses may involve verifying the correctness of statements or looking for evidence of specific types of mental processes, such as paraphrasing, synthesizing, comparing/contrasting, or evaluating. Open-ended assessments are typically scored using qualitative techniques, which may require developing a coding guide and validating the coding guide by measuring inter-annotator agreement on some subset of the data. Fourth, some methods (self-report) are easy to implement but may not reliably measure learning or may be impacted by characteristics of the individual. Finally, some methods cannot determine if the learner is able to engage in complex cognitive processes with the knowledge acquired. For example, counting the number of relevant pro and con arguments in an argumentative essay cannot determine whether a participant is able to engage in cognitive processes more complex than remember (i.e., recalling information).

Deciding how to measure prior knowledge and learning is an important decision in any SAL study. We suggest that future SAL studies consider the following recommendations.

3.6.1 Using Validated Assessment Materials

Multiple-choice tests are easy to grade but difficult to develop. They should ask questions and include answer choices that are grounded in common misconceptions that learners have within the domain of the task. Ideally, they should be developed by domain experts with pedagogical experience in the topic. SAL researchers often develop multiple-choice tests to match the topic of the search task. This is risky because there is no evidence that the test can reliably distinguish between people with different levels of knowledge. Researchers should consider taking a different approach. Rather than designing a multiple-choice test to match the topic of the search task, we recommend that researchers use a validated multiple-choice test and design a search task to match the test.

Urgo and Arguello (2024) adopted this approach. They used a multiple-choice test called the Osmosis and Diffusion Conceptual Assessment (ODCA) (Fisher *et al.*, 2011). The ODCA was developed by expert biology faculty. It includes both simple questions that ask about definitions and complex questions that ask about the expected outcomes in different scenarios. The ODCA includes 18 questions that are organized in pairs. Each pair includes a knowledge question and a reasoning question. Each knowledge question asks about specific concepts or processes and each reasoning questions asks for an explanation of the answer to the corresponding reasoning question. One might say that knowledge questions ask "what" and reasoning questions ask "why." In this respect, the assessment can be scored based on the percentage of pairs that are both correct. That is, participants get points only if they correctly know "what" and "why." In designing the assessment, a panel of students provided feedback. Additionally, the items in the ODCA have been shown to have high internal consistency across different student cohorts. That is, students with similar levels of knowledge tend to get the same items right and wrong. To match the ODCA, Urgo and Arguello (2024) asked participants to "learn everything you can about the biological concepts of osmosis and diffusion."

3.6.2 Consider the Cognitive Processes Measured by the Assessment

Learning can be characterized by the types of cognitive processes the learner is able to successfully engage in using the acquired knowledge. As discussed in Section 2, the cognitive process dimension from A&K's taxonomy is a useful framework to characterize how *deeply* someone learned something (Anderson *et al.*, 2000). It should be used to determine the types of learning being measured by an assessment or an assessment item (e.g., an open-ended prompt or a multiple-choice question). For example, is the assessment asking the participant to recall information verbatim (remember), summarize information in their own words (understand), use the acquired knowledge to complete a task (apply), compare different ideas (analyze), judge the significance of different ideas (evaluate), or use the acquired knowledge to generate something novel (create). By explicitly associating cognitive processes to assessment items, SAL researchers may be better able to determine the depth of knowledge acquired by participants during a search session. For example, a study might only find differences for items that ask participants to engage in complex cognitive processes with the acquired knowledge.

3.6.3 Combining Multiple Types of Assessments

SAL studies should use multiple types of assessments for two reasons. First, it helps counteract the different drawbacks associated with different assessment types. For example, consider a study that uses both a multiple-choice assessment and an assessment that asks participants to summarize everything they learned. The multiple-choice assessment can ask about topics that learners usually struggle with but may not capture everything that participants learned. Conversely, the knowledge summary gives participants the opportunity to convey everything they learned but differences in writing skills could be a confounding factor. Second, combining assessments may provide stronger evidence of learning. If participants assigned to one experimental condition improve their scores on two different types of assessments, this provides strong evidence that they learned more in that condition.

Key Takeaway

All assessment types have benefits and drawbacks. To mitigate the drawbacks, studies can combine multiple types of assessments.

3.6.4 Measuring Knowledge Retention

SAL studies typically measure learning immediately after the search task. SAL studies should also measure retention—a participant's ability to use the acquired knowledge in the long term. Retention assessments indicate what information has been moved into a learner's long-term memory storage. When saving information in long-term memory, Sousa (2017) explains that the brain has determined that the information has both *sense* (i.e., the learner has integrated the information into their existing knowledge structures) and *meaning* (i.e., the learner has determined that the information is relevant). Therefore, retention shows which information has been deeply learned, having both sense and meaning to the learner.

Learning retention assessment methods are designed to measure how much or how well knowledge has been integrated into long-term memory. This can be measured by administering a delayed post-test after the search session. Research has shown that the largest loss of newly acquired information or skills occurs within 18 to 24 hours (Sousa, 2017). For this reason, we recommend waiting at least 24 hours after the search session before administering a learning assessment that is meant to capture learning retention.

A few SAL studies have measured knowledge retention. Qiu *et al.* (2020) compared retention levels between participants using a conversational search interface versus a traditional search interface. Participants who used the conversational interface had lower levels of *information loss*, defined as the number of items answered correctly in the post-task test and incorrectly in a delayed retention test taken a few days later. Syed and Collins-Thompson (2018) explored retention in the context of a vocabulary learning search task. Results found that participants who interacted with documents with a higher density of difficult vocabulary words had higher retention rates of those words based on a test taken nine months after the search session. Urgo and Arguello (2024) explored that effects of goal-setting on learning and retention. Post-task assessments did not find significant differences between participants who set goals and participants who did not set goals. However, participants who set goals had higher levels of retention based on a test taken one week after the search session.

Key Takeaway

Retention assessments do not always show the same effects as assessments taken immediately after the search task.

3.6.5 Measuring Transfer of Learning

Transfer of learning involves applying knowledge in a new context or situation (Haskell, 2001). For example, suppose a searcher is asked to learn about Bernoulli's principle by understanding how it enables an airplane to fly. Transfer of learning would involve the searcher using this newly acquired knowledge to explain some other phenomenon, such as topspin in tennis. Assessments that target transfer of learning measure the learner's ability to use knowledge in a new context from the one encountered during the learning process. Essentially, transfer of learning measures a learner's ability to *generalize* from the learning context to the transfer context. Anderson *et al.* (2000) argued that being able to transfer knowledge to new situations or problems is a core tenet of meaningful learning. Research in psychology and education has found the students fail to transfer knowledge to new situations when they do not understand it deeply enough. In other words, depth of learning is a *prerequisite* for successful knowledge transfer.

Transfer of learning has largely gone unexplored in SAL studies. One possible reason is that designing assessment materials that measure transfer of learning is challenging. It requires the learning assessment to ask questions that are relevant to but different from the learning objective given to participants. We see two possible ways to do this, discussed below.

First, as recommended by Urgo and Arguello (2022b), one way is to ask questions that match the same cognitive process and knowledge type as the objective but focus on different material. For example, if the learning objective is to understand how concept A is exemplified by example B (understand/conceptual), the assessment might ask how concept A is exemplified by example C. Similarly, if the objective is to understand how procedure A can be used complete task B (apply/procedural), the assessment might ask how procedure A can be used to complete task C.

Key Takeaway

To better capture depth of learning, future studies should test participants' ability to transfer their new knowledge to novel contexts or scenarios.

A second possible way to measure a participant's ability to transfer knowledge is for the assessment to ask questions that involve cognitive processes more complex than the objective. For example, if the objective is to understand the differences between concepts A and B (analyze/conceptual), the assessment might ask which concept (A or B) best explains phenomenon C (evaluate/conceptual).

4

The Effects of Task and Searcher Characteristics on Learning

An important question in SAL research is: What are factors that impact learning during search? One obvious factor might be the system itself. In Section 6, we review studies focusing on experimental system features and/or tools to encourage and support learning during search. Beyond the system itself, however, there are other factors that might impact learning during search. Some factors may relate to the target learning objective—what is the searcher trying to learn? As discussed in Section 2, learning objectives can vary by complexity and can involve types of knowledge that vary along dimensions such as abstractness, interconnectedness, and subjectivity. Such task characteristics are likely to impact learning. Other task characteristics may relate to how the task or learning objective is structured. For example, one objective might involve learning about elements that are relatively independent (i.e., could be learned in any order), while another objective might involve learning about elements that are highly interconnected (i.e., must or should be learned in sequence). Another factor is the searcher's familiarity with the topical domain in question. Higher levels of domain knowledge may improve search efficacy, resulting in better learning outcomes. Finally, other factors may include characteristics of the individual searcher, which may include specific skills (e.g, reading comprehension ability), cognitive abilities (e.g., working memory capacity), or personality traits (e.g., need for cognition). In this section, we review SAL studies that have focused on understanding how such factors can impact learning during search.

4.1 Task Complexity

In Section 2.3.1, we reviewed studies that leveraged the cognitive process dimension from A&K's taxonomy to manipulate the complexity of search tasks assigned to participants. Most of those studies focused on the effects of task complexity on search behaviors and perceptions (not learning). In this section, we review the subset of studies that focused on learning as an important dependent variable (Ghosh *et al.*, 2018; Kalyani and Gadiraju, 2019; Liu *et al.*, 2019).

Ghosh *et al.* (2018) had participants complete tasks associated with the cognitive processes of remember/understand, apply, analyze, and evaluate. To measure learning, participants were asked about their knowledge of the task topic before and after searching. Participants reported positive knowledge gains for all tasks. However, participants were less confident about their prior knowledge during more complex tasks.

Kalyani and Gadiraju (2019) had participants complete tasks associated with all six cognitive processes from A&K's taxonomy. However, the authors only measured learning outcomes for tasks associated with the cognitive processes of remember, understand, apply, and analyze. Participants had lower knowledge gains for the most complex task (i.e., analyze) versus the third most complex task (i.e., apply).

Liu *et al.* (2019) asked participants to complete both a receptive task and a critical task. Receptive tasks were simpler (i.e., remember/understand) and critical tasks were more complex (i.e., evaluate). Participants were asked to develop mind maps to illustrate their evolution of knowledge throughout the search session. During receptive tasks, participants made edits to their mind maps throughout the *entire* search session. Conversely, during critical tasks, participants made more edits to their mind maps toward the *end* of the search session.

Key Takeaway

Studies suggest that searchers are more likely to achieve simple versus complex learning objectives.

4.2 Task Structure

Most SAL studies have investigated learning during a single search session. As one exception, Liu *et al.* (2013) investigated learning across multiple search sessions. Participants completed tasks that involved writing an article about hybrid cars. The study manipulated how the general task was decomposed into three subtasks that were completed across different search sessions. In the dependent subtask condition, the three subtasks were designed to build on each other. Conversely, in the parallel subtask condition, the three subtasks were largely independent (i.e., could hypothetically be completed in any order). To measure learning, participants rated their knowledge of the overall task topic before and after each search session. The manipulation of the task's decomposition had an interesting effect. In the dependent subtask condition, participants reported greater topic familiarity after all three search sessions. Conversely, in the parallel subtask condition, participants reported greater topic familiarity after the first two search sessions but not the third session. That is, during the third session, participants reported similar levels of topic familiarity before and after searching. This result suggests that participants perceived their knowledge to plateau more quickly in the parallel subtask condition.

4.3 Domain Knowledge

Several studies have explored the role of domain knowledge on learning during search.

Willoughby *et al.* (2009) conducted a study in which participants were asked to write essays on two topics. Topics were chosen such that participants had high prior knowledge about one topic and low prior knowledge about the other. Additionally, participants were grouped into two conditions. One group was allowed to search for 30 minutes before writing each essay and one group had to write their essays without searching. Participants who were allowed to search wrote essays with more relevant facts. However, this trend was only observed for the high-prior-knowledge topic. That is, for the low-prior-knowledge topic, being able to search did not improve learning based on the quality of essays written by participants. This result suggests that high prior knowledge may improve learning by improving search efficacy.

O'Brien *et al.* (2020) investigated the differences in learning outcomes between domain experts and non-experts. To measure learning, participants generated knowledge summaries before and after searching. Non-experts had *slightly* greater improvements in their summaries than experts. One possibility is that non-experts were more likely to encounter new information while searching.

Roy *et al.* (2020) asked participants to complete knowledge assessments intermittently throughout the search session. Domain knowledge affected *when* participants had greater knowledge gains. Novices learned more at the beginning of the search session while experts learned more toward the end. This result is congruent with those from O'Brien *et al.* (2020). One possibility is that experts had to search for longer to encounter new information that expanded their prior knowledge.

Key Takeaway

The relation between domain knowledge and learning is not straightforward. Domain experts may search more effectively but have less to learn.

4.4 Cognitive Abilities

Several studies have explored how different cognitive abilities may impact learning during search. Studies have considered cognitive abilities such as working memory capcity (Pardi *et al.*, 2020; Choi *et al.*, 2019a), perceptual speed (Azzopardi *et al.*, 2023), and an individual's propensity to become distracted while working on a task (Azzopardi *et al.*, 2023).

Pardi *et al.* (2020) explored the role of both working memory capacity and reading comprehension ability on learning during search. Working memory capacity refers to an individual's ability to hold and manipulate information in short-term memory, when it is no longer perceptually present (Diamond, 2013). To measure learning, participants were asked to write knowledge summaries before and after searching. Knowledge summaries were scored based on the number of relevant concepts included in the summary. After controlling for prior knowledge, both working memory capacity and reading comprehension ability had positive effects on learning.

Choi *et al.* (2019a) explored the effects of working memory capacity during search tasks of varying complexity. Participants completed decision-making tasks that asked them to compare a 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 gather information to complete a 2×2 or 4×4 table (depending on the task's complexity), choose an alternative based on the information gathered, and write a justification about their choice. Participants with higher working memory capacity wrote longer justifications, suggesting that they had better learning outcomes. Additionally, this trend was more pronounced for complex tasks.

Azzopardi *et al.* (2023) investigated the effects of two cognitive abilities on learning during search: (1) perceptual speed (PS) and (2) cognitive failure (CF). Perceptual speed refers to an individual's ability to quickly scan a visual display for information (French *et al.*, 1976). Cognitive failure refers to how easily someone can become distracted while performing a task (Broadbent *et al.*, 1982). Participants completed journalism tasks that asked them to find relevant examples of an event type for a hypothetical news article (e.g., write about recent tropical storms that caused widespread destruction). To measure learning, participants were asked to recall as many examples as possible after the task. High-PS participants were able to recall more examples than low-PS participants. Similarly, low-CF participants (less prone to become distracted) were able to recall more examples than high-CF participants (more prone to become distracted).

Key Takeaway

Studies suggest that cognitive abilities such as working memory, perceptual speed, and one's ability to avoid distractions have positive effects on learning during search.

4.5 Summary

SAL studies have investigated how different characteristics of the search task and the individual searcher can impact learning during search. In terms of task characteristics, studies have mostly focused on task complexity. The main trend is that complex tasks have slightly lower knowledge gains. This result is congruent with prior studies that have manipulated task complexity without measuring the effects on learning outcomes. Many of the studies previously surveyed in Section 2.3.1 found that complex tasks are perceived to be more difficult and require more search activity. Additionally, complex tasks have more divergent behaviors. For example, searchers issue similar queries during simple tasks and different queries during complex tasks. This divergence in search behaviors suggests that complex tasks are more open-ended—an effective approach to the task is less obvious.

In terms of the searcher, several studies have investigated the role of domain knowledge on learning during search. Here, results are less conclusive. On one hand, domain knowledge can help people search more effectively and achieve better learning outcomes. On the other hand, domain experts have less to learn than domain novices. In this respect, they may need to work harder to find information that extends their high levels prior knowledge.

So, how does domain knowledge impact learning during search? This remains an open question. Perhaps there is an inverted-U shaped relation between domain knowledge and learning during search. Perhaps searchers with the best knowledge gains are those who still have lots to learn but know enough about the domain to search effectively. Extreme novices and extreme experts may have lower knowledge gains. Extreme novices may have difficulty searching effectively and finding relevant information. Extreme experts have less to learn and may need to work harder to find information that extends their prior knowledge.

Studies have also considered the role of different cognitive abilities, such as working memory capacity, perceptual speed, and the searcher's susceptibility to becoming distracted while performing a task. Studies have found that such characteristics can impact learning outcomes.

People do not only vary by their abilities, but also by their personality. To our knowledge, SAL studies have not considered personality traits that may impact learning during search. Need for Cognition (NFC) is one personality trait worth exploring in future work. NFC refers to the extent to which a person enjoys cognitively effortful activities (Cacioppo *et al.*, 1996). Cacioppo *et al.* (1984) developed a scale for measuring NFC. In the context of information retrieval, Wu *et al.* (2014) investigated the role of NFC on search behaviors. NFC impacted participants' query-reformulation and SERP-scanning behaviors. When the top results were not relevant, high-NFC participants were more likely to reformulate the query than to keep scanning the SERP. Conversely, when the top results were not relevant, low-NFC participants were more likely to keep scanning the SERP than to reformulate the query. One possible explanation is the query-reformulation is more cognitively demanding than results evaluation. This might explain why high-NFC searchers are more likely to reformulate the query (vs. continue scanning the SERP) when it seemed like the best course of action. Given the effects of NFC on such behaviors, future SAL studies should consider the effects of NFC on learning during search.

5

Predicting Learning During Search

Many different SAL studies have explored whether specific search behaviors can predict learning during search. Most of these studies have considered measures generated from participants' queries, clicks, mouse events, and scrolls, as well as temporal measures such as the total session length and the time spent by participants on different activities (e.g., reading versus searching).

A natural question is: Why would we want to automatically predict learning during search? Perhaps the simplest answer is: evaluation. Understanding how search behaviors predict learning can help us automatically determine whether a particular system encourages and supports learning. A second (more ambitious) possibility is to develop interfaces and/or interventions that nudge searchers toward behaviors that promote learning and away from behaviors that hinder learning.

In this section, we review studies on search behaviors that can predict learning during search. Additionally, we review studies that have compared the search behaviors of domain experts versus non-experts. Predicting domain expertise is different from predicting learning during a search session. However, measures that predict domain expertise may be useful in future work that predicts learning of a topic over an extended time period.

It is important to note that the studies in this section exploit mere correlations between behaviors and learning outcomes. Correlation does not imply causation. Therefore, if a behavior predicts learning, it does not mean it *causes* learning.

There are two other possibilities. One alternative is that learning causes the behavior—the causal relation goes in the opposite direction. For example, as someone learns about a topic they may issue queries with more technical vocabulary. Therefore, the presence of technical query terms may predict learning during search. However, it is learning that causes the presence of technical query terms and not the other way around.

A second possibility is that a third, latent factor causes the behavior and learning to be correlated. For example, motivation may cause people to spend more time searching and have greater knowledge gains. Therefore, session length may predict learning. However, it may be that motivation is the root cause of both a lengthy search session and better learning outcomes.

5.1 Behaviors that Predict Learning During Search

Eickhoff *et al.* (2014) analyzed web search sessions automatically predicted to have either procedural or declarative knowledge acquisition intent.[1] As search sessions progressed, searchers issued more complex queries (i.e., at a higher reading level) that returned results from a wider range of web domains and a narrower range of topics. The authors also analyzed behaviors across multiple sessions for the same task, predicted based on query-term overlap. Some behaviors persisted across search sessions. For example, if query complexity *increased* during session $\mathcal{S}_t$, it also tended to be higher during session $\mathcal{S}_{t+1}$.

Bhattacharya and Gwizdka (2019) investigated search and reading behaviors that predict learning during search. Reading behaviors were generated using eye-tracking data. In terms of search behaviors, results found that participants who used more technical terms in their queries

[1] These were predicted using language models generated from queries with clicks in specific sources, Wikipedia for declarative knowledge and WikiHow for procedural knowledge.

had better learning outcomes. In terms of reading behaviors, results found that participants with more eye regressions (i.e., evidence of re-reading text) had worse learning outcomes. This result suggests that participants who learned less had more difficulty understanding the text encountered and had to re-read it.

Yu *et al.* (2018) used machine learning to predict both knowledge state and knowledge gain. Knowledge state was measured based on performance on a post-task assessment and knowledge gain was measured based on the improvement in performance between a pre-task and post-task assessment. The study considered a wide range of features generated from querying behaviors, as well as interactions on the SERP and pages visited during the search session. For both predictive tasks, participants were grouped into low, moderate, and high performers. Results found several trends. First, predicting knowledge gain was easier than predicting knowledge state. Second, the most predictive features were different for predicting knowledge state versus gain. For predicting knowledge state, the most predictive features were associated with query complexity and the title length of pages visited. Query complexity was measured using a dictionary of about 30,000 words and the average age at which native English speakers typically learn each word. For predicting knowledge gain, the most predictive features were associated with the amount of time participants spent on pages visited. Third, for both predictive tasks, it was easier to distinguish between participants in the low versus high performance groups. That is, it was more difficult to distinguish between participants in adjacent groups (i.e., low versus moderate and high versus moderate performers). Gadiraju *et al.* (2018) found similar results as Yu *et al.* (2018). Query complexity and the amount of time spent on visited pages were positively correlated with knowledge gains.

Otto *et al.* (2021) used machine learning to predict learning during search. The authors explored over a hundred features associated with characteristics of pages visited during the search session, including characteristics of the text and multimedia elements on the page (i.e., images and videos). Similar to Yu *et al.* (2018), participants were grouped into low, moderate, and high performance. The best prediction performance was obtained by combining both textual and multimedia features.

Collins-Thompson *et al.* (2016) conducted a study with three experimental conditions. In one condition (SQ), participants were only allowed to issue one query from a predefined set. In a second condition (MQ), participants were allowed to issue multiple queries from a predefined set. In a third condition (ID), participants were allowed to issue multiple queries from a predefined set that returned more diverse results than the queries in the MQ condition. Results found that more participants were able to achieve both factual and conceptual knowledge gains in the ID condition. Additionally, across experimental conditions, the amount of time spent per document was highly correlated with learning scores.

Lu and Hsiao (2017) investigated how students search in order to learn about computer programming. The study compared search behaviors and learning outcomes for novice versus advanced students. Results found that the amount of time participants spent reading pages had a positive effect on learning outcomes for both student groups. However, novice students had to spend more time reading to achieve similar outcomes as advanced students.

Lei *et al.* (2015) investigated learning by fifth graders using a video search system. Participants were asked to learn about the topic of animal courtship and learning was measured based on the number of relevant concepts recalled by participants. The total number of query-terms used during the search session had a significant *negative* correlation with learning outcomes. The authors observed that students with better learning outcomes were able to identify the core topic of the task and simply queried for "animal courtship" or "courtship behaviors", which yielded relevant results.

Syed and Collins-Thompson (2017a) experimented with retrieval algorithms to support vocabulary learning during search—helping people learn about a target list of unknown vocabulary words by searching for documents that show them in context. Results found that favoring documents with a higher density of unknown vocabulary words improves learning. Interestingly, it also increased the amount of time participants spent on documents relative to the number of words in the document.

Liu and Song (2018) investigated search behaviors that predict learning during two types of learning-oriented search tasks: receptive tasks and critical tasks. Receptive tasks involve gaining a deeper understanding of a topic. Critical tasks involve critiquing and evaluating ideas from different perspectives. Participants had better learning outcomes when they adapted their information source selection strategies depending on the task. Specifically, they had better learning outcomes when they gathered information from encyclopedic sources for receptive tasks and community Q&A sources for critical tasks. Receptive tasks require objective information (e.g., encyclopedic sources) and critical tasks require subjective information (e.g., community Q&A sources).

Palani *et al.* (2021) investigated search behaviors that predict learning during tasks that involve proposing new solutions to a problem (e.g., how should a city prepare itself for autonomous vehicles). Participants were students in a design-based course. To measure learning, participants completed knowledge summaries before and after the task. Knowledge summaries were scored along four dimensions: (1) inclusion of relevant facts, (2) synthesis of facts, (3) evaluation of facts, and (4) degree of problem definition. The fourth dimension measured the extent to which knowledge summaries described concrete problems to be solved. Different search behaviors were predictive of learning for different dimensions. The number of queries was predictive for all dimensions except for the "evaluation of facts" dimension. The average query length and the number of distinct query-terms across all queries was predictive for all dimensions except for the "inclusion of relevant facts" dimension. The number of pages visited was only predictive for the "inclusion of relevant facts" dimension.

Câmara and El-Zein (2022) developed a framework for predicting learning during search named RULK—Representing User Learning and Knowledge. The framework uses vectors to represent: (1) the ultimate learning objective, (2) the knowledge gained from a document visited during the search session, and (3) the searcher's state of knowledge at a given point during the search session. The learning objective was modeled as a vector representation of the Wikipedia article associated with the topic being learned. The user's current knowledge state was modeled by combining (i.e., summing) the vector representations of documents visited so far during the search session. Finally, learning was estimated by computing the cosine similarity between the vector representation of the participant's knowledge state at the end of the session and the vector representation of the learning objective. The study compared two implementations of the RULK framework. One implementation used a bag-of-words vectors and the other used BERT-based embeddings. Learning estimates were compared to *actual* learning outcomes based on participants' performance on a pre- and post-task assessment. Estimates from both RULK variants positively correlated with actual learning outcomes. These results suggest that searchers learn more when they visit documents that are relevant to the topic of the learning objective.

El Zein *et al.* (2023) experimented with a new implementation of the RULK framework named RULK_{NE}—Representing User Learning and Knowledge with Named Entities. Compared to both RULK implementations previously described, this new implementation used vectors of named entities to represent: (1) the learning objective, (2) the knowledge gained from a visited document, and (3) the searcher's current state of knowledge. By itself, the new RULK implementation did not predict learning as well as the older variants. However, linearly combining all three frameworks performed the best. This result suggests that searchers with better learning outcomes engage with documents that are relevant to the learning objective in terms keywords, latent topics, *and* named entities.

Key Takeaway

Studies have found two consistent trends. Searchers with better learning outcomes: (1) spend more time on pages visited during the session and (2) visit pages that are more relevant to the learning objective.

5.2 Behaviors that Predict Domain Expertise

The studies in the previous section directly focused on understanding search behaviors that predict learning during search. Studies have also investigated the differences in search behaviors between domain experts versus non-experts. Predicting domain expertise is not the same as predicting learning during search. However, the studies in this section provide "food for thought" about search behaviors that may predict learning during search.

In a longitudinal log-based study, White *et al.* (2009) investigated the differences in search behaviors between experts and non-experts in the domains of medicine, finance, law, and computer science. Results found that experts tend to issue more queries, visit more pages, and have longer search sessions. The authors noted two possible explanations for these trends. One explanation is that experts must work harder to find information that is novel to them. Conversely, non-experts have less prior knowledge. Therefore, they tend to seek information that is easier to find. A second possibility is that experts are more persistent during the search session because they seek information that is more important to them. Additionally, experts had more "branchy" search sessions. That is, they were more likely to revisit previously encountered pages in order to follow a new path forward. The authors conjectured that this behavior might be due to experts being better able to explore the space more systematically. Finally, experts visited more pages with technical content, visited fewer commercial websites, and had more successful search sessions. Search success was determined heuristically—sessions that ended with a clicked result were considered successful and sessions that ended with an abandoned query were considered unsuccessful.

Zhang *et al.* (2015) also compared the search behaviors of domain experts versus non-experts. Experts issued more queries, issued longer queries, visited and saved more pages, clicked on lower ranked results, had longer search sessions, and performed more actions during the search session.

Key Takeaway

Domain experts tend to have sessions with more search activity, complex queries, and tend to navigate the information space more systematically.

5.3 Summary

As we have seen, many different studies have investigated search behaviors that predict learning during search. Results from these studies suggest the following four trends. Our own hypothesis is that the first two trends are more likely to generalize across tasks and that the last two trends are more likely to be task-dependent.

First and foremost, searchers with better learning outcomes tend to spend more time on pages visited during the search session (Yu *et al.*, 2018; Gadiraju *et al.*, 2018; Collins-Thompson *et al.*, 2016; Lu and Hsiao, 2017; Syed and Collins-Thompson, 2017a). This is perhaps the most consistent trend across studies. As one might expect, it suggests that learning involves reading and internalizing information.

Second, searchers with better learning outcomes tend to visit pages that are relevant to the learning objective. Several studies point in this direction. Syed and Collins-Thompson (2017a) found that retrieving documents with a greater density of unknown vocabulary terms improves vocabulary learning. Liu and Song (2018) found that participants had better learning outcomes when they engaged with sources that are appropriate to the task—encyclopedic sources during a receptive task and community Q&A sources during a critical task. Critical tasks involve evaluation, which may require reading about people's opinions and subjective experiences. Câmara and El-Zein (2022) and El Zein *et al.* (2023) found that participants with better learning outcomes visited documents that are relevant to the learning objective in terms of keywords, latent topics, and named entities.

Third, searchers with better learning outcomes tend to issue more complex queries (Bhattacharya and Gwizdka, 2019; Yu *et al.*, 2018; Gadiraju *et al.*, 2018). Query complexity has been characterized from different perspectives. Bhattacharya and Gwizdka (2019) used technical terms related to the task. Yu *et al.* (2018) and Gadiraju *et al.* (2018) focused on reading level—complex query-terms tend to be learned by native English speakers at an older age. Palani *et al.* (2021) focused on the number of distinct query-terms used throughout the search session. This trend probably depends on the complexity of the topic associated with the learning objective. For example, Lei *et al.* (2015) had 5th graders learn about the topic of animal courtship and observed that participants had better learning outcomes when they issued a simple yet effective query (e.g., "animal courtship").

Finally, searchers with better learning outcomes visit more diverse pages (Collins-Thompson *et al.*, 2016). As with query complexity, this trend may also depend on the task. It may be true for learning objectives that involve a broad and multifaceted topic.

In this section, we also reviewed studies that have examined the differences in behaviors between domain experts and non-experts. Results suggest that experts expend more effort during their searches and explore the information space in a more calculated manner—returning to previously visited pages to follow a new path forward. Predicting domain expertise is not the same as predicting learning during search. However, future work should consider whether these behaviors may predict learning over an extended period of time. In other words, in the long term, searchers who learn more may start to behave more like experts than non-experts.

6

Tools to Support Learning During Search

Prior studies have explored different tools, visualizations, and ranking algorithms that may improve learning during search. Examples include note-taking and annotation tools, visualizations, goal-setting tools, self-assessment tools, and ranking algorithms that help vocabulary learners target specific keywords.

6.1 Note-Taking and Annotation Tools

Several SAL studies have investigated tools that allow searchers to take notes or highlight and annotate text within documents encountered during the search session.

Freund *et al.* (2016) investigated the effects of different reading environments on reading comprehension. Reading environments varied along two dimensions. The *text presentation* dimension manipulated whether articles were displayed using plain text versus HTML, which included potentially distracting elements such as ads. The *interactivity* dimension manipulated whether participants were provided with tools to highlight text and annotate documents with "sticky notes." Participants had better learning outcomes in the plain text versus HTML condition. However, access to the note-taking tools helped mitigate the negative effect of the HTML condition. That is, when given access to the note-taking tools, participants had similar learning outcomes in the HTML versus plain text condition.

Roy *et al.* (2021) investigated the effects of two tools on learning. One tool enabled participants to highlight text and see a summary of all highlights within a single "your highlights" panel. A second tool enabled participants to take notes, which included copy/pasting information. In isolation, access to either tool improved learning outcomes. In combination, however, access to both tools did not improve learning outcomes, possibly due to cognitive overload.

Qiu *et al.* (2020) conducted a study in which participants were assigned to four conditions. One manipulation involved having participants use a text-based conversational search interface versus a standard search interface. A second manipulation involved having participants take notes versus not take notes. Participants had the greatest knowledge gains when using the standard search interface and when instructed to take notes. The study also considered information loss, by comparing how much participants got correct in a post-task assessment but incorrect in a retention assessment completed several days after the search session. A small (marginally significant) trend found that participants forgot less when using the conversational search interface versus the standard search interface.

The studies above aimed to improve learning during search by enabling searchers to take notes. They did not, however, try to encourage searchers to take *high-quality* notes. Outside of SAL, research in education has examined the effects of note-taking instruction on learning. Chang and Ku (2015) summarizes prior work on the relationship between high-quality note-taking and learning. High-quality notes have several characteristics. First, quantity matters. High-quality notes involve more concepts (or even words) than low-quality notes. Second, high-quality notes make connections between elements, including prior knowledge. Finally, high-quality notes involve representations that are different from the original source.

Chang and Ku (2015) investigated the effects of note-taking instruction on reading comprehension. Students received 40 minutes of note-taking instruction per week for five weeks. Students who took the note-taking skills course improved their reading comprehension more than students in a control group.

Key Takeaway

Note-taking tools have consistently been found to improve learning outcomes during search.

6.2 Visualizations

Kammerer *et al.* (2009) investigated the effects of a search system that enabled participants to filter the search results using social tags. Participants had better learning outcomes with the experimental system versus a baseline system without social tags.

Câmara *et al.* (2021) explored the effects of an experimental search interface that presented participants with their coverage of subtopics explored during the search session. With the experimental system, participants explored more subtopics *superficially* and, ultimately, did *not* have better learning outcomes.

Salimzadeh *et al.* (2021) investigated the effects of displaying entity cards on the SERP. Web search engines typically display entity cards to the right of the search results. Entity cards typically combine information from different structured data sources. Results found that displaying entity cards did not improve learning outcomes.

Column-based faceted browsers are commonly used to explore music collections. Wilson *et al.* (2008) experimented with a "backward highlighting" visualization that highlighted the different paths that can lead to a specific facet category or item in the collection. Participants who used the novel interface were able to recall a larger number of facts about the music collection used in the experiment.

6.3 Goal-Setting Tools

Outside of SAL, prior research in education has shown that effective goal-setting improves learning. In the context of self-regulated learning, effective goal-setting promotes and supports more meta-cognitive monitoring and control.

Prior work has also identified ideal goal characteristics that make them more achievable. In particular, goals should be—(1) difficult; (2) specific; (3) proximal (short-term); (4) learning-oriented (versus performance-oriented); and (5) self-set (Locke and Latham, 1990; Locke and Latham, 2012; Locke and Latham, 2006; Locke and Latham, 2019). With respect to specificity, research has found that ideal goals include a specific action (i.e., cognitive process), content (i.e., material), standard (i.e., criteria to measure progress and success) and allotted timeframe (McCardle *et al.*, 2017).

Urgo and Arguello (2023) experimented with a tool called the Subgoal Manager. The tool was designed to help searchers break apart a learning objective into different subgoals. Using the tool, searchers can explicitly write subgoals, take notes with respect to subgoals, and mark subgoals as completed. Urgo and Arguello (2023) conducted a crowdsourced study with three conditions. In one condition, participants had access to the Subgoal Manager and were asked to set their own subgoals. In a second condition, participants had access to the Subgoal Manager with prepopulated subgoals. In this condition, participants could not add, delete, or modify the prepopulated subgoals. In a third condition, participants were not asked to set subgoals and were simply given a tool to take notes. Participants had slightly better learning outcomes when they had access to the Subgoal Manager and were able to set their own subgoals. Additionally, a qualitative analysis of subgoals found that participants set subgoals of varying quality. Specifically, participants in the self-set subgoals condition often neglected to set subgoals with specific criteria to measure progress and success. Results found that participants who set high-quality subgoals had better learning outcomes than those who set low-quality subgoals.

In a follow-up study (Urgo and Arguello, 2024), participants were assigned to one of two conditions. In one condition, participants had access to the Subgoal Manager and were able to set their own subgoals. In a second condition, participants were not asked to set subgoals and were simply given a tool to take notes. Knowledge gains were captured using a closed-ended multiple-choice test, as well as an open-ended assessment that asked participants to describe everything they learned. Participants completed these assessments immediately after the search task and one week later to measure knowledge retention. Participants had better learning outcomes in the Subgoal Manager condition, particularly with respect to knowledge retention. Additionally, the study used a think-aloud protocol. A qualitative analysis of think-aloud comments and behaviors found that participants in the Subgoal Condition had greater engagement with different self-regulated learning (SRL) processes. This result suggests that goal-setting improves learning during search at least in part *because* it promotes and supports greater engagement with different SRL processes (e.g., meta-cognitive monitoring and control).

Key Takeaway

Goal-setting tools have been found to improve learning partly because they encourange metacognitive monitoring and control.

6.4 Self-Assessment Tools

Syed *et al.* (2020) experimented with a reading environment that dynamically prompted participants to answer questions about paragraphs read during the session. As a starting point, the study did not involve participants searching a document collection. Instead, participants were asked to learn about a topic and were presented with a Wikipedia article. The reading environment automatically predicted which paragraphs participants read using eye-tracking measures. The study involved four experimental conditions. In one condition (Q_{auto}), participants were prompted to answer questions generated by a proprietary Automatic

Question Generation (AQG) system. In a second condition (Q_{human}), participants were prompted to answer manually curated questions originating from the SQUAD dataset (Rajpurkar *et al.*, 2018). In a third condition (Q^*_{human}), participants were prompted to answer manually curated questions that also included one question that required participants to synthesize information from different paragraphs. Finally, in a control condition (Q_{none}), participants were not prompted to answer questions while reading. To measure learning and retention, participants completed short-answer tests before the reading session, immediately after, and one week later. The study considered the effects of the experimental condition for low prior knowledge (low-PK) and high prior knowledge (high-PK) participants.

The study found several interesting trends. First, when prompted to answer questions (automatically generated or manually curated), low-PK participants had better knowledge retention scores. The same trend was not found for high-PK participants, suggesting the prompting people to answer questions is only beneficial for individuals with low prior knowledge. Second, participants scored higher in both the post-task and retention tests when asked automatically generated versus manually curated questions. One possible reason is that automatically generated questions turned out to be more specific than manually curated questions and may have prompted participants to re-read content more closely. Counter intuitively, including a manually-curated synthesis question did not improve learning outcomes.

Key Takeaway

Self-assessment tools help users test their own knowledge. Such tools have been found to benefit searchers with low prior knowledge.

6.5 Ranking Algorithms

Ranking algorithms are responsible for predicting the relevance of documents for a given query. Several studies have investigated whether specific ranking algorithms can improve learning outcomes directly or indirectly.

Syed and Collins-Thompson (2017b) experimented with a retrieval model that was specifically designed to support vocabulary learning. Participants had better learning outcomes when using different versions of the experimental system than when using a commercial web search engine.

Ranking algorithms exploit different types of evidence, including information about a user and a document. Syed and Collins-Thompson (2018) investigated whether the characteristics of documents read during a vocabulary learning search session can predict learning. In addition to measuring learning immediately after the session, the study also considered long-term retention using a delayed vocabulary test administered nine months later. Several document attributes were found to predict learning immediately after the session. For example, participants had higher vocabulary knowledge gains when they interacted with documents that had a higher density of unknown keywords, had contextually relevant (vs. distracting) images, and did not have too many images relative to the amount of text. Additionally, participants who interacted with documents with a greater density of unknown *difficult* keywords had higher retention rates of those same keywords. These results suggest that ranking algorithms to support vocabulary acquisition should exploit such user-document features.

Rokicki *et al.* (2022) used learning-to-rank (LTR) to re-rank search results based on how much they helped participants from a previous study learn during a search session. Documents were assigned gold-standard "relevance labels" based on participants' knowledge gains and how much time they spent on the document during the search session. In addition to features associated with the document's relevance to the query (e.g., BM25 score), the LTR model leveraged features associated with the document's reading level, structure/formatting, and linguistic characteristics (e.g., the presence of words associated with certain cognitive processes). Results found that the LTR model was able to re-rank documents based on how much they helped participants learn during the search session.

Studies have also investigated whether the types of documents visited during a search task affect knowledge gains. Pardi *et al.* (2020) had participants learn about the formation of thunderstorms, and found that participants had better learning outcomes when they interacted with textual documents versus videos. This result suggests that videos may not be as useful as textual documents during conceptual learning tasks.

Key Takeaway

Ranking algorithms can sometimes be optimized to improve learning during search.

6.6 Future Work

Several studies have examined the information-seeking processes of learners. Results from these studies provide insights about future systems that may better support learning during search.

Urgo and Arguello (2022c) investigated the typical "pathways" followed by searchers toward a specific learning objective. Learning objectives were manipulated across three knowledge types (factual, conceptual, procedural) and three cognitive processes (apply, evaluate, create) from A&K's taxonomy (Anderson *et al.*, 2000). Search sessions were recorded and analyzed using qualitative techniques. Search sessions were characterized as a sequence of *learning instances*—instances in which the participant set a new learning-oriented subgoal or seredipitously learned something new based on their think-aloud comments and actions. Among the results, the knowledge type of the objective influenced which cognitive processes were more or less common. For example, procedural objectives involved more creative processes (e.g., modifying or combining procedures). The knowledge type of the objective also influenced the types of transitions observed across "pathways." For example, for conceptual objectives, participants were more likely to transition from complex processes to understand-level processes. The authors proposed different tools to support learners with these processes and transitions.

Li *et al.* (2023) conducted a longitudinal study that investigated the cognitive processes that university students engage in while deciding on a topic for a research project (e.g., a thesis or independent study). Results found that participants typically engaged in the cognitive processes of understand, apply, analyze, and evaluate before deciding on a research topic. Additionally, common transition patterns included: (1) understand - apply - create; (2) understand - analyze - create; and (3) understand - analyze - evaluate - create. The authors proposed different tools that might support learners with these processes and transitions.

Finally, researchers have proposed novel search systems and architectures that may better support learning. Smith *et al.* (2022) proposed a multi-module search environment to support learning by students within the context of a school assignment. The authors proposed that such a system should be able to bias the search results toward documents that are relevant to the assignment. von Hoyer *et al.* (2022b) introduced a novel framework for describing the process through which people search for information in order to learn. The framework situates the learner in a specific learning context, enumerates cognitive activities that searchers typically engage in, and describes different factors that impact behavior. The authors proposed different tools that might support learning during search. For example, the authors proposed that search interfaces should highlight important relations between documents retrieved (e.g., whether information in one document corroborates or contradicts information in another document).

6.7 Summary

To summarize, SAL studies have investigated a wide range of tools to encourage and support learning during search. Study results suggest several important trends. First, note-taking and document annotation tools seem to consistently improve learning outcomes. Research in education suggests that tools to help searchers become more effective note-takers are worth exploring in future work. Second, tools and visualizations may not always result in better learning outcomes, for several reasons. Tools can lead to gamification. For example, in one study, visualizing the topical coverage of documents visited during the search session resulted in participants exploring more documents *superficially* to cover more subtopics. However, because participants did not engage in deep reading, they did not have better learning outcomes. Studies also suggest that access to too many tools can result in cognitive overload. Third, studies have found that tools may benefit some searchers but not others. For example, a tool that prompted participants to answer questions about passages read during the session improved retention for participants with low prior knowledge but not for participants with high prior knowledge.

Fourth, tools to support learning need not be complicated. For example, in one study, a tool that allowed participants to write subgoals, take notes with respect to subgoals, and mark subgoals as completed resulted in greater knowledge retention. Fifth, studies have found that engagement with some documents results in better learning outcomes. Interestingly, important document attributes may depend on the learning task. For example, one study found that textual documents improved learning more than videos for a conceptual learning task. Therefore, the types of documents that promote learning may depend on the type of learning objective. Finally, several papers have suggested future tools to support learning. For example, future tools could highlight relations between documents (e.g., one document corroborating or contradicting information in another document).

7

Self-Regulated Learning (SRL)

In SAL, researchers aim to better support learning during search. Self-regulated learning (SRL) is an active and reflective process in which learners engage in their own learning by selecting strategies, monitoring progress, and adapting when necessary to achieve their learning goals. SRL supports SAL in two critical ways. First, SRL has been shown to increase learning outcomes and, therefore, is likely to increase learning outcomes during search. Facilitating and encouraging SRL during search is an important area of focus for future research in SAL. Second, understanding the learning process during search is crucial for supporting learning. Capturing SRL during search can help researchers to understand *when* and *how* learning is occurring during search. In this section, we provide an overview of SRL, review different models and frameworks of SRL, and describe how SRL processes might be captured and studied in the context of SAL research.

7.1 Models of SRL

Researchers in the learning sciences have investigated the critical role of effective SRL in learning achievement (Boekaerts *et al.*, 1999; Sitzmann and Ely, 2011; Winne and Hadwin, 1998; Schunk and Swartz, 1993;

Zimmerman and Schunk, 2011). Prior work has found that SRL processes are important to improving learning outcomes (Zimmerman and Pons, 1986; Zimmerman and Martinez-Pons, 1988; Schunk, 1984; Schunk, 1981).

SRL is an active, reflective process in which a learner monitors and controls their learning to achieve their learning goals (Winne, 2001; Zimmerman and Schunk, 2011; Schunk, 2001). Through goal-setting, self-regulated learners generate feedback loops allowing for monitoring of progress and use of strategies when current approaches are not producing the desired learning outcomes. Several models of SRL emerged from prior work that investigated how learners engage in complex tasks (Winne and Perry, 2000; Pintrich, 2000; Zimmerman, 2000; Zimmerman, 2002; Boekaerts *et al.*, 1999).

Key Takeaway

Self-regulated learning (SRL) is an active, reflective process where learners monitor and control their own learning to achieve their learning goals.

Zimmerman's model of SRL (Zimmerman, 2000) is rooted in socio-cognitive theory (Bandura, 1986), highlighting the social foundations of human cognition and behavior. This perspective emphasizes internal and external contextual factors that influence SRL, arguing that learners' self-regulatory processes are shaped by their interactions with their social environment. Social interactions might include teachers providing feedback and encouragement, as well as students working together and implementing effective strategies to support their learning.

Pintrich's model of SRL (Pintrich, 2000) also comes from a socio-cognitive approach and focuses on learner motivation (Puustinen and Pulkkinen, 2001). This model includes elements of self-efficacy, task value, and goal orientation. Self-efficacy relates to a learner's confidence in their ability to achieve an objective (Locke, 2001). Task value includes a learner's interest in an objective, personal significance of an objective, or the usefulness of an objective toward future goals. Goal orientation relates to a learner's goals being learning-oriented versus performance-oriented. Learning-oriented goals are focused on internal growth. Performance-oriented goals are focused on external feedback and validation. This model proposes that motivation is a critical component in regulating one's learning (Pintrich and Groot, 1990).

Boekaerts's model of SRL (Boekaerts *et al.*, 1999) also focuses on learner motivation and was influenced by Action Control Theory (Kuhl, 1985). Action Control Theory examines how individuals manage their actions and goals, particularly in situations that require sustained effort and persistence (e.g., student working on a semester-long project). Boekaerts's model of SRL also highlights the role of volition (i.e., willpower) in maintaining motivation when faced with competing impulses, distractions, or habits (Boekaerts, 1995; Boekaerts and Cascallar, 2006).

The Winne & Hadwin model of SRL (Winne and Hadwin, 1998) builds on the work of Bandura and Zimmerman, highlighting metacognitive knowledge and metacognitive skills. Metacognitive knowledge includes a learner's knowledge about learning strategies and awareness of their own cognitive processes. Metacognitive skills include a learner's ability to monitor and control their own learning by adapting or changing strategies when necessary.

In this monograph, we focus on the Winne & Hadwin model of SRL for two main reasons. First, the Winne & Hadwin model emphasizes metacognitive knowledge and skills. In a classroom environment, feedback, guidance, and strategies can be provided by instructors and other students. In contrast, metacognitive knowledge and skills are particularly important in learning during search as learners must self-direct, reflect on, and make adaptations to support their own exploration and comprehension of new information. Second, the Winne & Hadwin model is supported by evidence from many empirical studies and provides a rich framework for understanding and capturing SRL processes (Greene and Azevedo, 2007; Boom *et al.*, 2007; Glogger *et al.*, 2012; Santhanam *et al.*, 2008; Bannert *et al.*, 2009; Kistner *et al.*, 2010; Azevedo *et al.*, 2002).

Key Takeaway

The Winne & Hadwin model of self-regulated learning emphasizes metacognitive knowledge and skills and is well-supported by empirical research demonstrating that the model's stages align with effective learning processes.

7.2 Winne & Hadwin (W&H) Model

The Winne & Hadwin (W&H) model of SRL is depicted in Figure 7.1. The W&H model of SRL consists of four weakly iterative phases—(1) task definition and understanding; (2) setting goals and plans; (3) selecting and enacting strategies and tactics; and (4) making adaptations. The phases of the W&H model of SRL are *weakly iterative* in that learners often move through the phases in sequence but may return to earlier phases when modifying goals, changing strategies (i.e., selecting a different learning strategy), or after making adaptations (i.e., making a large-scale shift or tactical approach to a learning plan).

In the first phase, the learner develops a working definition of the learning task. This definition is developed using external and internal resources. External resources may include a task description or a teacher's instruction. Internal resources may include prior knowledge and perceived abilities. The task definition phase influences all subsequent SRL phases. Since the initial task definition determines what the learner believes they need to accomplish, it directly shapes the types of goals they set, the strategies they choose to achieve those goals, and the standards with which they ultimately evaluate their performance in all later phases of the learning process.

In the second phase, the learner sets goals and makes a plan to complete the task as defined in phase one. Learners also choose cognitive

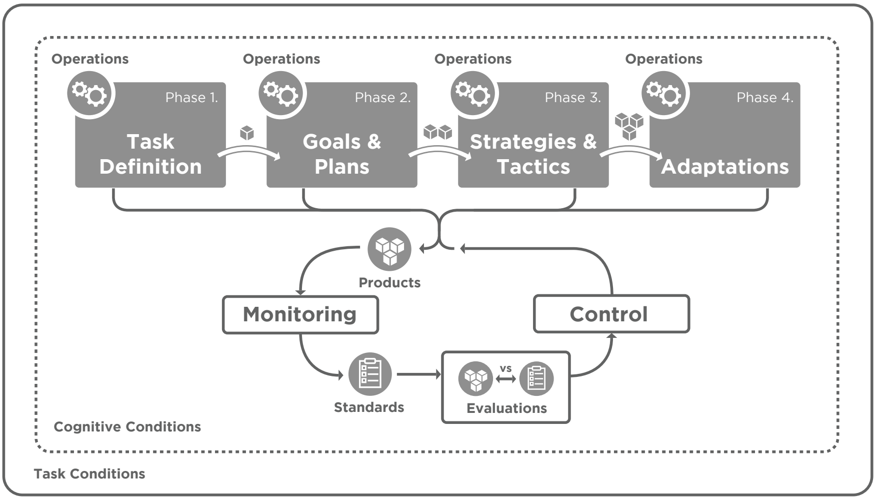

Figure 7.1: Conceptualization of the Winne and Hadwin Model of SRL.

strategies and tactics (e.g., find three examples of surrealism, be able to state the definition of automatism in my own words) deemed necessary to achieve the goals. These goals are modeled as "Standards" in Figure 7.1 and are the criteria used during metacognitive monitoring to make decisions about goal progress and if a goal has successfully been achieved. Goals and standards are dynamic. They may be updated in subsequent cycles as deeper understanding is acquired.

In the third phase, the learner uses the strategies and tactics identified in phase two. During this phase, learners construct information in their working memory. This involves processing the information relevant to the task, integrating the information into their existing knowledge/understanding, and internalizing the information in a way that enables effective recall and future use (Winne, 2022). In general, this phase involves a learner implementing activities or studying skills in order to learn the material. Feedback is a natural by-product of the third phase. As a learner implements strategies and tactics to achieve their goals, external and internal feedback may cause the learner to modify or update their goals. For example, in a time-constrained situation, external feedback might come in the form of a timer indicating how much time is left. On the other hand, internal feedback might involve a learner judging how well they understand something. Particularly relevant to SAL, Winne identified a common search behavior as an important phase three SRL process, "For example, a search query may be deemed unproductive because results were not what was expected or don't satisfy the standards for particular information" (Winne, 2022, p. 78).

The fourth phase involves learners making adaptations. This is a time of disengagement from the task itself to self-reflect, where a learner looks back and evaluates the success and/or failure across the previous phases. Based on the feedback from this phase, a learner may identify adaptive or maladaptive behaviors that helped or hindered the learning process. Adaptive behaviors are actions that include adjusting strategies or strategically adapting task perceptions and plans. Maladaptive behaviors include failing to interact productively, using the same strategies when negative judgments of learning have occurred, and ignoring experienced challenges (Sobocinski *et al.*, 2020). In the adaptation phase, a learner responds to this evaluation (i.e., what worked and what did not work) by making a large-scale adjustment. For example, a learner might stop working on a problem after realizing that gaps in their conceptual understanding are too large. In response, they might pivot to finding and internalizing information about relevant concepts. Prior work indicates that students may lack the necessary metacognitive skills or are not sufficiently motivated to engage in this complex and effortful process. Learners do not see the added value of engaging in these important self-reflective processes. Instead, students often focus on more immediate task completion processes rather than long-term learning improvement (Hadwin *et al.*, 2001; David *et al.*, 2024).

Throughout each phase, the learner enacts metacognitive monitoring and control. Metacognitive monitoring involves the learner comparing the products they have generated in the learning process to the standards they have set for themselves in their goals. For example, imagine a learner with a goal of being able to define a concept in their own words. At some point, the learner might monitor their progress and ask themselves: "At this point, can I define the concept in my own words?" Here, the learner might enact metacognitive monitoring to compare a product with a standard. The product is the definition they can produce in their own words and the standard is the correct definition of the concept. If the learner finds a mismatch (i.e., decides that their own definition deviates from the correct one), they might engage in metacognitive control. Controlling might include implementing a different strategy or tactic or modifying a goal. For example, the learner might decide to re-read their notes to better internalize information relevant to the concept or find a source that explains the concept in more detail.

Key Takeaway

The W&H model of SRL involves four phases. Throughout each phase, the learner monitors their goal progress and enacts control if they are not meeting their goal standards.

7.2.1 An Example of SRL in SAL

In order to better understand SRL in the context of SAL, consider a person using a search environment to learn about a topic. Imagine that a graduate student named Cara is tasked with completing a literature review on the impact of social media on mental health. Cara might move through the phases of SRL outlined in Figure 7.1 by doing the following:

- *Phase 1 - Task Definition:* Cara determines the scope and boundaries of the review, limiting the literature review to only the top five social media platforms and identifying the core concepts of anxiety, depression, and self-esteem to cover within the review (pulling from a discussion Cara had with her thesis advisor).
- *Phase 2 - Goals & Plans:* Cara sets specific goals for herself. For example, Cara writes that she will find and summarize findings from

at least 5 peer-reviewed references for each of the core concepts (i.e., anxiety, depression, and self-esteem) each week. Additionally, Cara identifies the keywords to combine during her search (e.g., "anxiety," "depression," "self-esteem," "mental health," "influence," "impact," "Facebook," "Instagram," "TikTok," "YouTube," "WhatsApp") and databases she will explore (e.g., Google Scholar, PubMed, APA PsycArticles) to complete her goals.

- *Phase 3 - Strategies & Tactics:* Cara begins by searching for information on the influence of social media on anxiety specifically as it is her first goal. Cara engages in various strategies such as skimming articles, summarizing findings in her notes, and citation chaining (i.e., investigating references cited in a highly relevant, methodologically rigorous publication).

- *Phase 4 - Adaptations:* Cara realizes there is a whole area of research exploring the impact of social media on sleep patterns. Cara decides to pause this area of investigation until she can discuss whether or not sleep patterns should be included in the review with her advisor (i.e., a more knowledgeable other, Vygotsky, 1980).

- *Metacognitive Monitoring & Control:* Cara engages in metacognitive monitoring and control across each of the phases:

  - In the *Task Definition* phase, Cara activates her prior knowledge and realizes that although she has some familiarity with her target mental health concepts, she is unclear of the boundary between anxiety and depression and how the topics of anxiety, depression, and self-esteem may influence one another (i.e., monitors understanding). She decides to add this investigation as a goal in the next phase (i.e., enacts control).
  - In the *Goals & Plans* phase, Cara monitors goal progress after reading several sources and decides that the review would not be comprehensive without the factor of addiction

and compulsive behavior and includes it as an additional goal (i.e., enacts control).

- In the *Strategies & Tactics* phase, Cara evaluates the usefulness of sources (i.e., monitors sources) when looking through results on the SERP after issuing the query "influence of social media on anxiety." Cara decides that the sources are too broad and decides to modify her search (i.e., enacts control) by adding specific social media platforms to each subsequent query.
- In the *Adaptations* phase, Cara meets with her advisor and realizes she cannot answer specific questions her advisor poses about particular study results and the definition of factors like compulsive behavior (i.e., monitors understanding). Cara decides to allocate an additional day in her schedule to review details about the factors and findings from the literature review before each meeting with her advisor (i.e., enacts control).

As demonstrated, Cara reviews and concretizes the task, sets specific goals, engages in useful strategies, adapts when necessary, and continuously engages in metacognitive monitoring and control to navigate her SAL research process. SRL in the context of SAL involves making strategic decisions throughout the learning process, such as determining the scope of her literature review. The above example shows how the W&H model of SRL helps in facilitating, organizing, and completing complex, evolving learning-oriented search tasks.

SAL research has begun to reveal the complex, iterative, and multidimensional nature of learning-oriented search tasks (Rieh *et al.*, 2016; Bhattacharya, 2023). In particular, prior SAL work has shown that learning-oriented search tasks contain multiple learning-oriented subgoals that must be achieved to complete the overall learning objective (Liu and Belkin, 2010; Urgo and Arguello, 2022c; Urgo and Arguello, 2023; Zhang and Liu, 2023). To better facilitate and improve learning during search, SAL researchers should focus on supporting learners' goals and subgoals. To this end, researchers can leverage both the W&H model of SRL, which emphasizes goals as a central function, along with decades of prior work in goal-setting that have determined goal characteristics that make goals more achievable. In the next section, we explore the important function of goals in SRL and research on goal characteristics that lead to better outcomes.

7.3 Goal-Setting and SRL

Goal-setting is identified across SRL models as an important part of the self-regulation process. Goals are particularly central to the W&H model of SRL, having three main functions. First, goals prompt learners to consider their understanding of the task. Second, goals direct learners' attention toward planning and choosing strategies to achieve their goals. Third, goals provide standards for monitoring and evaluating progress (McCardle *et al.*, 2017). Here, we provide a detailed overview of the underlying mechanisms of goals that account for improved learning outcomes and the qualities of goals that make them more achievable.

7.3.1 Ideal Goal Characteristics

Goals are broadly defined by a learner's purpose and are characterized by quantity, quality, and performance or attainment (Locke and Latham, 1990). Goal-setting is the creation of an objective that defines the aim of the learner's actions (Schunk, 2001). Locke and Latham spent more than half a century investigating goal-setting (Locke and Latham, 2012). Their analysis across hundreds of studies indicates several enduring and key goal characteristics that affect goal achievement: (1) difficulty; (2) specificity; (3) proximal versus distal goals (i.e., short-term versus long-term); and (4) learning versus performance goals (Locke and Latham, 1990; 2006; 2012; 2019). Here, we provide details on prior work that investigates these four goal-related factors that make goals more achievable.

Key Takeaway

Goals are more achievable when they are difficult, specific, short-term, and learning-oriented.

First, goal difficulty has an effect on achievement. Locke and Latham (1990) coined the *goal difficulty function* characterized by a positive linear function between the difficulty of a goal and performance. When completing goals of the highest difficulty, individuals exert the highest levels of effort and performance, with performance leveling off only when individuals reach the end of their ability. A natural question is: How is goal difficulty defined? Locke *et al.* (1989) generally define the degree of difficulty of a goal as the probability that a goal can be reached. Goal difficulty has typically been set relative to the context of the goal. For example, LePine (2005) set easy and difficult goals at approximately one standard deviation, respectively, below and above the mean performance level of a given goal. Additionally, Earley (1985) tasked participants in a pilot study with completing mock course schedules with particular constraints (e.g., no two natural science classes could be scheduled within 1 hour of another) to determine goal difficulty for a subsequent study. Pilot study results found that 4% of subjects could complete 10 products in 15 minutes and that no subjects could complete more than 3 products in 5 minutes. From this data, completing 10 schedules in 15 minutes was considered a *very difficult* goal. Much of goal research has been in either organizational or athletic settings (Latham, 2016). Less work has focused on academic goals (Acee *et al.*, 2012; McCardle *et al.*, 2017). Alessandri *et al.* (2020) examined the effect of academic goal difficulty on students' final grades. Results found a non-linear relationship between difficulty and final grades, where goal difficulty was a moderating factor between self-set academic goals and students' final course grade. Students who set moderately difficult goals tended to have higher daily study performance, leading to higher final course grades.

Second, goal specificity has an effect on achievement. Research has shown that difficult goals that are also *specific* lead to higher levels of performance (Locke and Latham, 2002). Locke *et al.* (1989) define the specificity of a goal in relation to the vagueness of a goal. While vague goals can be interpreted in various ways by different people, specific goals reduce variability of interpretation. This definition of goal specificity is in line with that of Hollenbeck and Klein (1987), who argue that "there are innumerable outcomes that could be consistent with a vague goal" (Hollenbeck and Klein, 1987, p. 214) Locke *et al.* (1989, p. 272) offer a continuum of example goals that start vague and increase in specificity:

- **Most vague**: improve the performance of your division;
- **Less vague**: increase the profits of your division;
- **Less specific**: increase profits by 10% or more;
- **Most specific**: increase profits by exactly 15%.

The most vague goal has multiple interpretations (i.e., "performance" can be defined in multiple ways). Conversely, the most specific goal limits the number of allowed actions (i.e., increase profits) and outcomes (i.e, by exactly 15%). Locke and Latham (2002) argue that less specific goals or "do your best" goals have no external referent, being defined subjectively rather than from some clear objective resource. In other words, "doing your best" will vary from person to person. Some will tend to satisfice and others will be quite thorough. In contrast, if a goal has an external referent or standard, then it will be objectively clear when the goal is achieved. Specific goals lack ambiguity, which reduces variance in performance. However, for performance to be high, Locke *et al.* (1989) argue that specific goals must also be difficult. As an example, Latham and Seijts (1999) used the following general goal versus specific and difficult goal:

- **General Goal**: Exerting high effort to make money typically results in high profits. Hence, it is important that you do your best to make as much money as possible.

- **Specific and Difficult Goal:** Exerting high effort to make money typically results in high profits. Hence, it is important that you commit to a specific difficult yet attainable goal to make money. In previous sessions, the average profit people earned was $8.71. Your goal should be to make $8.71 or more.

In terms of goal specificity, prior work has highlighted several *dimensions* of specificity that are important for goal achievement. McCardle et al. assessed micro-level goals that students set in a study session and identified four themes of specificity: *time*, *actions*, *standards*, and *content*, collectively called TASC. The first three dimensions (i.e., *time*, *actions*, and *standards*) are rooted in the literature on ideal goal characteristics. First, specifying a goal's *time* or timeframe relates to both specific and proximal goal characteristics, shown to be optimal in prior work (Locke and Latham, 2012). Goals that are more specific and goals that are short-term are more likely to be achieved. Second, specifying a goal's *actions* involves including cognitive processes that will be engaged to complete the goal, such as identify, apply, or evaluate. Prior work has found that learning goals that involve specific *actions* (e.g., "discover n shortcuts" or "produce n schedules") improve outcomes during complex tasks (Winters and Latham, 1996). Participants with action-specific goals also reported higher rates of self-efficacy (i.e., confidence in one's own ability to achieve a particular goal or task).

Third, *standards* relate to the specificity of success criteria for determining the degree to which a goal has been achieved. This type of specificity provides a clear reference point for judging progress in achieving a goal. For example, prior work has found that a "complete n correct class schedules" goal has higher rates of performance and self-efficacy than a "do your best" goal (Seijts and Latham, 2001). Finally, specifying goal *content* was introduced by McCardle et al. They argued that specifying a goal's *content* (e.g., facts, concepts, procedures) is the "foundation of effective learning goals because it focuses attention on the substance of learning rather than a sequence of tasks to complete" (McCardle *et al.*, 2017, p. 2156) Further, they argued that specific *content* helps learners focus on relevant material and actions necessary for learning. Overall, goals with all four TASC criteria (i.e., *time*, *action*, *standard*, and *content*) allow learners to—(1)

be aware of what and how they will learn; (2) monitor goal progress; and (3) notice discrepancies between standards and current outcomes and make adjustments to strategies and/or goals as needed.

Third, the time frame of the goal has an effect on achievement. Goals can be either proximal (short-term) or distal (long-term). Proximal goals support achievement by increasing motivation and increasing self-efficacy through improved detection of errors (Schunk, 1991). Proximal goals have been found to be particularly helpful when the primary distal goal is complex (Latham and Locke, 2007).

Distal or primary goals can be broken down into the related proximal goals or subgoals that help in attaining the distal goal. "Any distal or long-term goal can be segmented into several smaller and more immediate subgoals. These subgoals are intermediate steps to attaining the distal end goal and can be pursued in a sequential way" (Locke and Latham, 2012, p. 185). As an example, writing a literature review over the next two months (a complex distal goal) can be broken down into several proximal subgoals. These proximal subgoals might include: (1) finding and categorizing articles in week one, (2) summarizing articles in week two, and (3) making an outline in week three. Completing each proximal subgoal is a way to measure forward progress toward the larger distal goal. Overall, results have shown that setting proximal goals in support of a distal goal leads to higher achievement (Latham and Seijts, 1999).

Fourth, learning goals have a different effect than performance goals on achievement. Elliott and Dweck (1988) helped to define the categories of performance-oriented goals versus learning-oriented goals. Performance goals (also referred to as *outcome* goals (Locke and Latham, 2006; Latham and Brown, 2006)) are concerned with the outcome of a task. The individual is motivated by demonstrating their abilities and outperforming others. In contrast, learning goals are concerned with developing new skills and gaining knowledge. The individual is motivated by personal growth, development, and expanding their capabilities. Further, "students with learning goals are interested in acquiring new skills and improving their knowledge, even if they make some mistakes. On the other hand, students with performance goals are usually interested in obtaining positive evaluations of ability and avoiding negative evaluations" (Valle *et al.*, 2003, p. 72). Learning goals are established from someone's intention to develop or improve a skill and are rooted in a desire to learn. Conversely, performance goals are established from someone's intention to demonstrate competence to others and are rooted in a desire for positive external feedback. Learning goals and performance goals have been shown to effect learning outcomes of students.

Prior work has demonstrated the impact of learning goals versus performance goals on learning outcomes. Schunk (1996) investigated how learning and performance goals affect motivation and achievement outcomes. In one condition, students were given a learning goal from the teacher (i.e., "You'll be trying to *learn* how to solve fraction problems where the denominators are the same and you have to add the numerators."). In the other condition, students were given a performance goal from the teacher (i.e., "You'll be trying to *solve* fraction problems where the denominators are the same and you have to add the numerators"). While the differences were subtle, results found that students with learning goals had higher motivation and performed better on a math problem-solving assessment than students with performance goals. McNeil and Alibali (2000) investigated the effect of goal type on learning outcomes. Children in the study were either given performance goals, learning goals, or no goals toward problem-solving in math. While the differences leveled off after a two week period, results found that children who were given learning goals (versus performance goals) were more likely to gain conceptual knowledge as reflected in their performance on an immediate post-test. Additionally, children who were given goals of any type (versus no goals) were more likely to transfer their knowledge beyond the mathematical procedure introduced in the class.

Prior work has also investigated the influence of self-set versus externally assigned goals. The relationship between the source of goal-setting on learning outcomes, however, is not entirely straightforward. Locke and Latham (2012) examined a series of 11 studies on the effects of self-set versus assigned goals on performance. Results found that "when goal difficulty is held constant, an assigned goal is as effective as one that is set participatively" (Locke and Latham, 2012, p. 10). However, they noted that the logic or rationale behind assigned goals must be given for these results to hold true. When goals are self-set they are often selected by the probability of attainment determined by the learner. Results have also found a different affect. Azevedo *et al.* (2002) investigated differences between students that followed self-set goals versus teacher-set goals when learning using a web-based simulation environment. Results found that students with self-set goals had higher conceptual learning gains and engaged in more effective SRL processes like judging their own understanding, activating prior knowledge, and setting new subgoals. From this work, the authors argued that "allowing students to set learning goals can enhance their commitment to attaining them, which is necessary in order for goals to affect performance" (Azevedo *et al.*, 2002, p. 6).

Key Takeaway

Self-set goals have been shown to positively impact learning outcomes compared to assigned goals.

7.3.2 Mechanisms of Goals that Improve Performance

An important question is: Why do goals improve performance? Locke and Latham outline four mechanisms through which goals affect performance:

1. Goals direct attention and effort toward relevant activities and away from irrelevant activities.
2. Goals set at higher levels of difficulty lead to greater effort than goals set at a lower level of difficulty.
3. Goals affect persistence, with more difficult goals prolonging effort.
4. Goals affect action indirectly, instigating the "arousal, discovery, and/or use of task-relevant knowledge and strategies" (Locke and Latham, 2002, p. 707).

There are several ways in which goals may *indirectly* affect learners' actions. In response to goals, individuals may automatically recognize and implement relevant skills and knowledge they already possess. If relevant skills and knowledge are not available, then individuals may use skills and knowledge from related contexts that may apply to the current one. If a task is completely novel, individuals begin to develop new, relevant strategies for goal attainment. Self-efficacy impacts the effectiveness of these strategies. Self-efficacy refers to an individual's beliefs or confidence in their ability to achieve a particular goal (Bandura, 2010). Individuals with high self-efficacy develop strategies that are more effective than those with low self-efficacy (Bandura, 2010). Those with high self-efficacy approach challenging tasks with perseverance and effective strategies such as positive visualization and mental rehearsing. In contrast, those with low self-efficacy doubt their own capabilities and engage in counter-productive strategies such as dwelling on failures and personal shortcomings. Self-efficacy can be increased through practice, role modeling, encouragement, and re-contextualizing nervousness as excitement rather than fear (White and Locke, 2000).

Zimmerman (2008) explored motivational influences of goals in academic settings, further supporting the underlying mechanism of goals posed by Locke and Latham. Zimmerman underscored that students with goals put attention into task-relevant activities and away from non-relevant activities, produce higher levels of effort, have greater persistence over time, and higher levels of attentiveness, self-satisfaction, and lower defensiveness. Overall, his findings support that goals help learners to focus and apply more effort over time toward their ultimate learning objectives.

Feedback also plays a critical role in goal achievement. Feedback is the indicator of progress toward a goal. Without feedback, it may be difficult or impossible to engage in the metacognitive monitoring and control necessary to achieve a particular goal. If feedback indicates that goals are not being met, individuals may change strategies or increase their effort (Latham and Locke, 2007). Feedback can be internal or external. Internal feedback is initiated by metacognitive monitoring, comparing products or outcomes with standards (i.e., criteria for optimal completion of a goal). If a discrepancy between these is identified, internal feedback can prompt a learner to change strategies or tactics (metacognitive control). External feedback comes from outside the learner. This feedback might be delivered by teachers, teaching assistants, peers, or systems (Chou and Zou, 2020). Generally speaking, learners are more effective when they respond to external feedback on goal progress (Butler and Winne, 1995).

Key Takeaway

Internal and external feedback are important factors in goal achievement that allow learners to better understand their goal progress.

In summary, there are particular characteristics that make goals more achievable, namely that goals are *difficult*, *specific* (specifying *time*, *action*, *standard*, and *content*), *proximal*, and *learning-oriented*. Additionally, learning outcomes are improved by providing goal feedback and may be improved when goals are self-set rather than externally-assigned. Generally, goal-setting is an important part of the self-regulation process. Goals are particularly central to the W&H model of SRL, having three main functions. First, goals prompt learners to consider their task understanding. Second, goals direct attention toward planning and affect strategy choice for achievement. Third, goals provide standards for monitoring and evaluating progress (McCardle *et al.*, 2017).

7.4 Capturing SRL Processes During Search

In order to better understand when, where, how, and why SRL occurs, researchers have implemented one of two main strategies—(1) questionnaires and (2) think-aloud protocol.

With respect to questionnaires, researchers have developed various self-report inventories to capture SRL processes. The Metacognitive Awareness Inventory (MAI) (Schraw and Dennison, 1994) consists of two dimensions: *knowledge of cognition* and *regulation of cognition*. The first dimension aims to understand metacognitive knowledge or an individual's declarative and procedural self-knowledge. This dimension includes items such as, "I try to use strategies that have worked in the past" and "I understand my intellectual strengths and weaknesses." The second dimension aims to understand metacognitive skills or an individual's ability to monitor and control learning such as planning and evaluating one's own learning. This dimension includes items such as, "I ask myself questions about the material before I begin" or "I find myself pausing regularly to check my comprehension." The Motivated Strategies for Learning Questionnaire (MSLQ) consists of two dimensions, one of which focuses on SRL. This dimension assesses both *cognitive strategy use* (e.g., "When I study I put important ideas into my own words") and *self-regulation* (e.g., "I work on practice exercises and answer end of chapter questions even when I don't have to.") (Pintrich and Groot, 1990). Finally, the Learning and Study Strategies Inventory (LASSI) consists of 10 dimensions: anxiety, attitude, concentration, information processing, motivation, selecting main ideas, self-testing, test strategies, time management, and using academic resources (Weinstein *et al.*, 1987). While some SAL studies have implemented SRL questionnaires (Crescenzi *et al.*, 2021; Hoyer *et al.*, 2022a), researchers should implement this method with caution. Prior work has shown that learners may not accurately report their own SRL processing (Winne *et al.*, 2002).

In contrast to questionnaires, using think-aloud protocol allows researchers to capture SRL through in-the-moment observation of comments and behaviors during search (Urgo and Arguello, 2022a). With this method, researchers ask participants to think aloud while searching and record their search behaviors. Then, researchers review transcripts and videos of think-aloud comments and search behaviors to code instances of particular SRL processes. Greene et al. have coded think-aloud data into macro-SRL and micro-SRL processes (Greene *et al.*, 2012; Greene *et al.*, 2015; Greene *et al.*, 2018). Macro-SRL processes include those related to the main components of the W&H model of SRL (e.g., *planning*, *monitoring*, *strategy use*), while micro-SRL processes are categorized within macro-SRL processes (e.g., developing subgoals is a micro-SRL process within the macro-SRL process of *planning*).

Key Takeaway

SRL can be measured with questionnaires or through think-aloud data. Think-aloud data is reflective of *actual* SRL processes while questionnaires capture perceptions of such processes.

Urgo (2023) used the think-aloud method to capture SRL during search. Results found that particular SRL processes were more frequently engaged when participants had access to a goal-setting tool. These results helped contextualize *why* participants may have had higher learning retention outcomes in the study. This method of capturing SRL more directly matches the dynamic adaptive process of SRL, in that it allows researchers to observe SRL in real-time over a series of events across a learning task (Greene *et al.*, 2013).

Recently, Hadwin *et al.* (2025) developed the Self-Regulated Learning Profile and Self-Diagnostic (SRL-PSD) instrument that collects learners' beliefs about their self-efficacy and abilities, their perceptions of engagement in SRL processes, and the challenges they encountered during learning. Results found that perceptions of task understanding from the SRL-PSD predicted student learning outcomes. Hadwin *et al.* (2025) argue, "Rather than pitting self-reports against multi-modal or behavioral indicators of self-regulated learning, we advocate for advancing self-report measures as essential sources of information about the beliefs, actions and experiences learners perceive to be important during their studying that can inform the interpretation of other multimodal behavioral and physiological indicators" (Hadwin *et al.*, 2025, p. 2-3). In other words, rather than choosing between think-aloud protocols and traditional SRL questionnaires, SAL researchers might consider adapting the SRL-PSD framework to complement existing methods. This could provide a middle-ground approach that captures learners' recent search experiences, perceptions, and challenges while maintaining practical feasibility for larger-scale studies.

It is important for SAL researchers to understand the benefits and drawbacks of SRL methods in order to decide which methods are best suited for a study. On one hand, through in-the-moment observations, a think-aloud protocol allows researchers to better understand *when*, *where*, *why*, and *how often* participants engage in specific SRL processes. However, this method is quite time intensive as it requires multiple researchers to code and measure agreement to arrive at counts of SRL processes from search sessions. On the other hand, capturing SRL processes through a questionnaire requires relatively little time to capture and measure. However, the accuracy of such methods can be questionable as stand-ins for *actual* SRL and do not allow researchers insights into when, where, why, and how often SRL processes occur.

7.5 Tools to Support Effective SRL

Outside of SAL, SRL has been leveraged to support deeper learning in hypermedia environments. In the learning sciences, learning environments have integrated mechanisms to monitor and support SRL engagement. MetaTutor is a particularly effective, nuanced, and enduring example of such an SRL-supportive learning environment (Azevedo *et al.*, 2009; Azevedo *et al.*, 2012). This system and others are described in this section as exemplary prior work that demonstrate implementation of SRL support in computer-based environments. These systems provide valuable insights for understanding and advancing SRL integration into SAL environments.

MetaTutor is an intelligent tutoring system and hypermedia learning environment focused on supporting the acquisition of conceptual knowledge, specifically complex biological content (e.g, circulatory, digestive, and nervous systems). Research in SRL has shown that "although all students have the potential to regulate, few students do so effectively, possibly due to inefficient or insufficient cognitive or metacognitive strategies, knowledge, or control" (Azevedo *et al.*, 2013, p. 430). To facilitate effective SRL, MetaTutor has two main functions. First, MetaTutor is designed to teach and train self-regulation to students through modeling and scaffolding metacognitive monitoring, encouraging effective learning strategies, and facilitating goal-setting. Second, MetaTutor is a research tool that collects SRL trace data related to cognitive, affective, and metacognitive processes engaged by students while learning.

MetaTutor is grounded in SRL theory, situating learners as active constructors of knowledge. As mentioned above, SRL learners set goals toward a learning objective and attempt to monitor, regulate, and control cognitive and metacognitive processes in an effort to achieve those goals (Winne and Hadwin, 1998). MetaTutor provides a series of pedagogical agents designed to support specific SRL processes (e.g., planning, strategy use, and monitoring).

MetaTutor supports SRL processes in several specific ways. First, MetaTutor has four pedagogical agents (PAs) that guide and prompt students throughout the learning process. These four PAs are Gavin the Guide, Pam the Planner, Mary the Monitor, and Sam the Strategizer. Gavin the Guide provides guidance and explanations of the MetaTutor environment from start to end. Gavin also administers the pre- and post-test knowledge assessments and self-report measures. Pam the Planner supports and emphasizes planning, developing subgoals, and activating prior knowledge. Mary the Monitor prompts and supports monitoring processes. There are four main monitoring processes Mary supports: (1) judgment of understanding (i.e., a learner's awareness of what they do or do not understand); (2) feeling of knowing (i.e., a learner's awareness of having some understanding of a topic from the past); (3) content evaluation (i.e., monitoring relevance of content relative to a goal); and (4) monitoring progress toward goals, (i.e., assessing whether a previously set goal has been met) (Trevors *et al.*, 2014). Sam the Strategizer prompts students to summarize content and gives feedback on the quality of summaries. Summary quality is judged by length and number of keywords.

All PA actions are pre-programmed interactive sequences. Such PA actions are initiated according to a particular rule set in the MetaTutor system. For example, Mary the Monitor asks if a student has completed a subgoal if they have spent too much time on a particular subgoal.

This rule is initiated if a student spends more than 20 minutes on a subgoal. The subsequent action sequence initiated by Mary the Monitor is as follows:

1. Mary tells the student enough time has passed for the current subgoal and asks the student if they feel they know enough to complete the subgoal.
2. Mary provides the student a 10 question quiz to assess knowledge.
3. If the student scores at least 60% on this quiz, then the subgoal is marked as completed.

MetaTutor supports SRL processes through various components. One such component is the SRL palette of actions that allow students to express metacognitive monitoring and control processes. An SRL palette on the righthand side of the MetaTutor interface offers a variety of actions like "Assess how well I understand this" (i.e., judgment of understanding). This action allows students to state their understanding of the current page and then indicate their understanding on a scale. Also, students can select "Summarize" and write a summary of a page's content in a free text box (i.e., one type of SRL strategy use).

Harley *et al.* (2018) explored the impact of prompts and feedback from PAs in MetaTutor. The interactions between the student and the PAs were investigated across two conditions: prompt and feedback versus a baseline. In the prompt and feedback condition, PAs prompted students to use self-regulatory processes and provided feedback about students' use of these processes while learning. As part of this condition, Pam the Planner helped students to set appropriate subgoals and provided feedback if the goals were too specific or broad. In the baseline condition, students did not receive prompts or feedback from PAs. Results found that students followed prompts and feedback from PAs to set appropriate subgoals the majority of the time. Further, subgoals set *collaboratively* between students and the PA led to higher learning gains compared to those set less collaboratively.

Azevedo *et al.* (2012) also investigated prompts and feedback from PAs in MetaTutor, but with a finer level of granularity. The study involved both of the conditions of the Harley *et al.* (2018) study (i.e., prompt and feedback versus baseline), but also added a third intermediate condition called prompt. In the prompt condition, the PA prompted but did not provide feedback. In the two conditions with prompts, timing of prompts were adaptive to the individual using factors such as learning interaction, time on page, time on subgoal, number of pages visited, relevance of page to subgoal, etc. Results found that participants in the prompt and feedback condition generated *fewer* subgoals than the other conditions. Participants in the prompt and feedback condition also spent more time on each subgoal than those in the other conditions. A learning efficiency score was calculated by dividing a participant's post-test score by the number of minutes the participant spent learning. Participants in the prompt and feedback condition had significantly higher learning efficiency scores than those in the baseline condition.

Key Takeaway

SAL researchers can apply findings from computer-based learning environment research that leverages SRL to design better learning-supportive search environments.

Outside of SAL, supporting effective SRL has had positive effects on learning outcomes. However, few SAL studies have aimed to directly support effective SRL during learning-oriented search tasks. Urgo (2023) incorporated a tool called the Subgoal Manager in order to support SRL planning, strategy use, and monitoring in relation to goal-setting while learning during search. Results found that learners had better learning outcomes when using the Subgoal Manager. Importantly, learners had better learning retention over time and higher engagement in SRL processes when using the Subgoal Manager while learning during search (compared with a typical search system and word document to take notes). As discussed further in Section 8, future SAL studies should aim to understand how users engage in SRL processes during search and develop tools to encourage and support effective SRL.

7.6 Summary

It is important that SAL researchers investigate and integrate SRL for three main reasons. First, outside of SAL, SRL has been shown to improve learning outcomes. Researchers in SAL should facilitate SRL processes during search in order to improve learning outcomes. Second, while SRL is beneficial, effective SRL is challenging and takes effort and practice. SAL systems need to support learners in cultivating effective SRL practices for lifelong learning. Third, SRL will help researchers to better unpack the learning process during search. Capturing SRL during search uncovers *when*, *where*, and *how* learning occurs. Such insights are invaluable to supporting learning. In this section, we have put a considerable emphasis on goals and goal-setting as it plays an important role in effective SRL and, in turn, improves learning outcomes. The goal-setting literature has several key findings that should be considered in SAL research in order to help individuals achieve their learning objectives. In particular, there are four ideal goal qualities that make goals more achievable. Goals should be *difficult*, *specific* (specifying time frame, action, standard, and content), *proximal* (short-term rather than long-term), and *learning-oriented* (emphasizing personal growth of skill, understanding, and knowledge). Future researchers should also consider exploring characteristics of goals and the goal-setting process in support of learning during search.

8

Future Research Directions

Previous research in the learning sciences has underscored several crucial factors for supporting learning that SAL research has not yet thoroughly examined. In this section, we discuss opportunities for future SAL research.

First, transfer of learning—the ability for a learner to use knowledge in a novel context—is arguably the goal of meaningful learning (Council and Education, 2000; Anderson *et al.*, 2000; Haskell, 2001). However, limited work in SAL has investigated transfer of learning. Therefore, we discuss the importance of measuring transfer of learning in future SAL work.

Second, learning sciences research has highlighted the importance of the learning context (e.g., the topical domain, the learner's prior knowledge, and characteristics of the individual learner) on learning processes and outcomes (Sawyer, 2014). Therefore, we elaborate upon these contextual factors and highlight how they might be considered in future SAL environments.

Third, methodologically speaking, longitudinal studies are important for capturing the learning process over an extended time period. The majority of SAL work has been conducted in the context of a single search session that spans a short time period (e.g., 40 minutes). Therefore, we highlight the importance of longitudinal studies in future SAL work.

Fourth, SAL research is important to help people become better lifelong learners. Self-determined learning is an educational paradigm aimed at helping students learn to learn on their own. To help students become better lifelong learners, advocates of this trending paradigm have highlighted several important characteristics that educational environments should consider. These include having students actively participate in setting their own goals, regulate their own learning, and reflect on their own learning outcomes. We discuss these ideal characteristics of learning environments and their implications for future SAL research.

Fifth, most SAL studies have investigated learning within the context of non-controversial, agreed-upon topics (e.g., learn about diffusion and osmosis). Future studies should also investigate learning within the context of controversial, highly debated topics. Learning about a heavily debated topic may pose unique challenges, such as the presence of misinformation and disinformation; biased search results; and the individual searcher's cognitive biases.

Sixth, decades of prior work in the learning sciences have shown that effective self-regulated learning (SRL) improves learning outcomes. Therefore, we discuss the importance of capturing and scaffolding SRL processes within learning-supportive search environments. We emphasize the importance of providing scaffolding in a way that is thoughtful, adaptive, and nuanced. For example, system interventions should only offload the work that is *not* useful for learning.

Seventh, generative AI tools have paved the way for a new era of learning-supportive search environments. We discuss how SAL research should leverage and build upon existing generative AI technologies to help individuals achieve complex, learning-oriented goals.

Finally, SAL research has mostly focused on individuals learning on their own. The field of computer-supported collaborative learning (CSCL) investigates how people learn together using computers. While CSCL studies have focused on group learning, they have not focused on scenarios where information seeking is a critical activity performed by the group. Therefore, we discuss opportunities for future SAL research to study groups of people learning by searching.

Before moving onto these crucial areas of exploration, it is important to first address an issue that underpins all aspects of future work in SAL systems. There is a fundamental tension in the design of SAL systems: that the historical emphasis in information retrieval (IR) on search *efficiency* may actively work against *meaningful human learning*. Traditional IR systems were often optimized to minimize user time and effort, with speed and ease of access serving as primary measures of success. However, this approach conflicts with our understanding of how meaningful learning occurs. Learning is inherently effortful, requiring sustained engagement, careful reflection, and deliberate processing and organizing of information.

This tension presents a crucial design challenge for the next generation of SAL systems. Rather than prioritizing quick answers to meet individual goals, these systems need to be reconceptualized to scaffold and support sustained cognitive and metacognitive engagement. The goal shifts from minimizing user effort to optimizing it—creating desirable difficulty within one's ZPD (see Section 1) that enhances learning outcomes. This represents a paradigm shift from traditional search metrics to a model where prolonged, thoughtful interaction with information becomes a new measure of system success. By rethinking traditional goals, SAL systems can evolve to support not just the acquisition of information, but support meaningful learning, creating an environment where prolonged, thoughtful engagement becomes a defining feature of successful learning experiences.

8.1 Transfer of Learning

As discussed in Section 3, SAL studies have used a variety of assessment types to measure learning during search. However, assessments have largely focused on measuring learning within the same context in which the new knowledge was acquired. For example, one previous SAL study (Urgo *et al.*, 2020) provided participants with the following learning-oriented search task: "Determine which best explains lift acting on an airplane's wing and why: Bernoulli's principle or Newton's laws of motion?" In this task, participants were directed to learn about Bernoulli's principle and Newton's laws of motion within the context of lift. This led many participants to look at diagrams of airfoils in the context of aircraft flight. After searching, participants were provided with an assessment that asked them to decide which concept was best suited to explain lift and to provide a logical argument to support their claim. This assessment measured how well participants understood the application and components of Bernoulli's principle and Newton's Laws of Motion within the context of lift. However, these concepts can be applied to many other contexts. For example, Bernoulli's principle explains the physics behind: how the spin of a baseball makes the ball curve, how the curve on a sail makes a sailboat move forward, and how the narrow tube attached to a paint sprayer turns the paint to a fine mist. A very important part of meaningful learning involves using knowledge in a novel context. This type of learning is called *transfer of learning.*

Haskell defines transfer of learning as the "use of past learning when learning something new or the application of [past] learning to [...] new situations" (Haskell, 2001, p. xiii). Assessments that target transfer of learning measure a learner's ability to use knowledge in a new context from the one encountered during the learning process. Essentially, assessments that target transfer of learning measure a learner's ability to *generalize* beyond what was learned. Anderson *et al.* (2000) argued that being able to transfer knowledge to new situations or problems is a core tenet of meaningful learning.

Key Takeaway

Transfer of learning involves a learner generalizing their learning to a new context.

Although transfer of learning is central to meaningful learning, few SAL studies have explored transfer of learning. Heilman and Eskenazi

(2006) and Heilman *et al.* (2010) explored transfer of learning in the context of vocabulary acquisition during search. To measure learning, participants completed fill-in-the-blank sentences using target vocabulary words. These fill-in-the-blank sentences situated vocabulary terms within the same textual context encountered during the search session. Additionally, to measure transfer of learning, participants were also asked to generate their own sentences using target vocabulary words. These sentence generation questions required participants to situate a target vocabulary word in a novel context. Generated sentences were assessed based on correct grammar and the extent to which they signaled a complete and nuanced understanding of the vocabulary word.

SAL researchers may be hesitant to explore transfer of learning as the concept itself can be ambiguous and difficult to define. Although transfer of learning has been studied for more than a century, there is still no clear, unified model of transfer (Tuttle, 1955; Haskell, 2001; Barnett and Ceci, 2002). Because of the ambiguity of transfer of learning, it is challenging to develop assessments that measure transfer. To support researchers in implementing assessments to measure transfer in SAL, we present two frameworks. Both frameworks can be used in the development of transfer of learning assessments.

Using the A&K to Measure Complexity of Transfer: Urgo and Arguello (2022b) proposed the Anderson & Krathwohl (A&K) taxonomy (see Section 2) as a framework to develop assessment materials to measure transfer of learning. They show how transfer of learning assessment questions might be developed for the cognitive processes of understand, apply, analyze, evaluate, and create.

To demonstrate how to use the taxonomy for transfer, let us consider the learning objective provided at the beginning of this section from Urgo *et al.* (2020). The objective asked participants to decide whether Bernoulli's principle or Newton's Laws of Motion best explains how lift works. To measure transfer of learning across various cognitive processes, one could use the A&K taxonomy:

- To measure transfer along the understand cognitive process, a question could ask a participant to determine if Bernoulli's principle is exemplified in a new scenario: curveballs in baseball.

- To measure transfer along the apply cognitive process, a transfer of learning question might ask participants to use Bernoulli's principle to explain how a different phenomenon, a curveball, works.

- To measure transfer along the analyze cognitive process, a transfer of learning question might ask participants to compare the similarities and differences between Bernoulli's principle and a different related concept other than Newton's Laws of Motion, such as the Venturi Effect.[1]

- To measure transfer along the evaluate cognitive process, a transfer of learning question might ask participants to decide whether Bernoulli's principle or Newton's Laws of Motion best explain a different phenomenon, such as a curveball.

- To measure transfer along the create cognitive process, a transfer of learning question might ask participants to create a diagram of Bernoulli's principle applied to a different phenomenon, such as a curveball.

Designing transfer of learning questions with the A&K taxonomy affords researchers with the additional benefit of assessing complexity of transfer. For instance, perhaps a participant is able to answer the understand and apply transfer questions, but stumbles when encountering the questions of higher complexity (i.e., analyze, evaluate, and create). Understanding the complexity of transfer achieved may be useful in understanding how a system might better support particular kinds of learning during search.

Using Barnett and Ceci (2002) to Measure Dimensions of Near to Far Transfer: Transfer of learning has been researched for more than a century (Woodworth and Thorndike, 1901) and can be ambiguous in nature. When we talk about transfer, we discuss the application of learned information in a novel context. However, this leads to many subsequent questions. What makes a context novel?

[1]To this end, the assessment might need to provide some background information about the new related concept.

How close is the new context to the original context? Does the learner's knowledge evolve when it is applied to the novel context? This ambiguity has resulted in a myriad of interpretations into the idea of transfer of learning.

To organize these interpretations, Barnett and Ceci (2002) developed a two-dimensional framework to describe transfer of learning. The *context dimension* relates to similarities between the context in which the knowledge or skill was originally acquired and the novel context in which the knowledge or skill is being applied. The context dimension involves six sub-dimensions: (1) knowledge domain; (2) physical context; (3) temporal context; (4) functional context; (5) social context; and (6) modality.

The *content dimension* is more nuanced. The content dimension involves three different aspects that consider what is transferred, how well, and whether the learner is told which knowledge should be applied in the new scenario. The first aspect considers the *specificity* of the knowledge or skill being transferred. To illustrate, in a highly specific scenario, a learner might be expected to execute a step-by-step procedure in a novel context. In a more general scenario, a learner might be expected to apply a general principle or technique associated with the learned procedure. The second aspect considers *how well* transferred happened.[2] To illustrate, one might consider how quickly or how accurately a learner is able to execute a step-by-step procedure in a novel context. Finally, the third aspect considers the *complexity* of the transfer. In a simple scenario, the learner is told which prior knowledge they are expected to apply in the new context. In a scenario with a greater range of complexity, the learner is also expected to recognize which prior knowledge is relevant to the new context. In other words, in a simple transfer scenario, the learner is expected to recall and apply. Conversely, in a more complex transfer scenario, they are expected to recognize, recall, and apply. Next, we elaborate on the context dimension of transfer of learning, which focuses on differences between the context in which the learning took place and the context in which the knowledge or skill is being applied.

[2]Barnett and Ceci (2002) discuss this aspect as part of the content dimension. However, we view it as having more to do with how knowledge transfer might be scored.

In the definitions below, *learning context* refers to the context in which the new knowledge or skill was learned, and *transfer context* refers to the context in which the new knowledge or skill is being applied (or should be applied if transfer is successful). The context dimension of transfer has several sub-dimensions. Each sub-dimension describes a way in which the learning and transfer context can vary. When discussing similarities and differences between the learning and transfer contexts, researchers often use a distance metaphor. "Near" transfer involves transfer contexts with similar properties as the learning context and "far" transfer involves transfer contexts with different properties as the learning context. The sub-dimensions below should be viewed as orthogonal. That is, the learning and transfer contexts can be "near" along one dimension and "far" along another dimension. When we think about transfer of learning, we usually think about the knowledge domain (e.g., learning about Bernoulli's principle in the context of lift and then using it to explain curveballs in baseball). However, as it turns out, transfer of learning scenarios can vary along five other dimensions.

1. **Knowledge domain**: relates to the topic, subject area, or academic field associated with the learning context and the transfer context. While subject area and knowledge domain are quite intuitive to imagine, the topic can also vary. One might learn about a topic by seeing how it explains one phenomenon (e.g., Bernoulli's principle to explain lift) and later be required to use it to explain a different phenomenon (e.g., Bernoulli's principle to explain curveballs).

    - *Learning context*: A student learns the quadratic formula using step-by-step algebra tutorials online. Here, the student is learning the quadratic formula within the field of math.
    - *Near transfer context*: The student then uses the quadratic formula to solve problems in their physics class, like the trajectory of a projectile. Physics is quite close to the learning context knowledge domain of mathematics.
    - *Far transfer context*: The student then uses the quadratic formula in their biology class to model population growth

under certain conditions. Biology is further from the learning context knowledge domain of mathematics.

- *Farthest transfer context*: The student then uses the quadratic formula in their economics class to calculate the maximum profit on a homework assignment. Economics is the farthest academic field from the learning context knowledge domain of mathematics.

2. **Physical context**: relates to the physical setting or environment associated with the learning context and the transfer context.

    - *Learning context*: The student learns about how to conduct a titration experiment to determine the concentration of an acid solution by watching online lab demonstrations and tutorials. Here, the physical context in the videos is a high school chemistry lab.
    - *Near transfer context*: The student then successfully conducts the same titration experiment in the school lab. This is close to the physical learning context as the skill is also applied in a high school chemistry lab.
    - *Far transfer context*: The student then adapts the titration experiment to use household items (e.g., vinegar and baking soda) and successfully conducts the experiment at home. This is further from the physical learning context as the home environment is different from a high school lab.
    - *Farthest transfer context*: The student then uses titration methods to conduct a field experiment to determine the concentration of acidic pollutants in rainwater samples. The field environment is farthest from the physical learning context of a high school lab.

3. **Temporal context**: relates to the temporal difference between the learning context and the transfer context.

    - *Learning context*: A student learns how to create a presentation using Prezi by following step-by-step tutorials online. This is the temporal learning context.

- *Near transfer context*: The student immediately uses their understanding to create a presentation for a class project. This is close in time to the temporal learning context.
- *Far transfer context*: The student uses Prezi to create a presentation later in the semester. This is further in time from the temporal learning context.
- *Farthest transfer context*: The student uses Prezi years later to create a presentation. This is the farthest in time from the temporal learning context.

4. **Functional context**: relates to the goals associated with the learning context and the transfer context. One might learn about a topic for the purpose of passing a test (e.g., final exam in an academic course) and later be required to use the topic to solve a real-life problem (e.g., in a job interview). The objectives during the learning and transfer contexts are different.

    - *Learning context*: An engineer learns to use Fusion 360 to design a particular circuit by searching online forums and tutorials. Here, the functional context is the engineer learning to use the tool for a specific design task.
    - *Near transfer context*: The engineer successfully uses Fusion 360 to design a similar circuit to that from the online forums and tutorials for a research project. This is a similar circuit design task close to the functional learning context.
    - *Far transfer context*: The engineer uses Fusion 360 to design a circuit for a hobby project at home, building a custom amplifier. This is a different type of design task further from functional learning context.
    - *Farthest transfer context*: The engineer uses Fusion 360 in a real-world application, creating a circuit for a community center's new lighting system. This is the farthest type of design task from the functional learning context.

5. **Social context**: relates to the social situation associated with the learning context and transfer context.

- *Learning context*: A student learns Spanish using online courses and textbooks. Here, the student is learning Spanish in an individual social context.
- *Near transfer context*: The student reads articles, watches movies, and writes essays in Spanish. This context is slightly more social but close to the social learning context.
- *Far transfer context*: The student joins an online Spanish community and practices conversation in pairs and small groups. This is a more interactive social context further from the social learning context.
- *Farthest transfer context*: The student travels to Madrid and speaks with a variety of people across the city. This is the farthest social context from the social learning context (i.e., individual study session).

6. **Modality**: relates to the medium associated with the learning context and transfer context.

   - *Learning context*: A student learns about photosynthesis by reading online articles and textbooks. Here, the student is learning through reading text-based materials.
   - *Near transfer context*: The student explains photosynthesis in an open-ended exam using what they learned. This is a similar modality close to the text-based learning context.
   - *Far transfer context*: The student gives an oral presentation on photosynthesis to their classmates using what they learned. This is a different modality of learning further from the text-based learning context.
   - *Farthest transfer context*: The student creates a diagram or animation to illustrate photosynthesis using what they learned. This is the modality farthest from the text-based learning context.

Above, we illustrate how the learning context and the transfer context can vary along different dimensions. Additionally, along each dimension, both contexts can vary to different extents. These sub-dimensions illustrate both the complexity of transfer of learning and its vast opportunity. Using these sub-dimensions as a guide, SAL researchers can focus experimental investigations around specific dimensions of transfer. Clearly, it is too broad to simply state that a search system does or does not support "transfer" in general. It is important that a SAL study be designed with a specific dimension of transfer in mind. For example, imagine a tool that allows searchers to follow the search paths from other searchers who learned about a similar topic. Such a tool might support transfer across social contexts. As another example, imagine a tool that prompts searchers to answer quiz questions about a topic over several weeks. Such a tool might support transfer across temporal contexts. In either case, it is important that researchers develop assessments that are associated with the specific *dimension* of transfer that is being targeted. Additionally, it is important for SAL researchers to account for all six sub-dimensions even if they are not directly manipulated. For example, a SAL researcher may investigate if a novel search system facilitates knowledge domain transfer. In response, the SAL researcher develops an assessment that measures a learner's ability to use the topic across a variety of academic fields. However, the SAL researcher should also control for temporal context, social context, physical context, functional context, and modality in terms of transfer of learning.

Using these provided frameworks as a springboard, SAL researchers should develop instruments to measure transfer of learning. Further, future SAL studies should investigate factors (e.g., factors of the search system, individual, or learning task) that result in more effective transfer of learning.

Prior work in transfer of learning has also found that domain expertise has a significant effect on one's ability to transfer. For example, Day and Goldstone (2012) analyzed how physics students group a set of problems. PhD students grouped the problems by the general principle involved in the solution (e.g., Newton's laws of motion or conversation of energy). On the other hand, more novice students grouped the same problems by the concrete characteristics of mechanisms involved (e.g., the presence of pulleys or springs). Put simply, the greater one's depth of understanding of a topic, the greater one's ability to transfer that understanding across contexts. Therefore, assessments that consider transfer also measure how deeply someone understands something.

Key Takeaway

Transfer of learning is multi-dimensional. Particular dimensions of transfer should be mindfully targeted in SAL studies to support meaningful learning.

8.2 Context-Aware Systems

Research findings within SAL and beyond suggest that different contextual factors are likely to impact learning during search. Contextual factors can be grouped into at least five categories: (1) the topical domain; (2) the learning objective; (3) the searcher's prior knowledge of the topic; (4) characteristics of the individual searcher; and (5) the motivating task that prompted the searcher to gather information for the purpose of learning. Future SAL research should consider how these factors might influence the types of interventions and tools needed to encourage and support learning during search. Understanding how contextual factors affect the learning process may inform the design of context-aware search environments that improve learning.

Outside of SAL, research has found that the learning objective's topical domain can influence the learning process. Greene *et al.* (2015) conducted a study in which participants were asked to learn about topics in the domains of history or science using different digital libraries. The history task asked participants to learn about the construction of the Blue Ridge Parkway in the U.S. state of North Carolina and the science task asked participants to learn about the phase changes of matter. The study investigated the importance of different self-regulated learning (SRL) processes on learning outcomes across domains. Some

SRL processes (e.g., prior knowledge activation, corroborating sources, knowledge elaboration) were predictive of learning in both domains. However, the predictiveness of other SRL processes were domain-specific. For example, judging the relevance of content was positively predictive of learning for the science domain but not the history domain. This result suggests that the domain of the learning objective may impact the types of SRL processes that systems need to encourage and scaffold.

Learning objectives vary not only by topic (e.g., history vs. science) but also by knowledge type (e.g., factual vs. conceptual vs. procedural). As discussed in Section 2, SAL studies have found that the knowledge type of the objective can affect the cognitive activities involved in the learning process (Urgo *et al.*, 2020; Urgo and Arguello, 2022c). For example, conceptual objectives may involve more understanding and analyzing, and procedural objectives may involve more evaluating and creating (Urgo *et al.*, 2020). This trend suggests that learners may need different types of support depending on the knowledge type of the objective. For example, during conceptual objectives, searchers may benefit from seeing the relationships between the target concept and other related concepts. Conversely, during procedural objectives, searchers may benefit from seeing the pros and cons of a procedure and ways in which it could be modified based on personal preferences or situational constraints.

Key Takeaway

Studies have found that a learning objective's topical domain and knowledge type can influence the SRL and cognitive processes which are most useful for learning.

Furthermore, there are textual features such as vocabulary density, readability, and reading level that are important to explore in developing better context-aware SAL systems. Syed and Collins-Thompson (2018) found that keyword density significantly impacts long-term retention of challenging vocabulary. Context-aware SAL systems could leverage generative AI to dynamically adjust the density of key terms in responses, optimizing for more effective vocabulary acquisition.

Readability and reading level may also play a critical role in how users engage with and understand retrieved information. Readability includes textual factors such as sentence length, word difficulty, and syllables per word. Lee *et al.* (2025) investigated the readability of generative AI chatbot responses to frequently asked medical questions related to pregnancy. They found that Bard's answers, which were at a high school reading level, were more accessible compared to ChatGPT's, which were at a college level. Such findings underscore the potential importance of readability in searcher understanding. Moreover, Roegiest and Pinkosova (2024) highlighted the challenges posed by generative AI systems that predominantly produce responses at collegiate reading levels, which may exclude users with lower literacy. Future research could explore how SAL systems might increase inclusivity by dynamically adapting readability levels based on users' individual literacy and learning objectives to support deeper understanding.

Key Takeaway

Prior knowledge has been shown to affect search behaviors and focus during learning tasks.

As discussed in Section 4, SAL studies have found that prior knowledge can affect learning during search in different ways (Willoughby *et al.*, 2009; O'Brien *et al.*, 2020). On one hand, searchers with greater prior knowledge may be able to search more effectively. On the other hand, searchers with greater prior knowledge may have more difficulty finding novel information that expands their existing knowledge. Outside of SAL, Bernacki (2010) found that individuals with higher prior knowledge were more likely to stay focused and not be distracted by information that was not germane to the learning task. Future SAL studies should investigate the unique needs of searchers with different levels of prior knowledge.

Cognitive biases are systematic deviations from aspects of objective reality and can take different forms, such as projection bias, anchoring bias, and confirmation bias (Soprano *et al.*, 2024). Cognitive biases can impact how people search for and process information. To illustrate, projection bias may arise when individuals pose queries that assume positive or negative outcomes (e.g., "Does aloe vera cure cancer?"). Anchoring bias may occur when searchers assign disproportionate weight to results encountered early in the session regardless of their accuracy. Confirmation bias may involve favoring results supporting an individual's assumptions. Indeed, studies have found that searchers spend significantly more time viewing results that reinforce versus contradict their believes (Azzopardi, 2021).

With respect to cognitive biases, there are several questions for future SAL research to consider. How do specific cognitive biases influence the learning process during search? What evidence can we use to automatically detect when a specific cognitive bias is involved? What are contextual factors that may influence a searchers' susceptibility to a specific type of cognitive bias? What system features and interventions can we design to mitigate the negative effects of specific cognitive biases?

Beyond the factors of prior knowledge and cognitive bias, studies have found better learning outcomes for individuals with higher levels of working memory capacity (Pardi *et al.*, 2020; Choi *et al.*, 2019a), perceptual speed (Azzopardi *et al.*, 2023), and attention control (Azzopardi *et al.*, 2023). Outside of SAL, researchers have called for adaptive computer-based learning environments that provide tailored instruction based on the learner's cognitive abilities and learning styles (Vandewaetere *et al.*, 2011). Future SAL studies should investigate the unique needs of searchers with different types of abilities. As discussed in Section 4, searchers vary not only by their abilities but also by their personalities. Future work should also consider personality traits such as need for cognition (Wu *et al.*, 2014).

Finally, early calls for SAL research argued that we must better understand the contexts in which people search for the purpose of learning (Allan *et al.*, 2012). In their perspective paper, Smith *et al.* (2022) envisioned a search environment that can support learning by students within the context of an academic assignment. The authors envisioned a system with an *assignment model* that can process an assignment as input in order to: (1) identify target knowledge (e.g., relevant concepts to be learned); (2) identify relevant modalities (e.g., videos, images, text); and (3) update the user's knowledge state once the assignment is completed. Future research should consider search environments that allow users to provide information about the higher-level objective being pursued (e.g., the school assignment). Such a system could provide search and learning experiences that are tailored to the higher-level objective.

Key Takeaway

Search environments should consider characteristics of the task, the learner, and the retrieved information relative to the learner's knowledge state.

To summarize, context-aware SAL systems should consider characteristics of the learning task, the searcher's prior knowledge, characteristics of the individual, and characteristics of the higher-level objective being pursued. Automatically predicting these contextual factors may be overly ambitious. As a starting point, systems could enable searchers to explicitly input contextual information using the interface.

8.3 Longitudinal Studies

Most SAL studies to date have focused on learning during one search session. Longitudinal studies take place over an extended period of time and measurements are taken at different intervals (Kelly, 2009).

In real life, people often embark on learning journeys that extend over days, weeks, or even months. Researchers in the learning sciences have underscored the importance of longitudinal studies as a means to investigate: (1) the learning process over an extended time period and (2) the long-term effects of learning interventions (Sawyer, 2014). Longitudinal studies are often naturalistic. That is, they involve participants working on genuine (versus assigned) learning objectives. For example, they might involve participants learning about a topic for school, work, or based on their personal hobbies. In this respect, Harju *et al.* (2019) argue that longitudinal studies are more likely to consider the learning context, which is likely to impact the learning process. In the context of SAL, longitudinal studies may help us understand how learning unfolds over time across multiple search sessions.

Only a few SAL studies have investigated learning over an extended period. Bhattacharya (2023) investigated learning by a group of students enrolled in a university course over a semester. Results found that participants who had higher levels of engagement in metacognitive and self-regulatory processes (based on self-report data) searched more efficiently. As evidence, their final grades were similar to participants who had longer search sessions and visited more pages per session. Cole (2022) investigated learning by a set of user experience professionals over a five-day period. Using a learning diary, participants described their experiences while aiming to learn about topics relevant to their work. An important goal of the study was to investigate participants' perceptions of self-efficacy before and after the five-day period. Self-efficacy is defined as someone's confidence in their ability to achieve their goals. The study considered different dimensions of self-efficacy, including motivational, affective, and cognitive dimensions. Surprisingly, results found a decline in one dimension of self-efficacy, *schema training*, which is defined as someone's self-described ability to employ effective search strategies depending on their goals. Based on their feedback, participants worked on tasks that required finding highly detailed information. This prompted them to become more aware of gaps in their understanding of how search algorithms work. This result suggests that current search systems fall short in supporting long-term learning.

In the future, SAL researchers should consider conducting more longitudinal studies aimed at understanding long-term learning processes that involve information seeking. Open questions may include: What types of SRL processes do learners engage in at different phases of the learning process? How might a search environment support learners during different phases? What are the long-term influences of different learning interventions? Past SAL studies have mostly considered how people learn about a specific topic. In the future, longitudinal studies could investigate how people leverage search engines to learn a new skill (e.g., learn to program in Python), which is likely to require learning over an extended time period.

Key Takeaway

Longitudinal studies are rare in SAL research. However, they could provide valuable insights into how learning unfolds over time and the long-term impacts of system interventions.

Outside of SAL, IR research has investigated the reasons for why people conduct multi-session searches and the challenges they face. As an example, Li and Capra (2022) conducted a diary study in which 15 participants embarked on tasks that they expected to require multiple search sessions. Interestingly, 2 out of 15 participants embarked on tasks that involved learning as the primary objective. Participants conducted multi-session searches for reasons related to the task's structure (e.g., the task can be sub-divided into several sub-tasks) and the task's complexity (e.g., the task will require digesting the information found before moving forward). Additionally, participants commented on experiencing different challenges while resuming their tasks, such as losing track of past searches/information, losing momentum, and staying motivated. Future SAL studies could leverage past research on cross-session search to brainstorm ways in which searchers may need support during long-term learning tasks.

8.4 Self-Determined Learning

Within the field of education, researchers and practitioners continuously debate how students should learn and be taught. One of the most recent learning paradigms is called *self-determined learning* or *heutagogy*. Proponents of self-determined learning argue that, to be effective lifelong learners, students must learn to establish their own learning agendas and drive their own learning (Blaschke, 2012; Blaschke and Hase, 2016; Blaschke, 2018). In other words, to thrive in today's and tomorrow's workforce, educators must teach students how to teach themselves (Blaschke, 2012).

As argued by Blaschke (2012), to produce effective self-determined learners, the learning environment must pay attention to five important themes. Perhaps the most important theme is learner agency—the learner should be an active participant in: (1) defining their own learning objectives, (2) deciding how they will learn, and (3) assessing their own learning. The role of the educator is to guide students through these different processes. The second theme is the importance of developing both competence and capability. Competence is the ability to show that knowledge has been acquired. Capability is the ability to show that knowledge can be used to solve problems in new and unfamiliar situations (i.e., transfer of learning). The third theme is the importance of self-efficacy—a person's belief in their ability to solve problems, complete tasks, and learn new things. Proponents of self-determined learning argue that teaching students to transfer knowledge is one way to improve their confidence. The fourth theme is the importance of metacognition and reflection. Students should be taught to be aware of and regulate their own learning processes, as well as reflect on learning outcomes to better understand their unique learning styles and develop into better lifelong learners. The final important theme is authenticity. Authenticity argues that learners should learn by doing rather than simply seeing information. Additionally, learners should learn by performing tasks that are perceived as relevant to them.

To summarize, in education, self-determined learning is a relatively new paradigm that originated from a desire to help students become more effective lifelong learners. It argues that students should be active participants in all aspects of their learning; should be taught to transfer knowledge to new and unfamiliar situations; should be supported to grow more confident; should be taught to regulate their own learning processes and reflect on their learning styles; and should learn by doing things that are perceived as relevant to their lives.

Key Takeaway

Self-determined learning is a relatively new theory of learning that stands out for its emphasis on learner agency and authenticity in learning experiences.

Self-determined learning offers a unique perspective for future SAL studies to consider. Next, we describe how the five themes behind self-directed learning could influence future SAL research.

First and foremost, the goal of self-determined learning is to teach students to teach themselves. In Section 6, we surveyed prior SAL studies that have investigated how different tools can help people learn about a topic during search. Future studies should also consider tools to help people learn to learn. For example, future longitudinal studies (Section 8.3) should investigate how tools within a search environment can help improve people's ability to learn in the long term.

Second, self-determined learning argues that people should be actively involved in setting their own learning objectives, planning their own instructional activities, and assessing their own learning. In terms of learning objectives, prior studies have considered tools to help people set goals before searching to learn (Urgo and Arguello, 2023; Urgo and Arguello, 2024). Future work could investigate tools to help with this process (e.g., propose new goals, suggest ways to improve a goal, or reorder the goals to have a more logical flow). In terms of assessment, future work could investigate tools to help learners assess their own learning based on the information encountered and/or notes taken within the search environment.

Third, future systems could propose ways for users to test their ability to transfer new knowledge based on the information encountered and/or their notes. For example, suppose someone just learned how Bernoulli's principle enables a plane to fly. A system could prompt this user to explain how Bernoulli's principle causes a tennis ball to dip when it has topspin.

Fourth, future SAL systems should consider the effects of tools on perceptions of self-efficacy. Cole (2022) developed a questionnaire to measure perceptions of self-efficacy during a SAL study. The questionnaire considers different dimensions of self-efficacy, including motivation, emotion regulation, (meta)cognitive skills (i.e., planning, monitoring, evaluating), and search skills. Proponents of self-directed learning have argued that positive perceptions of self-efficacy are incredibly important for lifelong learning. SAL studies should continue measure learning from an external and objective perspective (i.e., by using pre- and post-task assessments). However, internal perceptions of learning, as well as changes to someone's confidence in their ability to learn, are equally important and should not be neglected.

Fifth, future SAL studies should consider tools to help searchers engage in effective SRL processes and to reflect on their learning outcomes. In Section 8.6, we describe possible tools to support SRL processes. To our knowledge, SAL studies have not considered how tools might help people reflect on their learning outcomes after a search session. Tools to encourage and support reflection might help people learn about their learning styles (i.e., what works for them and what does not). A longitudinal study could consider how tools that encourage and support reflection after a learning-oriented search session might improve learning in the long term.

Finally, the notion of authenticity argues that people should learn by *doing* rather than just *seeing*. Additionally, learners should perform educational tasks that are perceived as relevant to them. The notion of authenticity has two important implications for future SAL studies. First, it suggests that more SAL studies should involve people working on their own, genuine learning tasks (Cole, 2022). Studies that use genuine tasks might see larger effects than studies that use assigned tasks that may not be perceived as relevant to participants' lives. Second, search environments should help people learn by doing versus just seeing. In fact, this might help explain why note-taking tools have consistently improved learning outcomes during search (Freund *et al.*, 2016; Roy *et al.*, 2021; Qiu *et al.*, 2020). With the advent of generative AI technologies, future systems could automatically generate exercises for searchers to complete based on the content read and/or notes written within the search environment.

8.5 Learning within Highly Debated Topics

With few exceptions (e.g., Demaree *et al.*, 2020), most SAL studies have investigated learning within the context of non-controversial, agreed-upon topics. Examples include: (1) learn about the biological concepts of diffusion and osmosis (Urgo and Arguello, 2023); (2) learn about the formation of thunderstorms and lightning (Pardi *et al.*, 2020); and (3) learn about the history of gold rushes in British Columbia, Canada (O'Brien *et al.*, 2020). Future studies should also investigate learning within the context of controversial, heavily debated topics. Learning about a heavily debated topic may pose unique challenges, such as: (1) the presence of misinformation and disinformation; (2) biases in the document collection, ranking algorithm, or snippet-generation algorithm; and (3) learning about viewpoints that may be different from one's own (i.e., learning while overcoming one's own biases). Researchers have acknowledged that people often use search engines to learn about controversial topics and inform their viewpoints (Rieger *et al.*, 2024). Within this area, there are several research questions that may be worth exploring.

First, how do we measure learning within a controversial, heavily debated topic? In Section 3, we surveyed different types of learning assessments used in prior work. Closed-ended assessments such as multiple-choice or short-answer tests, which have objectively correct answers, may not be appropriate. Open-ended assessments may be more suitable. As mentioned in Section 3, one prior SAL study asked participants to learn about the extent to which nuclear power may help solve the climate crises (Demaree *et al.*, 2020). Learning was measured based on the number of correct pro and con statements included in argumentative essays written by participants after the search task. Examples include "nuclear power plants do not emit carbon dioxide" (pro) and "mining uranium produces a lot of carbon dioxide" (con). However, within some heavily debated topics, it may be difficult to determine whether pro or con arguments are factually and verifiably true or false. An alternative assessment could consider whether the searcher developed a more comprehensive or more nuanced understanding of the debated topic.

Second, future studies could consider the interplay between learning and viewpoint. Here, there are interesting questions to explore in both directions. First, how does viewpoint impact learning? For example, if asked to learn about both sides of a controversial topic, are searchers more likely to learn about the side they already favor? If so, why? Second, how does learning impact changes in viewpoint? For example, if asked to learn about both sides of a controversial topic, are searchers more likely to be sympathetic to people on both sides of the debate? Additionally, in both directions, studies could examine factors that impact the interplay between learning and viewpoint.

Key Takeaway

Future SAL studies should investigate learning within highly debated topical domains, which may pose a unique set of challenges.

Finally, future studies could explore different system features and tools to support learning within highly debated topics. Here, prior work has developed different resources and methods that may be useful. Draws *et al.* (2022) developed a topic-*independent* taxonomy for classifying argumentative text along two dimensions: stance and logic of evaluation. Stance relates to the strength of the argument, and logic of evaluation relates to *why* the stance is taken. To illustrate, some logic of evaluation categories include: (1) popular (i.e., "it is what people want"); (2) moral (i.e., "it is morally right"); (3) civic (i.e., "it is legal"); and (4) functional (i.e., "it works"). Additional, prior work has developed algorithmic approaches to tasks such as: (1) classifying text into viewpoint categories (ALDayel and Magdy, 2021); (2) detecting fake news (Pérez-Rosas *et al.*, 2018); (3) measuring viewpoint-related bias in search results (Draws *et al.*, 2021); and (4) diversifying search results by viewpoint (Tintarev *et al.*, 2018).

8.6 Capturing and Scaffolding SRL

To better understand the learning process and support learning during search, future SAL work should focus on capturing and scaffolding SRL. As described in Section 7, prior work in the learning sciences has found that effective SRL improves learning outcomes (Zimmerman and Pons, 1986; Zimmerman and Martinez-Pons, 1988; Schunk, 1984; Schunk, 1981; Deekens *et al.*, 2018; Greene *et al.*, 2020). Further, computer-based learning systems that encourage and scaffold effective SRL have been found to improve learning (Zheng, 2016). Specifically, computer-based learning systems that support SRL processes such as goal setting, strategizing, and monitoring have had the largest effects. Despite these encouraging results, few SAL studies have investigated SRL.

Crescenzi *et al.* (2021) investigated the effects of an auxiliary search tool called the OrgBox on participants' post-task *perceptions* of SRL engagement. The OrgBox tool allowed participants to save and organize information about a topic as they searched. Participants who used the OrgBox reported greater levels planning and monitoring. The researchers in this study did not *directly* capture participants' engagement in SRL processes and did not measure learning outcomes. Therefore, more work is needed to understand the influence of tools such as the OrgBox on SRL engagement and learning.

Urgo and Arguello (2024) provided searchers with an auxiliary search tool called the Subgoal Manager that supported various SRL processes, including *planning* subgoals, *monitoring progress* toward subgoals, *taking notes* associated with subgoals, and *summarizing* subgoal knowledge. Urgo and Arguello (2024) captured think-aloud comments and behaviors that were coded as specific SRL processes. Results found that participants who used the Subgoal Manager had higher levels of engagement in SRL processes that were *directly* supported by the Subgoal Manager, such as *planning* subgoals and *monitoring progress* toward subgoals. Additionally, participants who used the Subgoal Manager had higher levels of engagement in SRL processes that were *not directly* supported by the Subgoal Manager, such as *prior knowledge activation*. Finally, based on learning assessments completed one week after the search session, participants who used the Subgoal Manager had higher levels of retention. Given this result, future work should investigate other tools that might support effective SRL during search.

Key Takeaway

SRL has been shown to be critical to enhancing learning outcomes and, therefore, should be encouraged and facilitated by learning-supportive search systems.

More work is necessary to know specific ways in which a search environment could support specific SRL processes. Next, we provide several avenues to explore based on prior work. In some cases, we reference specific phases and processes associated with the Winne and Hadwin (W&H) model of SRL (Figure 7.1).

First, a search system could scaffold the first phase of the W&H model of SRL, *task understanding*. A search system could provide a simple table that prompts the learner to brainstorm, reflect, and write about the task and its internal and external constraints. Internal constraints could include cognitive conditions such as prior knowledge within the task domain and knowledge of learning strategies that might be useful during the task. External constraints could include task conditions such as available resources, temporal constraints, and instructor expectations.

This type of simple tool could help the learner reflect on what the task is asking, what they already know, and what resources they have available. Additionally, this reflection might help the learner have a more effective *planning* phase.

Second, a search system could scaffold SRL *planning*. Urgo and Arguello (2024) used the Subgoal Manager to assist in *planning* by prompting learners to set goals before searching. However, systems could further support planning by offering feedback to learners about the *quality* of their goals. As noted in Section 7, goals that include a specific action, content, standard, and timeframe are more likely to be achieved. The Subgoal Manager offered tooltips that reminded participants about these ideal goal characteristics. However, future tools could go one step further by offering dynamic feedback about ways in which a subgoal could be improved (e.g., "This subgoal lacks a specific standard to help you determine when the subgoal has been completed.")

Third, a search system could scaffold SRL *monitoring*. In particular, SAL systems could provide support for *judgment of understanding*. Judgments of understanding involve the learner asking themselves how well they currently understand target information. To support this process, a search system could intermittently nudge the learner by asking: "How well do you understand X?"

Finally, all of the preceding areas of exploration are somewhat dependent on knowing *when* support for different SRL processes is most effective. Although the W&H model of SRL is weakly sequential (i.e., task understanding, then planning, then strategizing, then adapting), SRL is highly iterative and subjective to the individual learner. Little is known about the optimal moment to prompt different types of SRL processes. Future work should gather more data on actual SRL processes to unpack when they occur.

Scaffolding of SRL should be done thoughtfully to ensure that only the work that is *not* beneficial to learning is offloaded. Systems should not automate processes that improve learning and retention. For example, prior work in psychology has demonstrated that the act of organizing and categorizing information is beneficial for learning and retention (Tulving, 1962; Berry, 2012). Although it may be *possible* for a SAL system to automatically organize and categorize relevant information, this might be detrimental to a learner's understanding and future retrieval of the information. SAL researchers should be cautious in implementing scaffolding, being mindful to "...offload work not productive for learning, provide strategic guidance, make the structure of the domain more transparent, and support articulation and reflection" (Sawyer, 2014, p. 51-52).

Key Takeaway

SAL researchers should be careful to support and encourage SRL rather than offloading important learning processes that are effortful and helpful to meaningful learning.

8.7 Generative AI Tools to Support SAL

In recent years, a massive wave of generative AI and large language model (LLM) powered tools has crashed into the realms of education, pedagogy, and student learning. This confluence is not just reshaping how information is delivered, but also how learners interact with and assimilate knowledge. The transformative potential of generative AI tools in educational contexts necessitates a thorough examination of their impact on learning outcomes and the strategies and methodologies that underpin SAL research. This section delves into five important dimensions of this intersection.

First, we explore the potential impact of generative AI tools on learning outcomes, emphasizing the urgency for targeted research within SAL. It is critical to understand how these tools can enhance or hinder learning to develop learning-supportive environments.

Second, we discuss the role of LLMs in helping users navigate the relationships among a collection of documents. This capability could significantly support a learner's ability to connect and synthesize disparate pieces of information, fostering a deeper understanding of the topic being learned.

Third, we discuss how generative AI tools could be used to promote self-regulated learning processes by adapting to a learner's current context and level of understanding.

Fourth, we examine how generative AI can support the generation of questions, facilitating learning-supportive search systems that are context-relevant and capable of assessing learning. The ability of AI to craft tailored questions could create personalized learning experiences that deepen learner understanding.

Finally, we discuss the use of generative AI tools for grading or scoring responses to open-ended learning assessments. Manually grading open-ended responses is time and resource intensive. Generative AI tools could help automate at least some parts of this process.

Together, these sub-sections provide an introduction to the range of possibilities for future research at the intersection of generative AI and SAL.

8.7.1 Impact of Generative AI on Learning Outcomes

Generative AI tools have recently made an enormous impact on the fields of both information retrieval and learning sciences. Both fields are exploring how tools like ChatGPT impact search processes and human learning processes. As an emergent technology, relatively few studies have focused specifically on the effect of generative AI tools on human learning outcomes. However, the studies that have investigated impacts on learning outcomes offer insights that are complex. Some studies have found that interventions involving generative AI tools can enhance learning outcomes (Albadarin *et al.*, 2024; Mai *et al.*, 2024; Yilmaz and Karaoglan Yilmaz, 2023). Other studies have found the opposite effect (Bastani *et al.*, 2024; Ju, 2023).

Yilmaz and Karaoglan Yilmaz (2023) investigated the impacts of ChatGPT on computational thinking skills, self-efficacy (i.e., confidence in completing a task), and motivation of students in a programming course. The computational thinking skills assessment is a validated instrument that measures learners' self-perception of computational thinking across five dimensions: creativity, algorithmic thinking, cooperativity, critical thinking, and problem solving (Korkmaz *et al.*, 2017). The study assigned students to one of two conditions, either a ChatGPT condition where students could use ChatGPT to complete their lab assignments or a baseline condition where students completed lab assignments without the use of ChatGPT. Results found that students in the ChatGPT condition had significantly higher levels of computational thinking skills, self-efficacy, and motivation. These findings indicate that ChatGPT may be useful in encouraging and motivating learners. However, the study did not consider the effects on any type of *objective* learning outcome (e.g., differences in grades). Therefore, the extent to which ChatGPT can help students learn programming remains an open question.

Other studies have found that generative AI tools can negatively impact learning outcomes. Bastani *et al.* (2024) investigated the impact of ChatGPT-based tools on learning outcomes. During the study, after receiving a math lecture, participants completed a set of practice math problems with different types of support. Nearly 1,000 students were assigned to one of three conditions. In the baseline condition, participants completed the practice math problems without ChatGPT support. Conversely, in the GPT Base and GPT Tutor conditions, participants had access to ChatGPT while completing the practice math problems. These two conditions differed based on the prompt provided to ChatGPT for each practice math problem. In the GPT Base condition, no prompt was provided. In the GPT Tutor condition, ChatGPT was provided with a prompt containing three different types of information. First, the prompt stated the correct answer to the math problem to help prevent ChatGPT from providing incorrect feedback. Second, the prompt described common misconceptions or mistakes associated with the practice problem. Third, the prompt instructed ChatGPT to provide hints without providing the correct answer. Finally, after completing the practice math problems, all students completed a closed-notes, closed-laptop exam.

Interestingly, results found that performance on the practice math problems was not the same as performance on the exam across conditions. On the practice math problems, performance was highest in the GPT

Tutor condition, followed by the GPT Base condition, followed by the baseline condition. Conversely, on the exam, performance was similar in the GPT Tutor and baseline conditions and *lowest* in the GPT Base condition. These results suggest two important trends. First, generative AI tools may help students complete learning exercises but may not necessarily help with learning and retention. Second, instructors may need to provide generative AI tools with deliberate safeguards (e.g., correct solutions and instructions to provide hints and not answers).

Key Takeaway

Studies have found mixed results on the impact of generative AI tools on learning outcomes. Further work in SAL is needed to better understand this relationship.

It is important that future work in SAL investigates the impact of generative AI tools on both (1) human learning in general and (2) human learning compared to learning with traditional search systems. Given the initial findings of Bastani *et al.* (2024), it may be particularly important that researchers investigate the types of human interventions and safeguards necessary to support meaningful learning. Additionally, it is important to investigate the role of generative AI tools on SRL processes during search. Such work will help researchers to better understand which SRL processes are currently supported by generative AI tools and which need additional scaffolding to help learners.

8.7.2 Generative AI Tools for Document Navigation

Current search systems do little to highlight relations between documents. Imagine a searcher who is learning about Bernoulli's principle. Suppose the searcher is currently looking at a formula of Bernoulli's principle. After realizing that they do not understand it, they may want to read a textual definition. Then, to deepen their understanding, they may want to see examples of Bernoulli's principle applied to everyday phenomena. Current search environments do not support this type of navigation between documents.

Future research should develop ways to link documents as a means for searchers to navigate a collection in a more intentional and structured manner. von Hoyer *et al.* (2022b) argued that, to support learning, future SAL environments should enable searchers to query the system for documents that may contradict the information being read. We argue that this idea should be expanded to include other document-to-document or passage-to-passage relations. Such tools might help searchers find documents or passages that share a specific relation to a current document/passage being read. Example relations include contradiction, corroboration, elaboration, simplification, and illustration, to name a few.

Outside of SAL, prior work has proposed this type of document linking in the context of procedural knowledge. Choi *et al.* (2023) conducted a survey in which people were asked about examples of procedural learning tasks conducted in real life. Part of the survey asked about the types of information participants sought. As expected, participants commented on needing step-by-step information. However, they also needed information about: (1) the input requirements of a step; (2) the outcome of a step (i.e., how the end result of a step should look like); (3) implementation details about a step; (4) the rationale behind a step (i.e., why it is necessary or beneficial); (5) alternative ways to execute a step; and (6) tips on executing the step. The authors argued that future systems should link procedural documents to enable searchers to find such relevant types of information without leaving the current step-by-step document.

Key Takeaway

Generative AI may be useful in developing SAL systems that infer inter-document relationships and support novel ways to navigate a document collection.

Progress in this area requires work in two directions. First, we need a taxonomy of document-to-document relations that is appropriate for SAL. Above, we provide some examples (e.g., contradiction, corroboration, elaboration, etc.). However, this list is not comprehensive nor empirically validated. Erikson and Erlandson (2014) developed a taxonomy of motivations for why academic papers cite each other, which might provide a starting point. Second, we need algorithms to link documents based on a specific taxonomy of relations. Generative AI technologies might help in this respect. For example, given a specific paragraph, one might be able to prompt an LLM to generate text exemplifying a specific relation. For example, "generate text that contradicts that following passage." Then, the system could search for passages in the corpus that are similar to the generated text.

8.7.3 Generative AI Tools to Support SRL Processes

SRL is incredibly important to increasing learning outcomes in SAL systems, as described in Section 7. In Section 8.6, we outlined several future directions for scaffolding SRL within search systems to support learning. Generative AI tools offer additional opportunities to directly and indirectly support SRL processes. However, little work to date has investigated generative AI system interventions to support SRL.

In Section 8.6, we outlined four avenues of exploration toward scaffolding SRL with search systems. Each of these avenues could be further enhanced by generative AI tools. For example, to further support learners in *task understanding*, a learner could use a generative-AI-based chatbot to determine if there are any gaps in their current understanding of the task. For example, a learner could upload their task or assignment to the chatbot. Next, the learner could explain their current understanding of the task and its internal and external constraints (e.g., learner's prior knowledge of the task topic, timeframe to complete task). Then, the chatbot could provide directions on anything the learner has missed (e.g,. "It looks like your instructor wants you to write this literature review in MLA format. Are you familiar with this format?")

Additionally, a generative AI tool could provide nuanced support for *goal-setting* and *planning*. In Section 8.6, we proposed a system that can provide feedback on subgoal characteristics. To provide enhanced support, researchers could develop a prompt for a generative-AI-based chatbot focused on ideal goal characteristics. This prompt would provide the chatbot with definitions of ideal subgoal characteristics and example subgoals with such characteristics. This would allow the learner to be supported with dynamic feedback during the *planning* phase of SRL as they develop subgoals. Also, this chatbot could offer areas of exploration not covered in the content of the subgoals (e.g., "Perhaps you should consider investigating the related subtopic X to make sure you write a comprehensive literature review.").

Finally, a generative-AI-supported search system could nudge learners based on the context of their current document or notes. For example, the generative AI tool could support the SRL monitoring process of *content evaluation*. Perhaps the tool could recognize that the current passage in a document is quite vague and present the learner with a series of related documents that clarify the current document's passage.

Key Takeaway

Generative AI agents can be designed to support and scaffold particular SRL phases like planning and goal-setting in order to improve learning outcomes.

Incorporating SRL scaffolding into search systems offers significant potential for increasing learning outcomes. However, the possibilities expand even further given the capabilities of generative AI tools. Generative AI tools can offer nuanced, targeted support of SRL processes such as *task understanding*, *planning*, and *monitoring* by providing personalized feedback and suggestions that align with the learner's needs and objectives. Although this area of work holds great promise, generative AI is a relatively new area of research. Thus, there is a gap in research exploring the application of generative AI tools to scaffold SRL, particularly in the context of SAL.

8.7.4 Generative AI Tools for Assessment Development

Two particular SRL processes important to improving learning outcomes are called Judgment of Understanding (JOU) and Judgment of Learning (JOL). JOU and JOL reflect a learner's current evaluations of their own understanding about a topic and their ability to use their knowledge to answer test questions (Efklides and Metallidou, 2020). As an example of JOU, after reading a dense mathematical explanation of Bernoulli's principle applied to lift, a learner might say "I don't get it." As an example of JOL, a learner might visit an online educational site with quiz questions and say, "I think I could answer other questions like this on my final exam." In order to support the important SRL processes of JOU and JOL, SAL systems should provide in-context quiz questions so that searchers can calibrate their own learning more accurately. In other words, providing learners with contextually relevant test questions will enable them to understand what they do and do not know. While context-relevant questions have been difficult to generate in the past, generative AI tools offer opportunities to more easily and efficiently integrate such questions into SAL environments.

Key Takeaway

Generative AI can help to support important SRL monitoring processes including JOU and JOL, which involve a learner reflecting on their understanding of information and their ability to use knowledge.

While there is room for improvement, current generative AI tools have shown promise in generating complex, context-relevant multiple-choice questions to test specific domain knowledge. As an example, within the field of medical education, Kiyak *et al.* (2024) used ChatGPT to generate 10 case-based multiple-choice questions to test medical students' knowledge of hypertension. The researchers asked an expert panel to review the set of questions in terms of clarity, accuracy, and relevance to the context (i.e., meeting all constraints that were specified in the prompt provided to ChatGPT). While all 10 questions were found by the experts to be clear and accurate, only 2 were deemed relevant to the context. These two questions were included without any modification in an exam taken by 99 fourth-year medical students. Results found that these two questions successfully differentiated between low- and high-performing students. However, one of the questions had three answer choices that were chosen by fewer than 5% of the students (evidence of a non-ideal multiple choice question). Therefore, as previously noted, there is promise but still room for improvement.

Researchers have provided two main cautionary tips for using ChatGPT to write assessment questions. First, questions generated by ChatGPT must still be scrutinized for accuracy (Han *et al.*, 2024). Second, question quality is tied to prompt quality. Researchers have used prompts that are highly specific in terms of the form of the desired questions (e.g., "provide a medical case that includes patient details, complaints, and medical test results for context"), the format of the answers, and the types of content that the questions and answers should contain (e.g., "include lab values"). For this reason, researchers have made their prompts available for generating questions (Kiyak, 2023; Kiyak *et al.*, 2024). Additionally, some researchers (e.g., Han *et al.*, 2024) have provided the sequence of prompts that led to the best one, demonstrating the *iterative* nature of prompt engineering to generate clear, accurate, and context-relevant questions.

One SAL study investigated the effects of prompting participants to answer questions about passages read during a learning session (Syed *et al.*, 2020). Results found that participants with low prior knowledge improved their learning outcomes when prompted to answer questions that were either automatically generated (in one condition) or manually curated (in another condition). The same trend was not observed for participants with high prior knowledge. Thus, these results suggest that learners with low prior knowledge may benefit the most from tools that prompt them to test their own understanding of content read during a search session.

8.7.5 Generative AI Tools for Grading Open-Ended Assessments

As described in Section 3, SAL studies have used a wide range of methods to measure prior knowledge and learning during search. Compared to closed-ended assessments (e.g., multiple-choice tests), open-ended assessments have several benefits. First, they are easy to develop. Knowledge summaries can simply ask participants to "describe everything you know or learned during the task." Second, because they do not target specific topics, open-ended assessments can help researchers capture everything that a participant learned during a search session. Third, open-ended assessments can help researchers capture depth of learning. For example, based on what is written, researchers can measure a participant's ability to engage in complex cognitive processes (e.g., synthesis, analysis, critical thinking, and creative thinking).

The main drawback of open-ended assessments is that they are difficult to score. Prior SAL studies have scored open-ended responses manually by: (1) counting the number of relevant and/or correct facts, concepts, or ideas included in the response (Collins-Thompson *et al.*, 2016; Abualsaud, 2017; Liu and Song, 2018; Urgo and Arguello, 2024; Demaree *et al.*, 2020; Willoughby *et al.*, 2009; Hornbæk and Frøkjær, 2003); (2) looking for evidence of complex cognitive processes (O'Brien *et al.*, 2020; Palani *et al.*, 2021); or (3) measuring the level of synthesis in the response (Salmerón *et al.*, 2020).

Future SAL studies should consider whether LLMs can be used to automatically score open-ended responses. Outside of SAL, a few recent studies have investigated the use of LLMs to score open-ended responses and provide feedback. Pinto *et al.* (2023) used ChatGPT to score open-ended responses to technical questions answered by 40 industry professionals. ChatGPT was prompted to score responses on a scale of 0-10 and provide feedback. As part of the prompt, ChatGPT was given an example answer written by an expert. Results found that experts largely agreed with the scores and feedback generated by ChatGPT. Henkel *et al.* (2024) used ChatGPT to mark open-ended responses to short-answer questions as correct or incorrect. Similar to Pinto *et al.* (2023), an example correct answer was included in the prompt. ChatGPT agreed with human assessors almost as much as human assessors agreed with each other (i.e., Cohen's $\kappa = .70$ versus .75).

8.8 Collaborative Learning in SAL

Within the learning sciences, computer-supported collaborative learning (CSCL) is a field that studies how people learn together using computers. CSCL research considers situations where people learn together in different settings, including synchronously, asynchronously, co-located, or remotely (Stahl *et al.*, 2005). SAL research has not investigated how people learn together within a collaborative search environment. Within IR, plenty of studies have evaluated search systems that help people collaborate on information-seeking tasks (see Shah, 2012, for a review). Collaborative search systems typically involve a search interface along with visualizations and tools for collaborators to communicate, share information, and gain awareness of each other's activities. Studies have found that such systems can help collaborators delegate tasks, avoid duplicating effort, review each other's work, and keep track of the group's progress. However, collaborative search research has not investigated situations in which learning is the group's main objective. That is, it has not investigated scenarios in which collaborators have a common learning objective and are trying to learn as a group. On the other hand, CSCL research has not investigated scenarios in which learning is mediated by a search system. Such a search environment could enable learners to search independently, have individual and shared workspaces, communicate synchronously (e.g., video conferencing) or asynchronously (instant messaging), and gain awareness of each other's search and learning activities.

CSCL systems are designed to help students learn together by providing, not only instructional materials, but also media for communication and scaffolding to encourage and support productive interactions between learners. During a typical CSCL study, participants learn together by sharing their learning experiences through messages and/or visual displays. Ideally, learning happens by participants identifying each other's knowledge gaps (i.e., unknown unknowns), externalizing and sharing their knowledge, and leveraging each other's distinct prior knowledge and skills. CSCL studies consider how social interactions influence the learning processes of individual group members. As noted by Stahl *et al.* (2005), because participants exchange their current understanding throughout the learning process, CSCL researchers can closely examine how each group member's knowledge evolves over the learning session. In this respect, paradoxically, understanding group learning can be easier than understanding learning by individuals working alone.

CSCL studies have investigated how different factors impact group learning (Miyake and Kirschner, 2014). Studies have considered factors related to: (1) the learning task; (2) the way that group member's interact with each other; (3) attitudes held by group members; and (4) the composition of the group. For example, in terms of the learning task, studies have investigated *task interdependence* and *outcome interdependence*. Task interdependence refers to the extent to which successfully completing one sub-task relies on successfully completing other sub-tasks. Studies have found that task interdependence leads to more communication, helping, and information sharing. Outcome interdependence refers to the extent to which personal gains depend on the success of other team members. Studies have found that outcome interdependence leads to group members being more open-minded, concerned about each other's outcomes, and more likely to compromise. In terms of group member interactions, studies have found that learners have better outcomes when they arrive at a shared understanding of the learning task and how to approach it. In terms of attitudes, studies have found that learning improves when collaborators see the group as being capable and effective. Finally, CSCL studies have also investigated team composition. For example, studies have considered how gender and skill diversity within the group impacts learning outcomes (Cen *et al.*, 2016).

Future SAL studies should investigate learning by groups in scenarios that involve information seeking as an important activity. Future SAL studies could investigate the challenges associated with group learning and develop tools and interventions to support group learning.

8.9 Summary

While prior work in SAL has provided many initial insights into learning during search, many important areas of investigation remain unexplored. First, limited work in SAL has investigated transfer of learning. SAL research should focus on developing transfer of learning instruments to measure transfer outcomes. Second, SAL researchers should develop search environments that consider characteristics of the learning context. These might include the topical domain, the learning objective, and characteristics of the individual searcher (e.g., prior knowledge, cognitive abilities, personality traits). Third, SAL studies have mostly focused on single search sessions. Future work in SAL should focus on longitudinal studies to better understand how to support searchers during complex learning-oriented tasks that involve an extended time period (e.g., learning a new skill). Fourth, future work should consider self-determined learning. This relatively new paradigm highlights important aspects of the learning environment to teach students how to teach themselves. Fifth, future SAL studies should investigate learning within the context of heavily debated topics, which may pose unique challenges and opportunities. Sixth, little work in SAL has focused on supporting SRL processes. Future work should focus on capturing and scaffolding of SRL in search environments. In particular, such support should be scaffolded in an intelligent and nuanced way so that only work that is *not* useful for learning is offloaded. Seventh, future research should leverage and build upon existing generative AI tools to help individuals achieve complex, learning-oriented goals. Finally, most SAL research to date has studied how people learn on their own. Future research should also consider how groups of searchers learn together. By building upon prior work in computer-supported collaborative learning (CSCL), future research might be able to develop tools and interventions to support group learning during tasks in which information seeking is a key component.

Search as learning research has revealed deep insights into how humans process, organize, and internalize knowledge while using search environments. Researchers have made great progress in understanding the learning process during search and the types of search system features that can benefit learning. We encourage researchers to push further to build on these discoveries. Methodologies from learning sciences such as capturing SRL and emerging technologies such as adaptive systems powered by generative AI are valuable tools for exploring new dimensions of SAL. Closing gaps in current SAL research will ultimately enhance how humans learn through search, empowering everyone, regardless of access, to tackle complex topics, think critically, and expand the boundaries of their knowledge.

References

Abualsaud, M. (2017). "Learning Factors and Determining Document-level Satisfaction In Search-as-Learning". *MA thesis*. Waterloo, Ontario, Canada: University of Waterloo.

Acee, T. W., Y. Cho, J.-I. Kim, and C. E. Weinstein. (2012). "Relationships among properties of college students' self-set academic goals and academic achievement". *Educational Psychology*. 32(6): 681–698. doi: 10.1080/01443410.2012.712795. (Accessed on 03/21/2022).

Agosti, M., N. Fuhr, E. Toms, and P. Vakkari. (2014). "Evaluation methodologies in information retrieval dagstuhl seminar 13441". *ACM SIGIR Forum*. 48(1): 36–41. doi: 10.1145/2641383.2641390. (Accessed on 04/28/2020).

Albadarin, Y., M. Saqr, N. Pope, and M. Tukiainen. (2024). "A systematic literature review of empirical research on ChatGPT in education". *Discover Education*. 3(1): 60. doi: 10.1007/s44217-024-00138-2. (Accessed on 07/30/2024).

ALDayel, A. and W. Magdy. (2021). "Stance detection on social media: State of the art and trends". *Information Processing & Management*. 58(4): 102597. doi: https://doi.org/10.1016/j.ipm.2021.102597.

Alessandri, G., L. Borgogni, G. P. Latham, G. Cepale, A. Theodorou, and E. De Longis. (2020). "Self-set goals improve academic performance through nonlinear effects on daily study performance". *Learning and Individual Differences*. 77(Jan.): 101784. DOI: 10.1016/j.lindif.2019.101784. (Accessed on 02/03/2021).

Aleven, V., B. McLaren, I. Roll, and K. Koedinger. (2006). "Toward Meta-cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor". *International Journal of Artificial Intelligence in Education*. 16(2): 101–128.

Allan, J., B. Croft, A. Moffat, and M. Sanderson. (2012). "Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne". *ACM SIGIR Forum*. 46(1): 2. DOI: 10.1145/2215676.2215678. (Accessed on 09/11/2018).

Anderson, J. R., C. F. Boyle, and B. J. Reiser. (1985). "Intelligent Tutoring Systems". *Science*. 228(4698): 456–462. URL: http://www.jstor.org/stable/1694721 (accessed on 01/18/2021).

Anderson, L. W., D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock. (2000). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition*. 1 edition. New York: Pearson.

Arguello, J., W.-C. Wu, D. Kelly, and A. Edwards. (2012). "Task Complexity, Vertical Display and User Interaction in Aggregated Search". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12*. New York, NY, USA: ACM. 435–444. DOI: 10.1145/2348283.2348343. (Accessed on 03/20/2019).

Azevedo, R., J. Harley, G. Trevors, M. Duffy, R. Feyzi-Behnagh, F. Bouchet, and R. Landis. (2013). "Using Trace Data to Examine the Complex Roles of Cognitive, Metacognitive, and Emotional Self-Regulatory Processes During Learning with Multi-agent Systems". In: *International Handbook of Metacognition and Learning Technologies*. Ed. by R. Azevedo and V. Aleven. *Springer International Handbooks of Education*. New York, NY: Springer. 427–449. DOI: 10.1007/978-1-4419-5546-3_28. (Accessed on 01/26/2021).

Azevedo, R., R. S. Landis, R. Feyzi-Behnagh, M. Duffy, G. Trevors, J. M. Harley, F. Bouchet, J. Burlison, M. Taub, N. Pacampara, M. Yeasin, A. K. M. M. Rahman, M. I. Tanveer, and G. Hossain. (2012). "The effectiveness of pedagogical agents' prompting and feedback in facilitating co-adapted learning with metatutor". In: *Proceedings of the 11th international conference on Intelligent Tutoring Systems. ITS'12*. Berlin, Heidelberg: Springer-Verlag. 212–221. DOI: 10.1007/978-3-642-30950-2_27. (Accessed on 01/30/2021).

Azevedo, R., S. Ragan, J. G. Cromley, and S. Pritchett. (2002). *Do Different Goal-Setting Conditions Facilitate Students' Ability to Regulate Their Learning of Complex Science Topics with RiverWeb?* URL: https://eric.ed.gov/?id=ED482509 (accessed on 04/23/2020).

Azevedo, R., A. Witherspoon, A. Chauncey, C. Burkett, and A. Fike. (2009). "MetaTutor: A MetaCognitive tool for enhancing self-regulated learning". In: *Cognitive and Metacognitive Educational Systems - Papers from the AAAI Fall Symposium, Technical Report*. 14–19. URL: https://asu.pure.elsevier.com/en/publications/metatutor-a-metacognitive-tool-for-enhancing-self-regulated-learn (accessed on 01/16/2021).

Azzopardi, L. (2021). "Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval". In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. Canberra ACT Australia: ACM. 27–37. DOI: 10.1145/3406522.3446023. (Accessed on 01/09/2025).

Azzopardi, L., D. Maxwell, M. Halvey, and C. Hauff. (2023). "Driven to Distraction: Examining the Influence of Distractors on Search Behaviours, Performance and Experience". In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. CHIIR '23*. Austin, TX, USA: Association for Computing Machinery. 83–94. DOI: 10.1145/3576840.3578298.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory. Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ, US: Prentice-Hall, Inc.

Bandura, A. (2010). "Self-Efficacy". In: *The Corsini Encyclopedia of Psychology*. American Cancer Society. 1–3. DOI: 10.1002/9780470479216.corpsy0836. (Accessed on 02/17/2021).

Bannert, M., M. Hildebrand, and C. Mengelkamp. (2009). "Effects of a metacognitive support device in learning environments". *Computers in Human Behavior*. Including the Special Issue: The Use of Support Devices in Electronic Learning Environments 25(4): 829–835. DOI: 10.1016/j.chb.2008.07.002. (Accessed on 02/03/2021).

Barnett, S. M. and S. J. Ceci. (2002). "When and where do we apply what we learn?: A taxonomy for far transfer." *Psychological Bulletin*. 128(4): 612–637. DOI: 10.1037/0033-2909.128.4.612. (Accessed on 04/07/2021).

Bastani, H., O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman. (2024). "Generative AI Can Harm Learning". *SSRN Scholarly Paper*. Rochester, NY. DOI: 10.2139/ssrn.4895486. (Accessed on 07/25/2024).

Belland, B. R., A. E. Walker, N. J. Kim, and M. Lefler. (2017). "Synthesizing Results From Empirical Research on Computer-Based Scaffolding in STEM Education: A Meta-Analysis". *Review of Educational Research*. 87(2): 309–344. DOI: 10.3102/0034654316670999. (Accessed on 10/28/2020).

Bernacki, M. (2010). "The influence of self-regulated learning and prior knowledge on knowledge acquisition in computer-based learning environments". *Ph.D.* United States – Pennsylvania: Temple University. URL: https://www.proquest.com/docview/609971335/abstract/B1B99D89827949E1PQ/1 (accessed on 07/19/2024).

Berry, J. L. (2012). "The Effects of Concept Mapping and Questioning on Students' Organization and Retention of Science Knowledge While Using Interactive Read-Alouds". *Thesis*. URL: https://oaktrust.library.tamu.edu/handle/1969.1/ETD-TAMU-2011-08-10164 (accessed on 08/09/2024).

Bhattacharya, N. (2023). "LongSAL: A Longitudinal Search as Learning Study with University Students". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. CHI EA '23*. New York, NY, USA: Association for Computing Machinery. 1–8. DOI: 10.1145/3544549.3583948. (Accessed on 06/26/2023).

Bhattacharya, N. and J. Gwizdka. (2019). "Measuring Learning During Search: Differences in Interactions, Eye-Gaze, and Semantic Similarity to Expert Knowledge". In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. Glasgow Scotland UK: ACM. 63–71. DOI: 10.1145/3295750.3298926. (Accessed on 12/14/2020).

Blaschke, L. M. (2012). "Heutagogy and lifelong learning: A review of heutagogical practice and self-determined learning". *The International Review of Research in Open and Distributed Learning*. 13(1): 56. DOI: 10.19173/irrodl.v13i1.1076. (Accessed on 01/26/2024).

Blaschke, L. M. (2018). "Self-determined Learning (Heutagogy) and Digital Media Creating integrated Educational Environments for Developing Lifelong Learning Skills". In: *The Digital Turn in Higher Education*. Ed. by D. Kergel, B. Heidkamp, P. K. Telléus, T. Rachwal, and S. Nowakowski. Wiesbaden: Springer Fachmedien Wiesbaden. 129–140. DOI: 10.1007/978-3-658-19925-8_10. (Accessed on 02/02/2024).

Blaschke, L. M. and S. Hase. (2016). "Heutagogy: A Holistic Framework for Creating Twenty-First-Century Self-determined Learners". In: *The Future of Ubiquitous Learning: Learning Designs for Emerging Pedagogies*. Ed. by B. Gros, Kinshuk, and M. Maina. *Lecture Notes in Educational Technology*. Berlin, Heidelberg: Springer. 25–40. DOI: 10.1007/978-3-662-47724-3_2. (Accessed on 01/26/2024).

Bloom, B. S. (1956). "Taxonomy of educational objectives. Vol. 1: Cognitive domain". *New York: McKay*: 20–24.

Boekaerts, M. (1995). "Self-regulated learning: Bridging the gap between metacognitive and metamotivation theories". *Educational Psychologist*. 30(4): 195–200. DOI: 10.1207/s15326985ep3004_4. (Accessed on 06/07/2024).

Boekaerts, M. and E. Cascallar. (2006). "How Far Have We Moved Toward the Integration of Theory and Practice in Self-Regulation?" *Educational Psychology Review*. 18(3): 199–210. DOI: 10.1007/s10648-006-9013-4. (Accessed on 06/07/2024).

Boekaerts, M., M. Zeidner, P. R. Pintrich, and P. R. Pintrich. (1999). *Handbook of Self-Regulation*. San Diego, UNITED STATES: Elsevier Science & Technology. URL: http://ebookcentral.proquest.com/lib/unc/detail.action?docID=300645 (accessed on 02/04/2021).

Boom, G. van den, F. Paas, and J. J. G. van Merriënboer. (2007). "Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes". *Learning and Instruction*. 17(5): 532–548. DOI: 10.1016/j.learninstruc.2007.09.003. (Accessed on 02/03/2021).

Brennan, K., D. Kelly, and J. Arguello. (2014). "The Effect of Cognitive Abilities on Information Search for Tasks of Varying Levels of Complexity". In: *Proceedings of the 5th Information Interaction in Context Symposium. IIiX '14*. New York, NY, USA: ACM. 165–174. DOI: 10.1145/2637002.2637022. (Accessed on 03/17/2019).

Broadbent, D. E., P. F. Cooper, P. FitzGerald, and K. R. Parkes. (1982). "The cognitive failures questionnaire (CFQ) and its correlates". *British journal of clinical psychology*. 21(1): 1–16.

Butler, D. L. and P. H. Winne. (1995). "Feedback and Self-Regulated Learning: A Theoretical Synthesis". *Review of Educational Research*. 65(3): 245–281. DOI: 10.3102/00346543065003245. (Accessed on 02/04/2021).

Cacioppo, J. T., R. E. Petty, J. A. Feinstein, and W. B. G. Jarvis. (1996). "Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition." *Psychological bulletin*. 119(2): 197.

Cacioppo, J. T., R. E. Petty, and C. Feng Kao. (1984). "The efficient assessment of need for cognition". *Journal of personality assessment*. 48(3): 306–307.

Câmara, A., N. Roy, D. Maxwell, and C. Hauff. (2021). "Searching to Learn with Instructional Scaffolding". In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. CHIIR '21*. New York, NY, USA: Association for Computing Machinery. 209–218. DOI: 10.1145/3406522.3446012. (Accessed on 03/15/2021).

Câmara, A. and D. El-Zein. (2022). "RULK: A Framework for Representing User Knowledge in Search-as-Learning". In: *3rd International Conference on Design of Experimental Search & Information REtrieval Systems*.

Capra, R., J. Arguello, A. Crescenzi, and E. Vardell. (2015). "Differences in the Use of Search Assistance for Tasks of Varying Complexity". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15*. New York, NY, USA: ACM. 23–32. DOI: 10.1145/2766462.2767741. (Accessed on 11/26/2018).

Capra, R., J. Arguello, H. O'Brien, Y. Li, and B. Choi. (2018). "The Effects of Manipulating Task Determinability on Search Behaviors and Outcomes". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. New York, NY, USA: ACM. 445–454. DOI: 10.1145/3209978.3210047. (Accessed on 09/15/2019).

Carbonell, J. R. (1970). "AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction". *IEEE Transactions on Man-Machine Systems*. 11(4): 190–202. DOI: 10.1109/TMMS.1970.299942.

Cen, L., D. Ruta, L. Powell, B. Hirsch, and J. Ng. (2016). "Quantitative approach to collaborative learning: Performance prediction, individual assessment, and group composition". *International Journal of Computer-Supported Collaborative Learning*. 11: 187–225.

Chang, W.-C. and Y.-M. Ku. (2015). "The Effects of Note-Taking Skills Instruction on Elementary Students' Reading". *The Journal of Educational Research*. 108(4): 278–291. DOI: 10.1080/00220671.2014.886175.

Chi, M. T. H., S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann. (2001). "Learning from human tutoring". *Cognitive Science*. 25(4): 471–533. DOI: https://doi.org/10.1207/s15516709cog2504_1. (Accessed on 02/22/2021).

Chi, Y., S. Han, D. He, and R. Meng. (2016). "Exploring knowledge learning in collaborative information seeking process". In: *CEUR Workshop Proceedings*. Vol. 1647. URL: http://d-scholarship.pitt.edu/32364/ (accessed on 07/27/2021).

Chi, Y. (2019). "Examining and Supporting Laypeople's Learning in Online Health Information Seeking". In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. CHIIR '19*. New York, NY, USA: Association for Computing Machinery. 425–428. DOI: 10.1145/3295750.3298975. (Accessed on 09/07/2020).

Choi, B. and J. Arguello. (2020). "A Qualitative Analysis of the Effects of Task Complexity on the Functional Role of Information". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. CHIIR '20*. Vancouver BC, Canada: Association for Computing Machinery. 328–332. DOI: 10.1145/3343413.3377992.

Choi, B., J. Arguello, and R. Capra. (2023). "Understanding Procedural Search Tasks "in the Wild"". In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. CHIIR '23*. Austin, TX, USA: Association for Computing Machinery. 24–33. DOI: 10.1145/3576840.3578302.

Choi, B., R. Capra, and J. Arguello. (2019a). "The Effects of Working Memory during Search Tasks of Varying Complexity". In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. CHIIR '19*. Glasgow, Scotland UK: Association for Computing Machinery. 261–265. DOI: 10.1145/3295750.3298948.

Choi, B., A. Ward, Y. Li, J. Arguello, and R. Capra. (2019b). "The Effects of Task Complexity on the Use of Different Types of Information in a Search Assistance Tool". *ACM Trans. Inf. Syst.* 38(1). DOI: 10.1145/3371707.

Chou, C.-Y. and N.-B. Zou. (2020). "An analysis of internal and external feedback in self-regulated learning activities mediated by self-regulated learning tools and open learner models". *International Journal of Educational Technology in Higher Education*. 17(1): 55. DOI: 10.1186/s41239-020-00233-y. (Accessed on 02/04/2021).

Cole, A. W. (2022). "Understanding self-efficacy in search as self-determined learning". *PhD thesis*. University of British Columbia. DOI: 10.14288/1.0416302. (Accessed on 06/26/2023).

Collins-Thompson, K., P. Hansen, and C. Hauff. (2017). "Search as Learning (Dagstuhl Seminar 17092)". Ed. by M. Herbstritt. DOI: 10.4230/dagrep.7.2.135.

Collins-Thompson, K., S. Y. Rieh, C. C. Haynes, and R. Syed. (2016). "Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. CHIIR '16*. New York, NY, USA: ACM. 163–172. DOI: 10.1145/2854946.2854972. (Accessed on 09/26/2018).

Conklin, J. (2005). "Review of A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives Complete Edition". *Educational Horizons*. 83(3): 154–159. URL: https://www.jstor.org/stable/42926529 (accessed on 05/16/2024).

Corbett, A. T., K. R. Koedinger, and J. R. Anderson. (1997). "Intelligent Tutoring Systems". In: *Handbook of Human-Computer Interaction*. Amsterdam: Elsevier. 849–874.

Council, N. R. and Education. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. National Academies Press.

Crescenzi, A., A. R. Ward, Y. Li, and R. Capra. (2021). "Supporting Metacognition during Exploratory Search with the OrgBox". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery. 1197–1207. URL: http://doi.org/10.1145/3404835.3462955 (accessed on 05/09/2022).

David, L., F. Biwer, R. Crutzen, and A. de Bruin. (2024). "The challenge of change: understanding the role of habits in university students' self-regulated learning". *Higher Education*. Apr. DOI: 10.1007/s10734-024-01199-w. (Accessed on 07/12/2024).

Davies, S., K. Butcher, and C. Stevens. (2013). "Self-Regulated Learning with Graphical Overviews: When Spatial Information Detracts from Learning". *Proceedings of the Annual Meeting of the Cognitive Science Society*. 35(35). URL: https://escholarship.org/uc/item/5kp8b0n9 (accessed on 07/27/2021).

Day, S. B. and R. L. Goldstone. (2012). "The Import of Knowledge Export: Connecting Findings and Theories of Transfer of Learning". *Educational Psychologist*. 47(3): 153–176. DOI: 10.1080/00461520.2012.696438. (Accessed on 09/10/2021).

Deekens, V. M., J. A. Greene, and N. G. Lobczowski. (2018). "Monitoring and depth of strategy use in computer-based learning environments for science and history". *British Journal of Educational Psychology*. 88(1): 63–79. DOI: 10.1111/bjep.12174. (Accessed on 04/15/2023).

Demaree, D., H. Jarodzka, S. Brand-Gruwel, and Y. Kammerer. (2020). "The Influence of Device Type on Querying Behavior and Learning Outcomes in a Searching as Learning Task with a Laptop or Smartphone". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. CHIIR '20*. New York, NY, USA: Association for Computing Machinery. 373–377. DOI: 10.1145/3343413.3378000. (Accessed on 08/26/2020).

Diamond, A. (2013). "Executive functions". *Annual review of psychology*. 64(1): 135–168.

Draws, T., O. Inel, N. Tintarev, C. Baden, and B. Timmermans. (2022). "Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics". In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval. CHIIR '22*. Regensburg, Germany: Association for Computing Machinery. 135–145. DOI: 10.1145/3498366.3505812.

Draws, T., N. Tintarev, and U. Gadiraju. (2021). "Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics". *SIGKDD Explor. Newsl.* 23(1): 50–58. DOI: 10.1145/3468507.3468515.

Earley, P. C. (1985). "Influence of information, choice and task complexity upon goal acceptance, performance, and personal goals". *Journal of Applied Psychology*. 70(3): 481–491. DOI: 10.1037/0021-9010.70.3.481. (Accessed on 02/03/2021).

Efklides, A. and P. Metallidou. (2020). "Applying Metacognition and Self-Regulated Learning in the Classroom". In: *Oxford Research Encyclopedia of Education*. URL: https://oxfordre.com/education/display/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-961 (accessed on 08/02/2024).

Eickhoff, C., J. Gwizdka, C. Hauff, and J. He. (2017). “Introduction to the special issue on search as learning”. *Information Retrieval Journal*. 20(5): 399–402. DOI: 10.1007/s10791-017-9315-9. (Accessed on 11/24/2018).

Eickhoff, C., J. Teevan, R. White, and S. Dumais. (2014). “Lessons from the journey: a query log analysis of within-session learning”. In: *Proceedings of the 7th ACM international conference on Web search and data mining - WSDM ’14*. New York, New York, USA: ACM Press. 223–232. DOI: 10.1145/2556195.2556217. (Accessed on 02/12/2019).

El Zein, D., A. Câmara, C. Da Costa Pereira, and A. Tettamanzi. (2023). “RULKNE: Representing User Knowledge State in Search-as-Learning with Named Entities”. In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. CHIIR ’23*. New York, NY, USA: Association for Computing Machinery. 388–393. DOI: 10.1145/3576840.3578330. (Accessed on 06/26/2023).

Elliott, E. and C. Dweck. (1988). “Goals: An approach to motivation and achievement.” *Journal of Personality and Social Psychology*. 54(1): 5–12. (Accessed on 02/01/2021).

Erikson, M. G. and P. Erlandson. (2014). “A taxonomy of motives to cite”. *Social studies of science*. 44(4): 625–637.

Evens, M. W., R.-C. Chang, Y. H. Lee, L. S. Shim, C. W. Woo, and Y. Zbang. (1997). “CIRCSIM-Tutor: An Intelligent Tutoring System Using Natural Language Dialogue”. In: *Fifth Conference on Applied Natural Language Processing: Descriptions of System Demonstrations and Videos*. Washington, DC, USA: Association for Computational Linguistics. 13–14. DOI: 10.3115/974281.974289. (Accessed on 09/06/2023).

Fisher, K. M., K. S. Williams, and J. E. Lineback. (2011). “Osmosis and Diffusion Conceptual Assessment”. *CBE—Life Sciences Education*. 10(4): 418–429. DOI: 10.1187/cbe.11-04-0038. (Accessed on 04/30/2021).

French, J., H. Harman, and D. Dermen. (1976). “Manual for kit of factor referenced cognitive tests”. *Educational Testing Service, Princeton, NJ*: 109–113.

Freund, L., J. He, J. Gwizdka, N. Kando, P. Hansen, and S. Y. Rieh. (2014). "Searching As Learning (SAL) Workshop 2014". In: *Proceedings of the 5th Information Interaction in Context Symposium. IIiX '14*. New York, NY, USA: ACM. 7–7. DOI: 10.1145/2637002.2643203. (Accessed on 11/24/2018).

Freund, L., R. Kopak, and H. O'Brien. (2016). "The effects of textual environment on reading comprehension: Implications for searching as learning". *Journal of Information Science*. 42(1): 79–93. DOI: 10.1177/0165551515614472. (Accessed on 04/23/2020).

Gadiraju, U., R. Yu, S. Dietze, and P. Holtz. (2018). "Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web". In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. CHIIR '18*. New York, NY, USA: ACM. 2–11. DOI: 10.1145/3176349.3176381. (Accessed on 11/29/2018).

Ghosh, S., M. Rath, and C. Shah. (2018). "Searching As Learning: Exploring Search Behavior and Learning Outcomes in Learning-related Tasks". In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. CHIIR '18*. New York, NY, USA: ACM. 22–31. DOI: 10.1145/3176349.3176386. (Accessed on 10/22/2018).

Glogger, I., R. Schwonke, L. Holzäpfel, M. Nückles, and A. Renkl. (2012). "Learning strategies assessed by journal writing: Prediction of learning outcomes by quantity, quality, and combinations of learning strategies". *Journal of Educational Psychology*. 104(2): 452–468. DOI: 10.1037/a0026683.

González-Betancor, S. M., A. Bolívar-Cruz, and D. Verano-Tacoronte. (2019). "Self-assessment accuracy in higher education: The influence of gender and performance of university students". *Active Learning in Higher Education*. 20(2): 101–114. DOI: 10.1177/1469787417735604. (Accessed on 02/16/2021).

Greene, J. A., K. R. Dellinger, B. B. Tüysüzoğlu, and L.-J. Costa. (2013). "A Two-Tiered Approach to Analyzing Self-Regulated Learning Data to Inform the Design of Hypermedia Learning Environments". In: *International Handbook of Metacognition and Learning Technologies*. Ed. by R. Azevedo and V. Aleven. *Springer International Handbooks of Education*. New York, NY: Springer. 117–128. DOI: 10.1007/978-1-4419-5546-3_8. (Accessed on 08/26/2020).

Greene, J. A., L. A. Hutchison, L.-J. Costa, and H. Crompton. (2012). "Investigating how college students' task definitions and plans relate to self-regulated learning processing and understanding of a complex science topic". *Contemporary Educational Psychology*. 37(4): 307–320. DOI: 10.1016/j.cedpsych.2012.02.002. (Accessed on 04/23/2020).

Greene, J. A., N. G. Lobczowski, R. Freed, B. M. Cartiff, C. Demetriou, and A. T. Panter. (2020). "Effects of a Science of Learning Course on College Students' Learning With a Computer". *American Educational Research Journal*. 57(3): 947–978. DOI: 10.3102/0002831219865221. (Accessed on 02/03/2023).

Greene, J. A. and R. Azevedo. (2007). "A Theoretical Review of Winne and Hadwin's Model of Self-Regulated Learning: New Perspectives and Directions". *Review of Educational Research*. 77(3): 334–372. DOI: 10.3102/003465430303953. (Accessed on 04/23/2020).

Greene, J. A., C. M. Bolick, W. P. Jackson, A. M. Caprino, C. Oswald, and M. McVea. (2015). "Domain-specificity of self-regulated learning processing in science and history". *Contemporary Educational Psychology*. 42(July): 111–128. DOI: 10.1016/j.cedpsych.2015.06.001. (Accessed on 01/04/2022).

Greene, J. A., D. Z. Copeland, V. M. Deekens, and S. B. Yu. (2018). "Beyond knowledge: Examining digital literacy's role in the acquisition of understanding in science". *Computers & Education*. 117(Feb.): 141–159. DOI: 10.1016/j.compedu.2017.10.003. (Accessed on 02/03/2023).

Gwizdka, J., P. Hansen, C. Hauff, J. He, and N. Kando. (2016). “Search As Learning (SAL) Workshop 2016”. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’16*. New York, NY, USA: ACM. 1249–1250. DOI: 10.1145/2911451.2917766. (Accessed on 11/24/2018).

Hadwin, A. F., R. Rostampour, and P. H. Winne. (2025). “Advancing Self-Reports of Self-Regulated Learning: Validating New Measures to Assess Students’ Beliefs, Practices, and Challenges”. *Educational Psychology Review*. 37(1): 8. DOI: 10.1007/s10648-024-09977-9. (Accessed on 01/31/2025).

Hadwin, A. F., P. H. Winne, D. B. Stockley, J. C. Nesbit, and C. Woszczyna. (2001). “Context moderates students’ self-reports about how they study”. *Journal of Educational Psychology*. 93(3): 477–487. DOI: 10.1037/0022-0663.93.3.477.

Hake, R. (2002). “Relationship of Individual Student Normalized Learning Gains in Mechanics with Gender , High-School Physics , and Pretest Scores on Mathematics and Spatial Visualization”. In: vol. 8. 1–14. URL: https://www.semanticscholar.org/paper/Relationship-of-Individual-Student-Normalized-Gains-Hake/ab557de0fdafe5def057a795c25264e74ac0e332.

Han, Z., F. Battaglia, A. Udaiyar, A. Fooks, and S. R. Terlecky. (2024). “An explorative assessment of ChatGPT as an aid in medical education: Use it with caution”. *Medical Teacher*. 46(5): 657–664. DOI: 10.1080/0142159X.2023.2271159. (Accessed on 08/02/2024).

Hanfstingl, B., A. Arzenšek, J. Apschner, and K. I. Gölly. (2021). “Assimilation and Accommodation”. *European Psychologist*. Nov. URL: https://econtent.hogrefe.com/doi/10.1027/1016-9040/a000463 (accessed on 01/09/2025).

Hansen, P. and S. Y. Rieh. (2016). “Editorial: Recent advances on searching as learning: An introduction to the special issue”. *Journal of Information Science*. 42(1): 3–6. DOI: 10.1177/0165551515614473. (Accessed on 11/24/2018).

Harju, V., A. Koskinen, and L. Pehkonen. (2019). “An exploration of longitudinal studies of digital learning”. *Educational Research*. 61(4): 388–407. DOI: 10.1080/00131881.2019.1660586.

Harley, J. M., M. Taub, R. Azevedo, and F. Bouchet. (2018). "Let's Set Up Some Subgoals: Understanding Human-Pedagogical Agent Collaborations and Their Implications for Learning and Prompt and Feedback Compliance". *IEEE Transactions on Learning Technologies*. 11(1): 54–66. DOI: 10.1109/TLT.2017.2756629.

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning. Transfer of learning: Cognition, instruction, and reasoning*. San Diego, CA, US: Academic Press. DOI: 10.1016/B978-012330595-4/50003-2.

Heilman, M., K. Collins-Thompson, J. Callan, M. Eskenazi, A. Juffs, and L. Wilson. (2010). "Personalization of Reading Passages Improves Vocabulary Acquisition". *International Journal of Artificial Intelligence in Education*. 20(1): 73–98. DOI: 10.3233/JAI-2010-0003. (Accessed on 07/27/2021).

Heilman, M. and M. Eskenazi. (2006). "Language Learning: Challenges for Intelligent Tutoring Systems". In: *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*.

Henkel, O., L. Hills, A. Boxer, B. Roberts, and Z. Levonian. (2024). "Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education". In: *Proceedings of the Eleventh ACM Conference on Learning @ Scale. L@S '24*. Atlanta, GA, USA: Association for Computing Machinery. 300–304. DOI: 10.1145/3657604.3664693.

Hersh, W. R., D. L. Elliot, D. H. Hickam, S. L. Wolf, and A. Molnar. (1995). "Towards new measures of information retrieval evaluation". In: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '95*. Seattle, Washington, United States: ACM Press. 164–170. DOI: 10.1145/215206.215355. (Accessed on 09/03/2020).

Hollenbeck, J. R. and H. J. Klein. (1987). "Goal commitment and the goal-setting process: Problems, prospects, and proposals for future research". *Journal of Applied Psychology*. 72(2): 212–220. DOI: 10.1037/0021-9010.72.2.212.

Hoppe, A., R. Yu, I. Brich, and J. Liu. (2021). “IWILDS’21: Second International Workshop on Learning During Web Search”. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM ’21*. New York, NY, USA: Association for Computing Machinery. 4880–4881. DOI: 10.1145/3459637.3482034. (Accessed on 01/08/2025).

Hoppe, A., R. Yu, Y. Kammerer, and L. Salmerón. (2020). “IWILDS’20: The 1st International Workshop on Investigating Learning during Web Search”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Virtual Event Ireland: ACM. 3535–3536. DOI: 10.1145/3340531.3414076. (Accessed on 10/27/2020).

Hoppe, A., R. Yu, and J. Liu. (2022). “IWILDS’22 – Third International Workshop on Investigating Learning During Web Search”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR ’22*. New York, NY, USA: Association for Computing Machinery. 3482–3484. DOI: 10.1145/3477495.3531698. (Accessed on 01/08/2025).

Hornbæk, K. and E. Frøkjær. (2003). “Reading patterns and usability in visualizations of electronic documents”. *ACM Transactions on Computer-Human Interaction*. 10(2): 119–149. DOI: 10.1145/772047.772050. (Accessed on 09/06/2020).

Hoyer, J. von, G. Pardi, Y. Kammerer, and P. Holtz. (2019). “Metacognitive Judgments in Searching as Learning (SAL) Tasks: Insights on (Mis-) Calibration, Multimedia Usage, and Confidence”. In: *Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information. SALMM ’19*. New York, NY, USA: Association for Computing Machinery. 3–10. DOI: 10.1145/3347451.3356730. (Accessed on 09/07/2020).

Hoyer, J. F. von, J. Kimmerle, and P. Holtz. (2022a). “Acquisition of false certainty: Learners increase their confidence in the correctness of incorrect answers after online information search”. *Journal of Computer Assisted Learning*. 38(3): 833–844. DOI: 10.1111/jcal.12657. (Accessed on 05/09/2022).

Hu, X. and N. Kando. (2017). "Task complexity and difficulty in music information retrieval". *Journal of the Association for Information Science and Technology*. 68(7): 1711–1723. DOI: 10.1002/asi.23803. (Accessed on 05/02/2019).

Jansen, B. J., D. Booth, and B. Smith. (2009). "Using the taxonomy of cognitive learning to model online searching". *Information Processing & Management*. 45(6): 643–663. DOI: 10.1016/j.ipm.2009.05.004. (Accessed on 02/08/2019).

Jones, N. A., H. Ross, T. Lynam, P. Perez, and A. Leitch. (2011). "Mental Models: An Interdisciplinary Synthesis of Theory and Methods". *Ecology and Society*. 16(1). URL: http://www.jstor.org/stable/26268859 (accessed on 03/23/2021).

Ju, Q. (2023). "Experimental Evidence on Negative Impact of Generative AI on Scientific Learning Outcomes". DOI: 10.48550/arXiv.2311.05629. (Accessed on 07/30/2024).

Kalyani, R. and U. Gadiraju. (2019). "Understanding User Search Behavior Across Varying Cognitive Levels". In: *Proceedings of the 30th ACM Conference on Hypertext and Social Media. HT '19*. New York, NY, USA: Association for Computing Machinery. 123–132. DOI: 10.1145/3342220.3343643. (Accessed on 09/07/2020).

Kammerer, Y., R. Nairn, P. Pirolli, and E. H. Chi. (2009). "Signpost from the masses: learning effects in an exploratory social tag search browser". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '09*. Boston, MA, USA: Association for Computing Machinery. 625–634. DOI: 10.1145/1518701.1518797. (Accessed on 04/23/2020).

Kelly, D. (2009). "Methods for Evaluating Interactive Information Retrieval Systems with Users". *Foundations and Trends® in Information Retrieval*. 3(1–2): 1–224. DOI: 10.1561/1500000012. (Accessed on 09/19/2018).

Kelly, D., J. Arguello, A. Edwards, and W.-c. Wu. (2015). "Development and Evaluation of Search Tasks for IIR Experiments Using a Cognitive Complexity Framework". In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval. ICTIR '15*. New York, NY, USA: ACM. 101–110. DOI: 10.1145/2808194.2809465. (Accessed on 10/22/2018).

Kim, N., M. Evens, J. A. Michael, and A. A. Rovick. (1989). "Circsim-tutor: An intelligent tutoring system for circulatory physiology". In: *Computer Assisted Learning*. Ed. by H. Maurer. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. 254–266. DOI: 10.1007/3-540-51142-3_64.

Kistner, S., K. Rakoczy, B. Otto, C. Dignath-van Ewijk, G. Büttner, and E. Klieme. (2010). "Promotion of self-regulated learning in classrooms: investigating frequency, quality, and consequences for student performance". *Metacognition and Learning*. 5(2): 157–171. DOI: 10.1007/s11409-010-9055-3. (Accessed on 02/03/2021).

Kiyak, Y. S. (2023). "A ChatGPT Prompt for Writing Case-Based Multiple-Choice Questions". *Revista Española de Educación Médica*. 4(3). DOI: 10.6018/edumed.587451. (Accessed on 08/02/2024).

Kiyak, Y. S., Ö. Coşkun, I. İ. Budakoğlu, and C. Uluoğlu. (2024). "ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam". *European Journal of Clinical Pharmacology*. 80(5): 729–735. DOI: 10.1007/s00228-024-03649-x. (Accessed on 08/02/2024).

Korkmaz, Ö., R. Çakir, and M. Y. Özden. (2017). "A validity and reliability study of the computational thinking scales (CTS)". *Computers in Human Behavior*. 72(July): 558–569. DOI: 10.1016/j.chb.2017.01.005. (Accessed on 07/30/2024).

Krathwohl, D. R. (2002). "A Revision of Bloom's Taxonomy: An Overview". *Theory Into Practice*. 41(4): 212–218. DOI: 10.1207/s15430421tip4104_2. (Accessed on 12/06/2018).

Kuhl, J. (1985). "From Cognition to Behavior: Perspectives for Future Research on Action Control". In: *Action Control: From Cognition to Behavior*. Ed. by J. Kuhl and J. Beckmann. *SSSP Springer Series in Social Psychology*. Berlin, Heidelberg: Springer. 267–275. DOI: 10.1007/978-3-642-69746-3_12. (Accessed on 02/04/2021).

Kuhlthau, C. C. (1994). "Students and the Information Search Process: Zones of Intervention for Librarians". In: *Advances in Librarianship*. Ed. by I. P. Godden. Vol. 18. *Advances in Librarianship*. Emerald Group Publishing Limited. 57–72. DOI: 10.1108/S0065-2830(1994)0000018004. (Accessed on 08/28/2020).

Latham, G. P. (2016). "Goal-Setting Theory: Causal Relationships, Mediators, and Moderators". In: *Oxford Research Encyclopedia of Psychology*. Oxford University Press. DOI: 10.1093/acrefore/9780190236557.013.12. (Accessed on 06/17/2024).

Latham, G. P. and T. C. Brown. (2006). "The Effect of Learning vs. Outcome Goals on Self-Efficacy, Satisfaction and Performance in an MBA Program". *Applied Psychology: An International Review*. 55(4): 606–623. DOI: 10.1111/j.1464-0597.2006.00246.x.

Latham, G. P. and E. A. Locke. (2007). "New Developments in and Directions for Goal-Setting Research". *European Psychologist*. 12(4): 290–300. DOI: 10.1027/1016-9040.12.4.290. (Accessed on 02/03/2021).

Latham, G. P. and G. H. Seijts. (1999). "The effects of proximal and distal goals on performance on a moderately complex task". *Journal of Organizational Behavior*. 20(4): 421–429. DOI: https://doi.org/10.1002/(SICI)1099-1379(199907)20:4<421::AID-JOB896>3.0.CO;2-#. (Accessed on 02/03/2021).

Lee, D., M. Brown, J. Hammond, and M. Zakowski. (2025). "Readability, quality and accuracy of generative artificial intelligence chatbots for commonly asked questions about labor epidurals: a comparison of ChatGPT and Bard". *International Journal of Obstetric Anesthesia*. 61(Feb.): 104317. DOI: 10.1016/j.ijoa.2024.104317. (Accessed on 01/24/2025).

Lee, H.-J., J. Lee, K. A. Makara, B. J. Fishman, and Y.-I. Hong. (2015). "Does higher education foster critical and creative learners? An exploration of two universities in South Korea and the USA". *Higher Education Research & Development*. 34(1): 131–146. DOI: 10.1080/07294360.2014.892477. (Accessed on 12/14/2020).

Lei, P.-L., C.-T. Sun, S. S. J. Lin, and T.-K. Huang. (2015). "Effect of metacognitive strategies and verbal-imagery cognitive style on biology-based video search and learning performance". *Computers & Education*. 87(Sept.): 326–339. DOI: 10.1016/j.compedu.2015.07.004. (Accessed on 02/22/2021).

LePine, J. A. (2005). "Adaptation of teams in response to unforeseen change: Effects of goal difficulty and team composition in terms of cognitive ability and goal orientation". *Journal of Applied Psychology*. 90(6): 1153–1167. DOI: 10.1037/0021-9010.90.6.1153. (Accessed on 02/17/2021).

Li, Y. and R. Capra. (2022). "Want or Need: Why Would Users Expect to Conduct Cross-Session Searches?" In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval. CHIIR '22*. Regensburg, Germany: Association for Computing Machinery. 327–331. DOI: 10.1145/3498366.3505829.

Li, Y., C. Liu, and P. Hansen. (2023). "Incubation and Verification Processes in Information Seeking: A Case Study in the Context of Autonomous Learning". In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. CHIIR '23*. New York, NY, USA: Association for Computing Machinery. 153–160. DOI: 10.1145/3576840.3578289. (Accessed on 06/26/2023).

Liu, C. and X. Song. (2018). "How do Information Source Selection Strategies Influence Users' Learning Outcomes". In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. CHIIR '18*. New York, NY, USA: Association for Computing Machinery. 257–260. DOI: 10.1145/3176349.3176876. (Accessed on 03/10/2021).

Liu, H., C. Liu, and N. J. Belkin. (2019). "Investigation of users' knowledge change process in learning-related search tasks". *Proceedings of the Association for Information Science and Technology*. 56(1): 166–175. DOI: https://doi.org/10.1002/pra2.63. (Accessed on 12/14/2020).

Liu, J. and N. J. Belkin. (2010). "Personalizing information retrieval for multi-session tasks: the roles of task stage and task type". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '10*. Geneva, Switzerland: Association for Computing Machinery. 26–33. DOI: 10.1145/1835449.1835457. (Accessed on 05/20/2020).

Liu, J., N. J. Belkin, X. Zhang, and X. Yuan. (2013). "Examining users' knowledge change in the task completion process". *Information Processing & Management*. 49(5): 1058–1074. DOI: 10.1016/j.ipm.2012.08.006. (Accessed on 05/15/2020).

Locke, E. A., D.-O. Chah, S. Harrison, and N. Lustgarten. (1989). "Separating the effects of goal specificity from goal level". *Organizational Behavior and Human Decision Processes*. 43(2): 270–287. DOI: 10.1016/0749-5978(89)90053-8. (Accessed on 04/24/2020).

Locke, E. A. (2001). "Self-set goals and self-efficacy as mediators of incentives and personality". In: *Work motivation in the context of a globalizing economy*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers. 13–26.

Locke, E. A. and G. P. Latham. (1990). *A theory of goal setting & task performance. A theory of goal setting & task performance*. Englewood Cliffs, NJ, US: Prentice-Hall, Inc.

Locke, E. A. and G. P. Latham. (2002). "Building a practically useful theory of goal setting and task motivation: A 35-year odyssey". *American Psychologist*. 57(9): 705–717. DOI: 10.1037/0003-066X.57.9.705. (Accessed on 04/24/2020).

Locke, E. A. and G. P. Latham. (2006). "New Directions in Goal-Setting Theory". *Current Directions in Psychological Science*. 15(5): 265–268. DOI: 10.1111/j.1467-8721.2006.00449.x. (Accessed on 01/21/2021).

Locke, E. A. and G. P. Latham. (2012). *New Developments in Goal Setting and Task Performance*. London, UNITED KINGDOM: Routledge. URL: http://ebookcentral.proquest.com/lib/unc/detail.action?docID=1104793 (accessed on 03/31/2020).

Locke, E. A. and G. P. Latham. (2019). "The development of goal setting theory: A half century retrospective". *Motivation Science*. 5(2): 93–105. DOI: 10.1037/mot0000127. (Accessed on 02/03/2021).

Lu, Y. and I.-H. Hsiao. (2017). "Personalized Information Seeking Assistant (PiSA): from programming information seeking to learning". *Information Retrieval Journal*. 20(5): 433–455. DOI: 10.1007/s10791-017-9305-y. (Accessed on 03/11/2021).

Mai, D. T. T., C. V. Da, and N. V. Hanh. (2024). "The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach". *Frontiers in Education*. 9(Feb.). DOI: 10.3389/feduc.2024.1328769. (Accessed on 07/30/2024).

Marchionini, G. (2006). "Exploratory search: from finding to understanding". *Communications of the ACM*. 49(4): 41. DOI: 10.1145/1121949.1121979. (Accessed on 09/26/2018).

Mariani, L. (1997). "Teacher support and teacher challenge in promoting learner autonomy". *Perspectives: A Journal of TESOL Italy*. XXIII (2). URL: http://www.learningpaths.org/papers/papersupport.htm (accessed on 09/25/2020).

McCardle, L., E. A. Webster, A. Haffey, and A. F. Hadwin. (2017). "Examining students' self-set goals for self-regulated learning: Goal properties and patterns". *Studies in Higher Education*. 42(11): 2153–2169. DOI: 10.1080/03075079.2015.1135117. (Accessed on 01/25/2021).

McNeil, N. M. and M. W. Alibali. (2000). "Learning mathematics from procedural instruction: Externally imposed goals influence what is learned". *Journal of Educational Psychology*. 92(4): 734–744. DOI: 10.1037/0022-0663.92.4.734. (Accessed on 05/01/2020).

McNeill, K. L., D. J. Lizotte, J. Krajcik, and R. W. Marx. (2006). "Supporting Students' Construction of Scientific Explanations by Fading Scaffolds in Instructional Materials". *Journal of the Learning Sciences*. 15(2): 153–191. DOI: 10.1207/s15327809jls1502_1. (Accessed on 02/01/2025).

Miyake, N. and P. A. Kirschner. (2014). "The Social and Interactive Dimensions of Collaborative Learning". In: *The Cambridge Handbook of the Learning Sciences*. Ed. by R. K. Sawyer. *Cambridge Handbooks in Psychology*. Cambridge University Press. 418–438.

Moraes, F., S. R. Putra, and C. Hauff. (2018). "Contrasting Search as a Learning Activity with Instructor-designed Learning". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18*. New York, NY, USA: Association for Computing Machinery. 167–176. DOI: 10.1145/3269206.3271676. (Accessed on 12/31/2020).

Nelson, L., C. Held, P. Pirolli, L. Hong, D. Schiano, and E. H. Chi. (2009). “With a little help from my friends: examining the impact of social annotations in sensemaking tasks”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery. 1795–1798. URL: https://doi.org/10.1145/1518701.1518977 (accessed on 07/27/2021).

Nersessian, N. J. (2002). “The cognitive basis of model-based reasoning in science”. In: *The cognitive basis of science*. New York, NY, US: Cambridge University Press. 133–153. DOI: 10.1017/CBO9780511613517.008.

O’Brien, H. L., A. Kampen, A. W. Cole, and K. Brennan. (2020). “The Role of Domain Knowledge in Search as Learning”. In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. CHIIR ’20*. Vancouver BC, Canada: Association for Computing Machinery. 313–317. DOI: 10.1145/3343413.3377989. (Accessed on 05/20/2020).

Otto, C., R. Yu, G. Pardi, J. von Hoyer, M. Rokicki, A. Hoppe, P. Holtz, Y. Kammerer, S. Dietze, and R. Ewerth. (2021). “Predicting Knowledge Gain During Web Search Based on Multimedia Resource Consumption”. In: *Artificial Intelligence in Education*. Ed. by I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova. *Lecture Notes in Computer Science*. Cham: Springer International Publishing. 318–330. DOI: 10.1007/978-3-030-78292-4_26.

Palani, S., Z. Ding, S. MacNeil, and S. P. Dow. (2021). “The "Active Search" Hypothesis: How Search Strategies Relate to Creative Learning”. In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. CHIIR ’21*. New York, NY, USA: Association for Computing Machinery. 325–329. DOI: 10.1145/3406522.3446046. (Accessed on 03/15/2021).

Pardi, G., S. Gottschling, P. Gerjets, and Y. Kammerer. (2023). “The moderating effect of knowledge type on search result modality preferences in web search scenarios”. *Computers and Education Open*. 4: 100126. DOI: https://doi.org/10.1016/j.caeo.2023.100126.

Pardi, G., J. von Hoyer, P. Holtz, and Y. 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*. Vancouver BC Canada: ACM. 378–382. DOI: 10.1145/3343413.3378001. (Accessed on 08/26/2020).

Pennycook, G., R. M. Ross, D. J. Koehler, and J. A. Fugelsang. (2017). "Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence". *Psychonomic Bulletin & Review*. 24(6): 1774–1784. DOI: 10.3758/s13423-017-1242-7. (Accessed on 02/16/2021).

Pérez-Rosas, V., B. Kleinberg, A. Lefevre, and R. Mihalcea. (2018). "Automatic Detection of Fake News". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by E. M. Bender, L. Derczynski, and P. Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics. 3391–3401. URL: https://aclanthology.org/C18-1287/.

Persky, A. M., E. Lee, and L. S. Schlesselman. (2020). "Perception of Learning Versus Performance as Outcome Measures of Educational Research". *American Journal of Pharmaceutical Education*. 84(7): ajpe7782. DOI: 10.5688/ajpe7782. (Accessed on 10/05/2020).

Piaget, J. and M. Cook. (1952). *The origins of intelligence in children*. Vol. 8. International Universities Press New York.

Pinto, G., I. Cardoso-Pereira, D. Monteiro, D. Lucena, A. Souza, and K. Gama. (2023). "Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT". In: *Proceedings of the XXXVII Brazilian Symposium on Software Engineering. SBES '23*. Campo Grande, Brazil: Association for Computing Machinery. 293–302. DOI: 10.1145/3613372.3614197.

Pintrich, P. R. (2000). "Chapter 14 - The Role of Goal Orientation in Self-Regulated Learning". In: *Handbook of Self-Regulation*. Ed. by M. Boekaerts, P. R. Pintrich, and M. Zeidner. San Diego: Academic Press. 451–502. DOI: 10.1016/B978-012109890-2/50043-3. (Accessed on 04/23/2020).

Pintrich, P. R. and E. V. de Groot. (1990). “Motivational and self-regulated learning components of classroom academic performance”. *Journal of Educational Psychology*. 82(1): 33–40. DOI: 10.1037/0022-0663.82.1.33.

Puntambekar, S. and R. Hübscher. (2005). “Tools for Scaffolding Students in a Complex Learning Environment: What Have We Gained and What Have We Missed?: Educational Psychologist: Vol 40, No 1”. *Educational Psychologist*. URL: https://www-tandfonline-com.libproxy.lib.unc.edu/doi/abs/10.1207/s15326985ep4001_1?casa_token=vQ2z4sXRAL4AAAAA:QradSMPeb8oeMa3ikjX6DmI7ak8qsDAOpW0FxSBsS7b1Bfk9hjUmqPtfs6Iw7O7gr0CGW3iYwH5T& (accessed on 10/19/2020).

Puustinen, M. and L. Pulkkinen. (2001). “Models of Self-regulated Learning: A review”. *Scandinavian Journal of Educational Research*. 45(3): 269–286. DOI: 10.1080/00313830120074206. (Accessed on 02/04/2021).

Qiu, S., U. Gadiraju, and A. Bozzon. (2020). “Towards Memorable Information Retrieval”. In: *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. ICTIR ’20*. New York, NY, USA: Association for Computing Machinery. 69–76. DOI: 10.1145/3409256.3409830. (Accessed on 01/05/2021).

Rajpurkar, P., R. Jia, and P. Liang. (2018). “Know What You Don’t Know: Unanswerable Questions for SQuAD”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne, Australia: Association for Computational Linguistics. 784–789. DOI: 10.18653/v1/P18-2124.

Rieger, A., T. Draws, N. Mattis, D. Maxwell, D. Elsweiler, U. Gadiraju, D. McKay, A. Bozzon, and M. S. Pera. (2024). “Responsible Opinion Formation onÂ Debated Topics inÂ Web Search”. In: *Advances in Information Retrieval*. Ed. by N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, and I. Ounis. Cham: Springer Nature Switzerland. 437–465.

Rieh, S. Y., K. Collins-Thompson, P. Hansen, and H.-J. Lee. (2016). "Towards searching as a learning process: A review of current perspectives and future directions". *Journal of Information Science*. 42(1): 19–34. DOI: 10.1177/0165551515615841. (Accessed on 11/24/2018).

Rieh, S. Y., J. Gwizdka, L. Freund, and K. Collins-Thompson. (2014). "Searching as learning: Novel measures for information interaction research". *Proceedings of the American Society for Information Science and Technology*. 51(1): 1–4. DOI: 10.1002/meet.2014.14505101021. (Accessed on 11/24/2018).

Roegiest, A. and Z. Pinkosova. (2024). "Generative Information Systems Are Great If You Can Read". In: *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval. CHIIR '24*. New York, NY, USA: Association for Computing Machinery. 165–177. DOI: 10.1145/3627508.3638345. (Accessed on 01/24/2025).

Rokicki, M., R. Yu, and D. Hienert. (2022). "Learning to Rank for Knowledge Gain".

Roy, N., F. Moraes, and C. Hauff. (2020). "Exploring Users' Learning Gains within Search Sessions". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. CHIIR '20*. Vancouver BC, Canada: Association for Computing Machinery. 432–436. DOI: 10.1145/3343413.3378012. (Accessed on 05/20/2020).

Roy, N., M. V. Torre, U. Gadiraju, D. Maxwell, and C. 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. CHIIR '21*. New York, NY, USA: Association for Computing Machinery. 229–238. DOI: 10.1145/3406522.3446025. (Accessed on 03/15/2021).

Salimzadeh, S., D. Maxwell, and C. Hauff. (2021). "On the Impact of Entity Cards on Learning-Oriented Search Tasks". In: *Proceedings of the 2021 ACM SIGIR on International Conference on Theory of Information Retrieval*. ACM. 10.

Salmerón, L., P. Delgado, and L. Mason. (2020). "Using eye-movement modelling examples to improve critical reading of multiple webpages on a conflicting topic". *Journal of Computer Assisted Learning*. 36(6): 1038–1051. DOI: https://doi.org/10.1111/jcal.12458. (Accessed on 02/22/2021).

Santhanam, R., S. Sasidharan, and J. Webster. (2008). "Using Self-Regulatory Learning to Enhance E-Learning-Based Information Technology Training". *Information Systems Research*. 19(1): 26–47. DOI: 10.1287/isre.1070.0141. (Accessed on 02/03/2021).

Sawyer, R. K. (2014). "The Cambridge Handbook of the Learning Sciences edited by R. Keith Sawyer". DOI: 10.1017/CBO9781139519526. (Accessed on 12/06/2018).

Schraw, G. and R. S. Dennison. (1994). "Assessing Metacognitive Awareness". *Contemporary Educational Psychology*. 19(4): 460–475. DOI: 10.1006/ceps.1994.1033. (Accessed on 01/24/2022).

Schunk, D. (1991). "Self-Efficacy and Academic Motivation." *Educational Psychologist*. 26(3 & 4): 207–231.

Schunk, D. H. (2001). "Self-Regulation Through Goal Setting". *ERIC Digest*: 2.

Schunk, D. H. (1981). "Modeling and attributional effects on children's achievement: A self-efficacy analysis". *Journal of Educational Psychology*. 73(1): 93–105. DOI: 10.1037/0022-0663.73.1.93.

Schunk, D. H. (1984). "Sequential attributional feedback and children's achievement behaviors". *Journal of Educational Psychology*. 76(6): 1159–1169. DOI: 10.1037/0022-0663.76.6.1159.

Schunk, D. H. (1996). "Goal and Self-Evaluative Influences During Children's Cognitive Skill Learning". *American Educational Research Journal*. 33(2): 359–382. DOI: 10.3102/00028312033002359. (Accessed on 05/01/2020).

Schunk, D. H. and C. W. Swartz. (1993). "Goals and Progress Feedback: Effects on Self-Efficacy and Writing Achievement". *Contemporary Educational Psychology*. 18(3): 337–354. DOI: 10.1006/ceps.1993.1024. (Accessed on 02/04/2021).

Seijts, G. H. and G. P. Latham. (2001). "The effect of distal learning, outcome, and proximal goals on a moderately complex task". *Journal of Organizational Behavior*. 22(3): 291–307. DOI: 10.1002/job.70. (Accessed on 06/17/2020).

Shah, C. (2012). *Collaborative information seeking: The art and science of making the whole greater than the sum of all*. Vol. 34. Springer Science & Business Media.

Shapiro, A. M. (2004). “How including Prior Knowledge as a Subject Variable May Change Outcomes of Learning Research”. *American Educational Research Journal*. 41(1): 159–189. URL: http://www.jstor.org/stable/3699387.

Sharma, P. and M. J. Hannafin. (2007). “Scaffolding in technology-enhanced learning environments”. *Interactive Learning Environments*. 15(1): 27–46. DOI: 10.1080/10494820600996972. (Accessed on 10/28/2020).

Sitzmann, T. and K. Ely. (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. DOI: 10.1037/a0022777.

Smith, C., K. Urgo, J. Arguello, and R. Capra. (2022). “Learner, Assignment, and Domain: Contextualizing Search for Comprehension”. In: *ACM SIGIR Conference on Human Information Interaction and Retrieval*. Regensburg Germany: ACM. 191–201. DOI: 10.1145/3498366.3505819. (Accessed on 03/15/2022).

Smith, S. G. and B. A. Sherwood. (1976). “Educational Uses of the PLATO Computer System”. *Science*. 192(4237): 344–352. URL: http://www.jstor.org/stable/1742096 (accessed on 01/18/2021).

Sobocinski, M., S. Järvelä, J. Malmberg, M. Dindar, A. Isosalo, and K. Noponen. (2020). “How does monitoring set the stage for adaptive regulation or maladaptive behavior in collaborative learning?” *Metacognition and Learning*. 15(2): 99–127. DOI: 10.1007/s11409-020-09224-w. (Accessed on 02/16/2021).

Soprano, M., K. Roitero, D. La Barbera, D. Ceolin, D. Spina, G. Demartini, and S. Mizzaro. (2024). “Cognitive Biases in Fact-Checking and Their Countermeasures: A Review”. *Information Processing & Management*. 61(3): 103672. DOI: 10.1016/j.ipm.2024.103672. (Accessed on 01/09/2025).

Sousa, D. A. (2017). *How the Brain Learns*. Fifth Edition. Corwin Press.

Stahl, G., T. Koschmann, and D. D. Suthers. (2005). “Computer-Supported Collaborative Learning”. In: *The Cambridge Handbook of the Learning Sciences*. Ed. by R. K. Sawyer. *Cambridge Handbooks in Psychology*. Cambridge University Press. 409–426.

Syed, R. and K. Collins-Thompson. (2017a). "Optimizing search results for human learning goals". *Information Retrieval Journal*. 20(5): 506–523. DOI: 10.1007/s10791-017-9303-0. (Accessed on 11/24/2018).

Syed, R. and K. Collins-Thompson. (2017b). "Retrieval Algorithms Optimized for Human Learning". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17*. New York, NY, USA: ACM. 555–564. DOI: 10.1145/3077136.3080835. (Accessed on 02/12/2019).

Syed, R. and K. Collins-Thompson. (2018). "Exploring Document Retrieval Features Associated with Improved Short- and Long-term Vocabulary Learning Outcomes". In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval. CHIIR '18*. New Brunswick, NJ, USA: Association for Computing Machinery. 191–200. DOI: 10.1145/3176349.3176397. (Accessed on 08/02/2020).

Syed, R., K. Collins-Thompson, P. N. Bennett, M. Teng, S. Williams, D. W. W. Tay, and S. Iqbal. (2020). "Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation". In: *Proceedings of The Web Conference 2020. WWW '20*. New York, NY, USA: Association for Computing Machinery. 1693–1703. DOI: 10.1145/3366423.3380240. (Accessed on 08/11/2020).

Talja, S., K. Tuominen, and R. Savolainen. (2005). ""Isms" in information science: constructivism, collectivism and constructionism". *Journal of Documentation*. 61(1): 79–101. Ed. by B. Hjørland. DOI: 10.1108/00220410510578023. (Accessed on 08/27/2020).

Tintarev, N., E. Sullivan, D. Guldin, S. Qiu, and D. Odjik. (2018). "Same, Same, but Different: Algorithmic Diversification of Viewpoints in News". In: *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization. UMAP '18*. Singapore, Singapore: Association for Computing Machinery. 7–13. DOI: 10.1145/3213586.3226203.

Trevors, G., M. Duffy, and R. Azevedo. (2014). "Note-taking within MetaTutor: interactions between an intelligent tutoring system and prior knowledge on note-taking and learning". *Educational Technology Research and Development*. 62(5): 507–528. DOI: 10.1007/s11423-014-9343-8. (Accessed on 08/29/2020).

Tulving, E. (1962). “Subjective organization in free recall of "unrelated" words”. *Psychological Review*. 69(4): 344–354. DOI: 10.1037/h0043150.

Tuttle, H. S. (1955). “Ambiguous Is the Word for “Transfer””. *The Educational Forum*. 19(2): 159–164. DOI: 10.1080/00131725509341781. (Accessed on 07/16/2024).

Urgo, K. (2023). “Investigating the Influence of Subgoals on Learning During Search”. *Ph.D.* United States – North Carolina: The University of North Carolina at Chapel Hill. URL: https://www.proquest.com/docview/2854268320/abstract/9393604F5EFD4DB5PQ/1 (accessed on 09/22/2023).

Urgo, K. and J. Arguello. (2022a). “Capturing Self-Regulated Learning During Search”. In: *Proceedings of the Third International Workshop on Investigating Learning During Web Search (IWILDS‘22) co-located with the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’22)*. Madrid, Spain.

Urgo, K. and J. Arguello. (2022b). “Learning assessments in search-as-learning: A survey of prior work and opportunities for future research”. *Information Processing & Management*. 59(2): 102821. DOI: 10.1016/j.ipm.2021.102821. (Accessed on 01/12/2022).

Urgo, K. and J. Arguello. (2022c). “Understanding the "Pathway" Towards a Searcher’s Learning Objective”. *ACM Transactions on Information Systems*. 40(4): 77:1–77:42. DOI: 10.1145/3495222. (Accessed on 01/12/2022).

Urgo, K. and J. Arguello. (2023). “Goal-setting in support of learning during search: An exploration of learning outcomes and searcher perceptions”. *Information Processing & Management*. 60(2): 103158. DOI: 10.1016/j.ipm.2022.103158. (Accessed on 11/22/2022).

Urgo, K. and J. 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*. Sheffield United Kingdom: ACM. 278–290. DOI: 10.1145/3627508.3638348.

Urgo, K., J. Arguello, and R. Capra. (2019). "Anderson and Krathwohl's Two-Dimensional Taxonomy Applied to Task Creation and Learning Assessment". In: *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval - ICTIR '19*. Santa Clara, CA, USA: ACM Press. 117–124. DOI: 10.1145/3341981.3344226. (Accessed on 11/02/2019).

Urgo, K., J. Arguello, and R. Capra. (2020). "The Effects of Learning Objectives on Searchers' Perceptions and Behaviors". In: *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. Virtual Event Norway: ACM. 77–84. DOI: 10.1145/3409256.3409815. (Accessed on 10/29/2020).

Valle, A., R. G. Cabanach, J. C. Núnez, J. González-Pienda, S. Rodríguez, and I. Piñeiro. (2003). "Multiple goals, motivation and academic learning". *British Journal of Educational Psychology*. 73(1): 71–87. DOI: https://doi.org/10.1348/000709903762869923. (Accessed on 02/01/2021).

Vandewaetere, M., P. Desmet, and G. Clarebout. (2011). "The contribution of learner characteristics in the development of computer-based adaptive learning environments". *Computers in Human Behavior*. Current Research Topics in Cognitive Load Theory 27(1): 118–130. DOI: 10.1016/j.chb.2010.07.038. (Accessed on 07/19/2024).

von Hoyer, J., A. Hoppe, Y. Kammerer, C. Otto, G. Pardi, M. Rokicki, R. Yu, S. Dietze, R. Ewerth, and P. Holtz. (2022b). "The Search as Learning Spaceship: Toward a Comprehensive Model of Psychological and Technological Facets of Search as Learning". *Frontiers in Psychology*. 13. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2022.827748 (accessed on 03/16/2022).

Vygotsky, L. S. (1980). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Vygotsky, L. S. and A. Kozulin. (1962). *Thought and Language*. MIT Press.

Weingart, N. and C. Eickhoff. (2016). "Retrieval Techniques for Contextual Learning". *SAL @ SIGIR*: 5.

Weinstein, C., D. Palmer, and A. Schulte. (1987). "Learning and Study Strategies Inventory (LASSI)". *Clearwater, FL: H & H Publishing*.

White, R. W., S. T. Dumais, and J. Teevan. (2009). "Characterizing the Influence of Domain Expertise on Web Search Behavior". In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining. WSDM '09*. New York, NY, USA: ACM. 132–141. DOI: 10.1145/1498759.1498819. (Accessed on 12/07/2018).

White, S. S. and E. A. Locke. (2000). "Problems with the pygmalion effect and some proposed solutions". *The Leadership Quarterly*. 11(3): 389–415. DOI: 10.1016/S1048-9843(00)00046-1. (Accessed on 04/24/2020).

Willoughby, T., S. A. Anderson, E. Wood, J. Mueller, and C. Ross. (2009). "Fast searching for information on the Internet to use in a learning context: The impact of domain knowledge". *Computers & Education*. 52(3): 640–648. DOI: 10.1016/j.compedu.2008.11.009. (Accessed on 03/10/2021).

Wilson, M. J. and M. L. Wilson. (2013). "A comparison of techniques for measuring sensemaking and learning within participant-generated summaries". *Journal of the American Society for Information Science and Technology*. 64(2): 291–306. DOI: 10.1002/asi.22758. (Accessed on 09/06/2020).

Wilson, M. L., P. André, and m. schraefel mc. (2008). "Backward highlighting: enhancing faceted search". In: *Proceedings of the 21st annual ACM symposium on User interface software and technology. UIST '08*. New York, NY, USA: Association for Computing Machinery. 235–238. DOI: 10.1145/1449715.1449754. (Accessed on 09/06/2020).

Winne, P. H. (2001). "Self-regulated learning viewed from models of information processing". In: *Self-regulated learning and academic achievement: Theoretical perspectives, 2nd ed.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers. 153–189.

Winne, P. H. (2022). "Learning Analytics for Self-Regulated Learning". In: *The Handbook of Learning Analytics*. Ed. by C. Lang, G. Siemens, and A. F. Wise. 2nd ed. SOLAR. 78–85. DOI: 10.18608/hla22.008. (Accessed on 06/07/2024).

Winne, P. H. and A. F. Hadwin. (1998). "Studying as self-regulated engagement in learning". In: *Metacognition in educational theory and practice*.

Winne, P. H., D. Jamieson-Noel, and K. Muis. (2002). "Methodological issues and advances in researching tactics, strategies, and self-regulated learning." In: *New directions in measures and methods*. Ed. by P. R. Pintrich and M. L. Maehr. 1. ed. *Advances in motivation and achievement*. No. 12. Amsterdam: JAI, An Imprint of Elsevier Science.

Winne, P. H. and N. E. Perry. (2000). "Measuring self-regulated learning". In: *Handbook of self-regulation*. San Diego, CA, US: Academic Press. 531–566. DOI: 10.1016/B978-012109890-2/50045-7.

Winters, D. and G. P. Latham. (1996). "The effect of learning versus outcome goals on a simple versus a complex task". *Group & Organization Management*. 21(2): 236–250. DOI: 10.1177/1059601196212007.

Woo, C. W., M. W. Evens, R. Freedman, M. Glass, L. S. Shim, Y. Zhang, Y. Zhou, and J. Michael. (2006). "An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning". *Artificial Intelligence in Medicine*. Intelligent Medical Training Systems 38(1): 25–46. DOI: 10.1016/j.artmed.2005.10.004. (Accessed on 01/18/2021).

Woodworth, R. S. and E. L. Thorndike. (1901). "The influence of improvement in one mental function upon the efficiency of other functions. (I)". *Psychological Review*. 8(3): 247–261. DOI: 10.1037/h0074898.

Wu, W.-C., D. Kelly, and A. Sud. (2014). "Using information scent and need for cognition to understand online search behavior". In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. *SIGIR '14*. Gold Coast, Queensland, Australia: Association for Computing Machinery. 557–566. DOI: 10.1145/2600428.2609626.

Xu, L., X. Zhou, and U. Gadiraju. (2020). "How Does Team Composition Affect Knowledge Gain of Users in Collaborative Web Search?" In: *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. *HT '20*. New York, NY, USA: Association for Computing Machinery. 91–100. DOI: 10.1145/3372923.3404784. (Accessed on 09/07/2020).

Yilmaz, R. and F. G. Karaoglan Yilmaz. (2023). "The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation". *Computers and Education: Artificial Intelligence*. 4(Jan.): 100147. DOI: 10.1016/j.caeai.2023.100147. (Accessed on 07/30/2024).

Yu, R., U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, and S. Dietze. (2018). "Predicting User Knowledge Gain in Informational Search Sessions". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18*. New York, NY, USA: ACM. 75–84. DOI: 10.1145/3209978.3210064. (Accessed on 11/29/2018).

Zhang, X., J. Liu, M. Cole, and N. Belkin. (2015). "Predicting users' domain knowledge in information retrieval using multiple regression analysis of search behaviors". *Journal of the Association for Information Science and Technology*. 66(5): 980–1000. DOI: https://doi.org/10.1002/asi.23218. (Accessed on 01/21/2021).

Zhang, X. and C. Liu. (2023). "Examination of Information Problem Decomposition Strategies: A New Perspective for Understanding Users' Information Problems in Search as Learning". In: *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. SIGIR-AP '23*. New York, NY, USA: Association for Computing Machinery. 84–94. DOI: 10.1145/3624918.3625326. (Accessed on 11/27/2023).

Zhang, Y. and C. Liu. (2020). "Users' Knowledge Use and Change during Information Searching Process: A Perspective of Vocabulary Usage". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. JCDL '20*. New York, NY, USA: Association for Computing Machinery. 47–56. DOI: 10.1145/3383583.3398532. (Accessed on 09/07/2020).

Zheng, L. (2016). "The effectiveness of self-regulated learning scaffolds on academic performance in computer-based learning environments: a meta-analysis". *Asia Pacific Education Review*. 17(2): 187–202. DOI: 10.1007/s12564-016-9426-9. (Accessed on 08/08/2024).

Zimmerman, B. (2008). "Goal Setting: A Key Proactive Source of Academic Self-Regulation". In: *Motivation and Self-Regulated Learning: Theory, Research, and Applications*. Vol. 267.

Zimmerman, B. J. (2000). "Chapter 2 - Attaining Self-Regulation: A Social Cognitive Perspective". In: *Handbook of Self-Regulation*. Ed. by M. Boekaerts, P. R. Pintrich, and M. Zeidner. San Diego: Academic Press. 13–39. DOI: 10.1016/B978-012109890-2/50031-7. (Accessed on 04/23/2020).

Zimmerman, B. J. (2002). "Becoming a Self-Regulated Learner: An Overview". *Theory Into Practice*. 41(2): 64–70. DOI: 10.1207/s15430421tip4102_2. (Accessed on 03/29/2020).

Zimmerman, B. J. and M. Martinez-Pons. (1988). "Construct validation of a strategy model of student self-regulated learning". *Journal of Educational Psychology*. 80(3): 284–290. DOI: 10.1037/0022-0663.80.3.284.

Zimmerman, B. J. and M. M. Pons. (1986). "Development of a Structured Interview for Assessing Student Use of Self-Regulated Learning Strategies". *American Educational Research Journal*. 23(4): 614–628. DOI: 10.3102/00028312023004614. (Accessed on 02/04/2021).

Zimmerman, B. J. and D. H. Schunk. (2011). *Handbook of self-regulation of learning and performance. Handbook of self-regulation of learning and performance*. New York, NY, US: Routledge/Taylor & Francis Group.