Information Processing and Management 59 (2022) 102821

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Learning assessments in search-as-learning: A survey of prior work and opportunities for future research

Kelsey Urgo, Jaime Arguello *

School of Information and Library Science, University of North Carolina at Chapel Hill, 216 Lenoir Drive, CB #3360, 100 Manning Hall, Chapel Hill, NC 27599-3360, United States of America

ARTICLE INFO

Keywords:
Search as learning
Learning assessment
Interactive information retrieval

ABSTRACT

People often search for information in order to learn something new. In recent years, the "search-as-learning" movement has argued that search systems should be better designed to support learning. Current search systems (especially Web search engines) are largely designed and optimized to fulfill simple look-up tasks (e.g., navigational or fact-finding search tasks). However, they provide less support for searchers working on complex tasks that involve learning. Search-as-learning studies have investigated a wide range of research questions. For example, studies have aimed to better understand how characteristics of the individual searcher, the type of search task, and interactive features provided by the system can influence learning outcomes. Learning assessment is a key component in search-as-learning studies. Assessment materials are used to both gauge prior knowledge and measure learning during or after one or more search sessions. In this paper, we provide a systematic review of different types of assessments used in search-as-learning studies to date. The paper makes the following three contributions. First, we review different types of assessments used and discuss their potential benefits and drawbacks. Second, we review assessments used outside of search-as-learning, which may provide insights and opportunities for future research. Third, we provide recommendations for future research. Importantly, we argue that future studies should clearly define learning objectives and develop assessment materials that reliably capture the intended type of learning. For example, assessment materials should test a participant's ability to engage with specific cognitive processes, which may range from simple (e.g., memorization) to more complex (e.g., critical and creative thinking). Additionally, we argue that future studies should consider two dimensions that are understudied in search-as-learning: 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).

1. Introduction

While current search systems are effective in helping users complete simple look-up tasks (e.g., navigational or fact-finding tasks), they provide less support for searchers working on complex tasks that involve learning. In recent years, the "search-as-learning" research community has argued that learning is an important outcome of search. In this respect, search systems should be designed and evaluated as tools to support learning.

Within the last decade, several summits have taken place to develop research agendas at the intersection of interactive information retrieval (IIR) and learning sciences. Participants at SWIRL 2012 advocated that future research should aim to: (1)

* Corresponding author.
E-mail addresses: kurgo@unc.edu (K. Urgo), jarguell@unc.edu (J. Arguello).

https://doi.org/10.1016/j.ipm.2021.102821
Received 26 April 2021; Received in revised form 4 November 2021; Accepted 8 November 2021
Available online 11 January 2022

understand the cognitive biases fostered by existing search systems, (2) help searchers be more critical consumers of information, and (3) develop search tools to better support learning (Allan, Croft, Moffat, & Sanderson, 2012). Similarly, participants at the Dagstuhl Seminar on Search as Learning proposed that future research should aim to: (1) understand the contexts in which people search to learn; (2) investigate search as a learning process; (3) develop search system tools and interventions to promote learning; and (4) explore and expand upon methods to measure learning during search (Collins-Thompson, Hansen, & Hauff, 2017).

An important question is: What makes a study a "search-as-learning" study? Generally speaking, a search-as-learning study is designed to better understand learning during search. The common thread is that learning is an important component of the study. In many cases, this means that learning objectives and/or outcomes are important experimental variables in the study. In other cases, it means that an important goal of the study is to understand the learning *process* during search. Studies in the area of search-as-learning have investigated a wide range of research questions. Many studies have investigated how different factors can influence learning during search. Specifically, studies have investigated characteristics of the individual searcher (O'Brien, Kampen, Cole, & Brennan, 2020; Pardi, von Hoyer, Holtz, & Kammerer, 2020; Roy, Moraes, & Hauff, 2020; Willoughby, Anderson, Wood, Mueller, & Ross, 2009), characteristics of the searcher's learning objective (Ghosh, Rath, & Shah, 2018; Kalyani & Gadiraju, 2019; Liu, Belkin, Zhang, & Yuan, 2013; Liu, Liu, & Belkin, 2019), and characteristics of the search system (Câmara, Roy, Maxwell, & Hauff, 2021; Chi, Han, He, & Meng, 2016; Demaree, Jarodzka, Brand-Gruwel, & Kammerer, 2020; Freund, Kopak, & O'Brien, 2016; Heilman, Collins-thompson, Callan, & Eskenazi, 2010; Heilman & Eskenazi, 2006; Hersh, Elliot, Hickam, Wolf, & Molnar, 1995; Kammerer, Nairn, Pirolli, & Chi, 2009; Palani, Ding, MacNeil, & Dow, 2021; Qiu, Gadiraju, & Bozzon, 2020; Roy, Torre, Gadiraju, Maxwell, & Hauff, 2021; Syed & Collins-Thompson, 2017; Weingart & Eickhoff, 2016; Wilson, André, & schraefel, 2008; Xu, Zhou, & Gadiraju, 2020). Additionally, studies have investigated the relation between specific behaviors and learning outcomes (Abualsaud, 2017; Bhattacharya & Gwizdka, 2019; Chi et al., 2016; Collins-Thompson, Rieh, Haynes, & Syed, 2016; Gadiraju, Yu, Dietze, & Holtz, 2018; Lei, Sun, Lin, & Huang, 2015; Liu & Song, 2018; Lu & Hsiao, 2017; Palani et al., 2021; Xu et al., 2020; Yu et al., 2018).

Learning assessment is a critical component in search-as-learning studies. For example, in order to study how specific factors can influence learning during search, it is necessary to *measure* learning. As it turns out, studies have used a wide range of methods to measure learning during search. For example, some studies have simply focused on self-reported *perceptions* of learning (Capra, Arguello, O'Brien, Li, & Choi, 2018). Other studies have measured learning more *objectively*, by administering tests before and after the search session. Studies have used closed-ended assessments with predefined correct answers, such as multiple-choice (Gadiraju et al., 2018) and short-answer tests (Hersh et al., 1995). Other studies have used more open-ended methods, for example, by asking participants to summarize what they learned during the search process (Collins-Thompson et al., 2016). To measure learning from open-ended responses, research has proposed different dimensions to consider when manually grading open-ended responses (Wilson & Wilson, 2013).

In this paper, we present a systematic review of different types of learning assessments used in prior search-as-learning studies. We identified nine different types of assessments. Importantly, we discuss some of the benefits and drawbacks of each type of assessment. For example, in terms of benefits, multiple-choice tests have predefined correct answers and are therefore easy to grade. Learning can be measured as the difference between pre- and post-test scores, and these differences can be easily compared across participants in different experimental conditions (e.g., participants exposed to different tools). However, in terms of drawbacks, multiple-choice tests ask specific questions that may not capture *everything* a participant has learned during the search process. Open-ended assessments (e.g., knowledge summaries) provide searchers with the opportunity to convey everything they learned. However, grading requires more effort. Open-ended responses are usually graded along different subjective criteria (e.g., the presence of evaluative statements that convey the learner's ability to think critically about the subject). The grading of open-ended responses requires clearly defining assessment criteria that yield sufficiently high levels of intercoder agreement. A low agreement suggests that the grading criteria are vaguely defined and therefore unreliable.

The paper is organized as follows. Section 2 reviews relevant background. Importantly, we introduce Anderson and Krathwohl's taxonomy of learning (Anderson et al., 2001). In the field of education, the A&K taxonomy was developed to precisely define learning objectives for students. Additionally, it was developed to ensure that instructional and assessment materials are aligned with the objective(s). Studies in search-as-learning have leveraged this taxonomy to systematically manipulate learning objectives for study participants and develop learning assessments to measure whether specific types of learning took place. Furthermore, the A&K taxonomy is a useful framework for understanding the caveats and limitations of specific types of assessments. For example, argumentative essays ask participants to enumerate pro and con arguments about a given stance or proposition (e.g., climate change is primarily caused by human activities). This type of assessment reliably measures a participant's ability to *recall* arguments encountered during the search process. However, it does not reliably measure a participant's ability to *critique* the validity or importance of different arguments. After introducing the A&K taxonomy, Section 2 reviews some of the research questions and results from previous search-as-learning studies.

Section 3 describes our methodology for identifying relevant papers to include in our review. As an important contribution of our work, we discuss common and uncommon types of assessments used in prior work. In Section 3, we also describe how learning assessments were manually categorized.

Section 4 reviews the different types of assessments used in search-as-learning studies. We discuss how participants' pre- and post-test responses were graded and compared in order to measure learning during a search session. Additionally, we discuss benefits and drawbacks from different assessment types. Learning assessment also plays a critical role in studies in education and psychology. Therefore, in Section 5, we review assessment techniques that have *not* yet been considered in search-as-learning studies.

Selecting a type of assessment and developing assessment materials may seem like daunting tasks. In Section 6, we propose recommendations for future work that may help to alleviate these challenges. First, we discuss the importance of clearly defining

the learning objective of a search task. What exactly is a participant being asked to learn as they search? Having precisely defined objectives can help researchers develop assessment items that measure specific types of learning. Learning is inherently multidimensional. Therefore, we argue that researchers should consider four dimensions when developing assessment items. Leveraging the A&K taxonomy, we argue that assessment items should focus on a specific type of knowledge and cognitive process. Additionally, we argue that future studies should consider two additional dimensions: (1) *retention* (i.e., being able to retrieve what was learned from long-term memory) and (2) *transfer of learning* (i.e., being able to use what was learned in a new context or to support *new* learning-oriented activities).

In Section 6, we also discuss practical considerations that may help a researcher choose a type of assessment. For example, perhaps a study aims to investigate whether self-assessment during the search process improves learning. If timely feedback is important, then multiple-choice, short-answer, and free-recall assessments are good candidates that can be graded quickly and even automatically. Finally, we discuss strategies for mitigating some of the drawbacks of certain types of assessments.

2. Background and related work

When people search to learn, they typically have a specific learning objective in mind—"I need to find information that enables me to do <learning objective>". Learning objectives can vary along different dimensions. In Section 2.1, we review the Anderson & Krathwohl (A&K) taxonomy of learning (Anderson et al., 2001).[1] In the field of education, the A&K taxonomy was developed to help educators *precisely* define learning objectives for students. In other words, what exactly is the student expected to learn as the result of an instructional exercise? Additionally, it was developed to help educators align the learning *assessment* materials with the target *objective*. Simply put, the taxonomy was developed to help educators avoid ambiguity when: (1) setting goals for students and (2) assessing whether the goals were successfully met.

Characterizing learning objectives is an important part of search-as-learning research. Several studies in search-as-learning have leveraged the A&K taxonomy to systematically manipulate learning-oriented search tasks and develop learning assessment materials (Collins-Thompson et al., 2016; Kalyani & Gadiraju, 2019; Wilson & Wilson, 2013). We believe this trend is likely to continue. Additionally, the A&K taxonomy is a useful framework for understanding the *types* of learning that can be reliably assessed by different methods. For these reasons, we review the A&K taxonomy early in the paper.

Following our review of the A&K taxonomy, in Section 2.2, we review prior work in the area of search-as-learning. First, we discuss some of the early summits, workshops, and publications that called for future research in search-as-learning. Then, we review prior studies in search-as-learning. Studies have investigated a wide range of research questions. Our goal is to provide an informative foundation on what we know so far about learning during search. For example, what are factors that impact learning during search? How do specific factors impact learning? Are there certain behaviors that correlate with learning outcomes? Naturally, learning assessment has played a critical role in all these studies. However, in Section 2.2, we primarily focus on research questions and findings with respect to learning outcomes. Later, in Section 4, we provide a comprehensive review of learning assessments used in all search-as-learning studies to date.

2.1. The Anderson and Krathwohl (A&K) taxonomy

The A&K taxonomy was developed to help educators: (1) precisely define *learning objectives* for students, (2) develop *instructional exercises* that directly teach to the objectives, and (3) develop *assessment materials* that reliably measure whether the objectives were successfully met (Anderson et al., 2001). The A&K taxonomy was developed in 2001 as a revision of Bloom's taxonomy (Bloom, 1956), a taxonomy developed in 1956 that continues to be relevant after more than six decades.[2] The A&K taxonomy situates learning objectives at the intersection of two orthogonal dimensions: (1) the cognitive process dimension and (2) the knowledge type dimension (Table 1).

The *cognitive process* dimension defines the types of mental activities associated to the learning objective. In other words, it defines the types of cognitive processes the learner will be able to successfully perform once the objective is met. The A&K taxonomy defines six cognitive processes. A *remember* objective involves rote memorization—being able to regurgitate information verbatim. An *understand* objective involves being able to summarize and exemplify. An *apply* objective involves using knowledge to perform a task. An *analyze* objective involves understanding the similarities, differences, and/or relations between elements. An *evaluate* objective involves critiquing, judging, evaluating, and/or prioritizing elements or alternatives. Finally, a *create* objective involves inventing a novel solution to a problem or a novel representation of knowledge. The cognitive process dimension runs along a continuum from simple (remember) to complex (create). Within this dimension, there is an important relationship between understand, analyze, and evaluate. Analyzing is often an extension to understanding and a prelude to evaluating (Anderson et al., 2001, p. 79). To

[1] Although other taxonomies of learning exist (e.g., (Biggs & Collis, 1982; Fink, 2013)), the A&K taxonomy is arguably the most established and enduring. The A&K taxonomy and its foundational taxonomy, Bloom's taxonomy (developed in 1956), together have over 75,000 citations. Additionally, the A&K taxonomy has been used across a large number of search studies to both characterize learning-oriented search tasks and/or to assess learning after search (e.g., Collins-Thompson et al. (2016), Jansen, Booth, and Smith (2009), Kalyani and Gadiraju (2019), Wilson and Wilson (2013), Wu, Kelly, Edwards, and Arguello (2012)).

[2] The A&K taxonomy differs from Bloom's taxonomy in three ways: (1) cognitive processes are articulated as verbs instead of nouns; (2) the *synthesis* cognitive process from Bloom's taxonomy is called *create* and is situated as the most complex cognitive process; and (3) an additional dimension (i.e., knowledge type) characterizes the type of knowledge associated with an objective.

Table 1
The Anderson and Krathwohl taxonomy of learning.

Knowledge Dimension	Cognitive process dimension					
	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual						
Conceptual						
Procedural						
Metacognitive						

put it simply, analyzing the relations between elements typically requires first understanding each element in isolation. Similarly, evaluating alternatives typically requires first analyzing their similarities and differences.

The *knowledge type* dimension defines the type of knowledge associated with the learning objective. The taxonomy considers four types of knowledge. *Factual* knowledge is defined as declarative knowledge about discrete, isolated elements. *Conceptual* knowledge relates to concepts, categories, models, principles and theories. Anderson and Krathwohl argue that factual knowledge relates to bits of information, while conceptual knowledge relates to concepts people can use to organize bodies of knowledge in a systematic and interconnected manner (Anderson et al., 2001, p. 42). *Procedural* knowledge relates to step-by-step (or "how to") knowledge about performing a task. Finally, *metacognitive* knowledge relates to knowledge about one's own cognition or cognition in general. 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. In prior work, we also argued that the knowledge type dimension runs along continuums from objective to subjective and isolated to interrelated (Urgo, Arguello, & Capra, 2020). To illustrate, let us consider the difference between facts and concepts. Facts are objective bits of information about the external world. Conversely, working with conceptual knowledge may involve more subjectivity. For example, deciding whether a painting exemplifies an artistic style may require a judgment call based on context or personal experience. Additionally, facts are typically isolated bits of information that "are believed to have value in and of themselves" (Anderson et al., 2001, p. 42). Conversely, understanding a concept (e.g., a specific artistic style) may rely on understanding how it relates to other concepts (e.g., other artistic styles).

The A&K taxonomy can be used to categorize learning objectives at the intersection of these two orthogonal dimensions. The cognitive process dimension is represented by the "verb" of a learning objective and the knowledge type dimension is represented by the "noun". For example, the learning objective "the learner will be able to *distinguish* between the arts movements of *surrealism* and *Dadaism*" is 'analyze/conceptual'. The objective is analyze because it asks the learner to *distinguish* between elements. The objective is conceptual because the art movements of *surrealism* and *Dadaism* are categories or classifications of art. The A&K taxonomy can also be used to design assessment questions that align with a specific objective. For example, after an instructional session, a learner could be asked: "What are characteristics that *distinguish* between *surrealism* and *Dadaism*?"

2.2. Search-as-learning: Emergence and prior work

The search-as-learning movement is rooted on the idea that learning is an important outcome of search. In this respect, search systems can (and should) be designed and evaluated as tools to support learning. In the last decade, two summits have taken place to develop research agendas in the area of search-as-learning. In 2012, the 3-day SWIRL workshop emphasized the importance of supporting learning during search as one of many emerging topics (Allan et al., 2012). In 2017, Dagstuhl Seminar 17092 was entirely dedicated to search-as-learning (Collins-Thompson et al., 2017).

At SWIRL 2012, researchers identified three key directions to pursue: (1) going beyond a simple ranked list of results, (2) developing search tools to support learning, and (3) modeling contextual factors that may impact learning during search. With respect to contextual factors, researchers asserted that important factors are likely to vary across individuals and evolve for the same individual over time. Dagstuhl Seminar 17092 (Collins-Thompson et al., 2017) brought together researchers from various backgrounds. Attendees asserted that current search systems are effective for simple lookup tasks but do not support complex search tasks that require "exploration and learning, user collaborations, and involve different [...] search strategies" (Collins-Thompson et al., 2017, p. 135). Attendees discussed three critical roadblocks impeding the advancement of search-as-learning research: (1) the reliance on small-scale lab studies that may lack ecological validity, (2) the lack of awareness of relevant research in other disciplines, and (3) the lack of shared research infrastructure. Discussions focused on addressing these bottlenecks. Additionally, attendees discussed four directions to explore in future work: (1) understanding search as a learning process, (2) understanding how contextual factors can influence learning processes, (3) developing materials to measure learning, and (4) developing search tools to support learning. In additional to these summits, several conference workshops (Freund et al., 2014; Gwizdka, Hansen, Hauff, He, & Kando, 2016) and journal special issues (Eickhoff, Gwizdka, Hauff, & He, 2017; Hansen & Rieh, 2016) have been exclusively devoted to search-as-learning.

Studies in the area of search-as-learning have investigated a wide range of research questions. Some studies have investigated factors that influence learning during search. Specifically, studies have focused on how learning is impacted by: (1) characteristics of the individual searcher, (2) characteristic of the search task (i.e., the learning objective), and (3) characteristics of the search system. Additionally, studies have investigated the relation between specific behaviors and learning outcomes.

As described in detail in Section 4, studies have adopted a wide range of methods to assess learning during search. Some studies have measured learning by administering pre- and post-tests with predefined correct answers, including: (1) true-or-false (Freund et al., 2016; Gadiraju et al., 2018; Kalyani & Gadiraju, 2019; Nelson et al., 2009; Qiu et al., 2020; Yu et al., 2018), (2) multiple-choice (Davies, Butcher, & Stevens, 2013; Freund et al., 2016; Heilman et al., 2010; Heilman & Eskenazi, 2006; Kalyani & Gadiraju, 2019; Syed & Collins-Thompson, 2017; Weingart & Eickhoff, 2016) and (3) short-answer tests (Abualsaud, 2017; Câmara et al., 2021; Collins-Thompson et al., 2016; Davies et al., 2013; Hersh et al., 1995; Roy et al., 2020, 2021). Other studies have asked participants to complete more *open-ended* exercises. Specifically, studies have measured learning by asking participants to: (1) list relevant key phrases and facts (Bhattacharya & Gwizdka, 2019; Kammerer et al., 2009); (2) create visual representations of a domain (Liu, Liu, & Belkin, 2019); (3) enumerate arguments for and against a specific proposition (Demaree et al., 2020); and (4) summarize their knowledge of a topic (Abualsaud, 2017; Collins-Thompson et al., 2016; Davies et al., 2013; Kalyani & Gadiraju, 2019; Lei et al., 2015; Liu & Song, 2018; O'Brien et al., 2020; Palani et al., 2021; Pardi et al., 2020; Salmerón, Delgado, & Mason, 2020; Willoughby et al., 2009). To assess learning from open-ended responses, studies have adopted grading strategies that involve: (1) counting relevant concepts or facts (Abualsaud, 2017; Bhattacharya & Gwizdka, 2019; Collins-Thompson et al., 2016; Kammerer et al., 2009; Palani et al., 2021; Willoughby et al., 2009); (2) counting relevant pro/con arguments (Demaree et al., 2020); and (3) counting statements that show evidence of mental processes such as critical or creative thinking (Abualsaud, 2017; Collins-Thompson et al., 2016; Liu & Song, 2018; O'Brien et al., 2020; Palani et al., 2021; Salmerón et al., 2020). Finally, studies have also considered self-reported *perceptions* of learning (Collins-Thompson et al., 2016; Freund et al., 2016; Ghosh et al., 2018; Heilman et al., 2010; Kammerer et al., 2009; Liu, Liu, & Belkin, 2019) and *behavioral measures* that are assumed to provide evidence of learning (Chi et al., 2016).

In the following sections, we summarize findings from prior work. In our review, we focus primarily on key takeaways with respect to learning outcomes.

2.2.1. The effects of individual characteristics on learning during search

Several studies have investigated the effects of domain knowledge on learning during search (O'Brien et al., 2020; Roy et al., 2020; Willoughby et al., 2009). O'Brien et al. (2020) asked participants to produce knowledge summaries before and after completing three search tasks on the same general topic. Compared to domain experts, novices had slightly greater improvements in their summary scores, possibly because they uncovered more *new* information while searching. Willoughby et al. (2009) asked participants to produce knowledge summaries on a given domain. Some participants were instructed to search for 30 min before producing their summaries and other participants were instructed to produced their summaries without searching. Participants in the *search condition* produced summaries with more accurate facts. Interestingly, however, this effect was stronger for participants with greater prior knowledge. The authors hypothesized that participants with greater prior knowledge were able to search more effectively. Roy et al. (2020) investigated the role of domain knowledge on learning *during* the search session. To this end, participants completed quick vocabulary learning assessments at regular intervals throughout the search session. Prior knowledge influenced *when* participants had greater knowledge gains—novices had greater knowledge gains towards the start of the session and experts had greater knowledge gains towards the end.

Beyond domain knowledge, studies has also considered how individual abilities impact learning during search. Pardi et al. (2020) considered the impact of working memory capacity and reading comprehension ability. Both abilities had positive effects on learning, which was measured based on the number of relevant concepts included in knowledge summaries produced by participants before and after searching.

2.2.2. The effects of task characteristics on learning during search

Task complexity is a characteristic that has been found to influence search behaviors and perceptions (Capra, Arguello, Crescenzi, & Vardell, 2015; Jansen et al., 2009; Kelly, Arguello, Edwards, & Wu, 2015; Wu et al., 2012). Several studies have investigated the effects of task complexity on learning outcomes. Importantly, to manipulate task complexity, studies have leveraged the cognitive process dimension from the A&K taxonomy (Ghosh et al., 2018; Kalyani & Gadiraju, 2019).

Ghosh et al. (2018) had participants complete tasks associated with the cognitive processes of understand, apply, analyze, and evaluate. Participants self-reported significant knowledge gains across all tasks. Additionally, participants selected different 'action verbs' when describing their mental activities during the task (e.g., 'define' for remember, 'demonstrate' for apply, and 'relate' for analyze and evaluate tasks). Kalyani and Gadiraju (2019) had participants complete tasks associated with all six cognitive processes from the A&K taxonomy. Learning was measured using closed-ended tests for simple tasks and open-ended tests for complex tasks. Participants had lower knowledge gains for complex tasks (i.e., apply < analyze). Liu, Liu, and Belkin (2019) had participants complete two tasks of varying cognitive complexity: a *receptive* (i.e., remember or understand) task and a *critical* (i.e., evaluate) task. To measure learning, participants constructed mind maps (i.e., graphical domain representations) before each task, and modified their mind maps throughout the search session. During the receptive (i.e., simpler) task, participants made structural changes to their mind maps throughout the whole session. Conversely, during the critical (i.e., more complex) task, participants made more structural changes towards the end of the session.

Beyond task complexity, research has also studied learning during multi-session search. Liu et al. (2013) had participants complete three subtasks on the same general topic. In the dependent subtask condition, the three subtasks built on each other. Conversely, in the parallel subtask condition, the three subtasks were largely independent of each other. To measure learning, participants rated their own familiarity with the general topic after each subtask. As expected, participants reported greater topic familiarity after each subtask. Interestingly, however, this increase in topic familiarity plateaued faster in the parallel (vs. dependent) subtask condition. This result suggests that learners may benefit from subtasks that build on each other.

2.2.3. The effects of system characteristics on learning during search

Studies have also investigated how search systems and features can impact learning. Studies have considered different system characteristics: (1) the type of device used to search, (2) the presence of novel search tools, and (3) the retrieval algorithm.

Demaree et al. (2020) compared learning outcomes from participants searching on a smartphone versus laptop computer. Participants were asked to gather information on a controversial topic and write an argumentative essay. Learning was measured by counting pro and con arguments included in the essay. The search device had a significant effect on participants' behaviors but not their learning outcomes.

Freund et al. (2016) investigated the impact of two factors on participants' reading comprehension of pre-selected articles: (1) whether articles were displayed in plain text versus HTML, which included distracting elements (e.g., ads), and (2) whether participants could add "sticky notes" to articles. Without the "sticky notes" tool, participants had higher reading comprehension scores in the plain text versus HTML condition. Conversely, with the "sticky notes" tool, participants performed equally well in both conditions.

Kammerer et al. (2009) evaluated a system that enabled users to use social tags to filter search results and explore the collection. To measure learning, participants completed tests that required them to summarize their knowledge and recall domain-relevant keywords. Participants scored higher on both tests with the experimental system versus a baseline system without social tags.

Roy et al. (2021) investigated the impact of two interface features that enabled participants to highlight text and add notes. To measure learning, participants wrote post-task summaries that were analyzed based on the number of facts and subtopics included. Results found benefits from each tool in isolation—the highlighting tool resulted in summaries with more subtopics and the note-taking tool resulted in summaries with more facts. Interestingly, participants did not have positive knowledge gains when given access to *both* tools, possibly due to cognitive overload.

Câmara et al. (2021) evaluated different interface features to support learning: (1) displaying subtopics in the task domain and (2) displaying the searcher's level of coverage across subtopics during the session. Interestingly, these novel features did not significantly improve learning. Instead, they influenced participants to explore more subtopics *superficially*. As evidence, when given feedback about their topical coverage, participants viewed more search results but had shorter dwell times. Importantly, this trend suggests that feedback features can have unintended effects—they can influence searchers to pursue strategies that undermine their *depth* of learning.

In terms of tools to support learning, studies have found mixed results. In general, studies have found benefits from tools that: (1) convey more information about the items in the collection (Kammerer et al., 2009) and (2) enable searchers to annotate documents (Freund et al., 2016; Roy et al., 2021). However, results also suggest that tools can have unintended effects—they can lead to cognitive overload (Roy et al., 2021) and encourage searchers to focus on breadth of learning at the expense of depth (Câmara et al., 2021).

Studies have also considered the effects of the retrieval algorithm on learning. In the context of vocabulary learning, Syed and Collins-Thompson (2017) evaluated a retrieval algorithm that favored documents with a greater *density* of target vocabulary words. Participants had better learning outcomes with the experimental versus baseline system. Weingart and Eickhoff (2016) explored the impact of several well-established retrieval techniques on learning. To measure learning, participants completed multiple-choice tests after each task. Passage (vs. document) retrieval had a positive effect on learning, possibly because passages have a higher *density* of query-related content than whole documents. On the other hand, query expansion had a negative effect on learning, possibly due to topic drift from the original query.

2.2.4. The relation between behaviors and learning

Finally, studies have investigated how specific behaviors relate to learning outcomes. Most of these studies have focused on search behaviors captured by the search system (Abualsaud, 2017; Bhattacharya & Gwizdka, 2019; Collins-Thompson et al., 2016; Gadiraju et al., 2018; Liu & Song, 2018; Lu & Hsiao, 2017; Palani et al., 2021; Yu et al., 2018). Studies have found that searchers with better learning outcomes have a tendency to: (1) spend more time reading documents (Collins-Thompson et al., 2016; Gadiraju et al., 2018; Lu & Hsiao, 2017; Syed & Collins-Thompson, 2017; Yu et al., 2018); (2) issue queries with more advanced or uncommon vocabulary (Bhattacharya & Gwizdka, 2019; Collins-Thompson et al., 2016; Gadiraju et al., 2018); (3) issue more diverse queries within the session (Palani et al., 2021); (4) review more search results that are relevant and novel (Abualsaud, 2017; Collins-Thompson et al., 2016); and (5) visit sources that are more suitable to the task, such as encyclopedic sources during *receptive* tasks and Q&A sources during *critical* tasks (Liu & Song, 2018).

Other studies have considered behaviors that are more difficult to capture within existing search environments. Using eye-tracking, Bhattacharya and Gwizdka (2019) found that participants with better learning outcomes had fewer eye regressions (i.e., less re-reading of text). Lei et al. (2015) examined the search behaviors of 5th graders in the context of a mock school assignment involving video search. An analysis of post-search interviews found that students with better learning outcomes engaged in more metacognitive planning (e.g., setting objectives), monitoring (e.g., tracking progress), and evaluating (e,g., reconsidering strategies) during their searches.

3. Methodology

One of our goals in this paper is to provide a comprehensive review of learning assessment methods used in search-as-learning studies. Our methodology for selecting relevant research papers was adapted from Cooper (1998), Kelly and Sugimoto (2013), and Liu (2021). Our process involved four key steps: (1) defining inclusion and exclusion criteria for selecting papers to review; (2) defining search strategies for finding potentially relevant papers; (3) developing and validating a coding scheme for categorizing learning assessments used in previous studies; and (4) synthesizing our findings.

Table 2
Cohen's κ agreement across assessment types used to measure prior knowledge (pre-test) and learning (post-test). No papers used implicit measures, sentence generation, or argumentative essays to assess prior knowledge.

	Pre-test	Post-test
Self-report	0.94	1.00
Implicit measure	–	1.00
Multiple-choice	1.00	1.00
Short-answer	1.00	0.92
Free recall	1.00	0.65
Sentence generation	–	1.00
Mind map	1.00	1.00
Argumentative essay	–	1.00
Summary & Open-ended	1.00	1.00

3.1. Inclusion & exclusion criteria

The first step involved defining inclusion criteria for selecting papers to review. We decided to include papers that satisfy two criteria:

1. The research paper involved active information seeking for the purpose of learning.
2. The research paper used a measurement of learning (or perceptions of learning) as a result of an information-seeking session.

The first criterion ensured that each paper involved study participants actively engaged in information seeking. This is important because we are primarily interested in how learning assessments have been operationalized in *search* studies. We excluded studies that measured learning in non-search scenarios (e.g., learning assessed from students enrolled in a MOOC (Moreno-Marcos, Pong, Muñoz Merino, & Kloos, 2020) or from students in a classroom setting equipped with electronic note-taking tools (Kim, Turner, & Pérez-Quiñones, 2009)).

The second criterion required each paper to measure learning outcomes either objectively by administering one or more pre- and post-tests or subjectively by capturing perceptions of learning. We excluded studies that only used a pre-test to measure and study the role of prior knowledge during search (e.g., Wildemuth (2004)). Such studies were excluded because they did not measure learning outcomes. Additionally, we excluded studies that administered a post-test to measure learning but did *not* analyze the assessment data.

3.2. Search strategies

We gathered potentially relevant papers from a set of academic publication databases and search engines, including Google Scholar, ACM Digital Library, Science Direct, Academic Search Premier, and Microsoft Academic. These search engines and databases provided access to articles from the major conferences and journals that publish research on search-as-learning (e.g., SIGIR, CHIIR, CIKM, TOIS, JASIS, ASIS&T). Queries were formulated using combinations of search terms such as: search-as-learning, information seeking, search, user study, web search, online search learning, learning assessment, learning evaluation, learning outcome, learning measurement, learning metric, etc. Keyword searches yielded an initial set of relevant articles. We used two methods to identify additional papers—(1) we considered other work from authors in the initial set; and (2) we followed chains of citations from papers in the initial set.

Based on our inclusion criteria, three papers were considered borderline cases (Freund et al. (2016), Hornbæk and Frøkjær (2003), and Shi, Otto, Hoppe, Holtz, and Ewerth (2019)). In these papers, participants did not proactively search a document collection or the open Web. Instead, participants interacted with a preselected set of documents for the purpose of learning. Ultimately, all three papers were included because they explicitly discuss implications for designing search environments to support learning. Our final review includes 40 research papers focusing on learning outcomes during and/or after information seeking (and related activities).

3.3. Categorizing learning assessments

All 40 papers included in our review used some form of learning assessment. All studies administered some form of assessment *after* participants completed search tasks. Additionally, many studies administered tests of prior knowledge *before* participants completed search tasks. Using an inductive qualitative coding process, we identified nine learning assessment types (defined in Section 4): (1) self-report, (2) implicit measure, (3) multiple-choice, (4) short-answer, (5) free recall, (6) sentence generation, (7) mind map, (8) argumentative essay, and (9) summary & open-ended.

In Section 4, we report on which assessment types have been used frequently (and infrequently) in prior work. To support this analysis, assessment types were defined as unambiguously as possible.[3] To test the reliability of our assessment type definitions, both authors *independently* annotated all 40 papers. Each author reviewed each paper and indicated which types of assessments

[3] The full coding guide with the complete set of assessment type definitions and examples can be downloaded from https://www.kelseyurgo.com/ipm-2021/.

were used to measure prior knowledge (i.e., pre-test) and learning (i.e, post-test). Some studies used multiple assessment types in the pre- and/or post-task phases. Table 2 shows interannotator agreement based on Cohen's κ. In most cases, we observed perfect agreement ($\kappa = 1.0$). In two cases, we observed agreement at the level of "almost perfect" ($0.80 < \kappa < 1.0$) (Landis & Koch, 1977). In one case, we observed agreement at the level of "substantial" ($0.60 < \kappa \leq 8.0$) (Landis & Koch, 1977). The lowest agreement ($\kappa = 0.65$) was observed for "free recall" assessments used in the post-task phase. This can explained by the rarity of this assessment category. Ultimately, only three studies were coded as using a free recall assessment in the post-task phase. Initially, both authors agreed on two of those three.[4] After this independent coding phase, disagreements were resolved through discussion. For example, the assessment under dispute in the "free recall" category was from Wilson et al. (2008). This assessment asked participants to recall and write facts learned during the search process. One author coded the assessment as "free recall" and the other coded it as "summary & open-ended". After some discussion, the authors agreed to label the assessment as "free recall" because it was graded based on the number of facts included in the summary.

4. Learning assessments in search-as-learning

The learning assessment is an integral component of a search-as-learning study. In most prior work, learning assessments have been used to measure prior knowledge and learning as an *outcome* of the search process. Additionally, a few studies have used learning assessments to investigate the learning *process* during search (Liu, Liu, & Belkin, 2019; Roy et al., 2020).[5]

Selecting a particular type of learning assessment is a powerful choice that affects the study results and their implications. The assessment method and materials determine the *type(s)* of learning being measured. Learning can be characterized from different perspectives. For example, studies have considered: (1) the types of knowledge acquired (e.g., factual vs. conceptual vs. procedural), (2) the learner's ability to use the acquired knowledge to engage in specific cognitive processes (e.g., recalling vs. differentiating vs. critiquing), and (3) the learner's ability to use the acquired knowledge over time (e.g., short- vs. long-term retention). The choice of learning assessment also affects the implications of a study. To illustrate, consider a study in which participants interact with two different systems and learning is measured by how many topic-related keywords participants are able to recall after searching. Any trends observed in the results would only support recommendations for designing systems to help users *recall* information after searching.

There are also practical issues to consider when selecting a type of assessment. First, assessment types vary based on how easy they are to prepare, administer, and grade. Second, some assessment types ask specific questions that may not capture everything participants learned during a search session. For example, multiple-choice tests are easy to administer and grade but may not capture everything learned. Finally, some assessment types are more open-ended than others. Open-ended assessments have the potential to capture breadth and depth of learning. However, open-ended assessments also require additional steps to ensure that test scores are reliable and directly comparable across participants. For example, they typically require grading rubrics that are clearly defined and unambiguous.

In this section, we categorize learning assessments used in search-as-learning studies to date and review their theoretical and practical benefits and drawbacks.

4.1. Self-report

One way to measure learning is to ask participants directly or indirectly. Prior studies have asked participants to self-report how much they learned during a search session or to rate their topic familiarity before and after searching.

Several studies have measured learning using self-reported perceptions. 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 grade their own learning performance on a scale from 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 rated their familiarity with the topic of the subtask and general task.

Studies have also used self-report as a metric *in addition* to other types of learning assessments (Freund et al., 2016; Heilman et al., 2010; Kammerer et al., 2009; Liu, Liu, & Belkin, 2019; Wilson & Wilson, 2013; Zhang & 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.

Self-report assessments have three main benefits. First, there is no grading required. Participants typically rate their own learning (or pre-/post-task topic familiarity) using a Likert scale. Second, this type of assessment offers insight into subjective perceptions of learning. For example, Collins-Thompson et al. (2016) and Abualsaud (2017) found perceptions to be correlated with more objective

[4] Cohen's κ is a sensitive metric for rare categories.

[5] In prior work, the learning *process* has been defined as *what* is learned and *when*. Thus far, such studies have used mind maps (Liu, Liu, & Belkin, 2019) and short-answer assessments (Roy et al., 2020). While we discuss these learning process studies in the following sections, it is important to note that *any* learning assessment could potentially be used to investigate the learning process. For example, Roy et al. (2020) presented participants with short-answer assessments every two minutes throughout the search session to better understand what was being learned and when. While the authors used short-answer assessments, any other assessment category (e.g., free recall) could have been used.

measures of learning. Finally, assessment materials are easy to develop. In prior work, perceptions of learning have been measured using a single questionnaire item (Capra et al., 2018; Ghosh et al., 2018) or several items (Collins-Thompson et al., 2016).

Self-report assessments have two main drawbacks. First, depending on their wording, questionnaire items designed to measure perceptions of learning may *not* provide insights into the *types* of learning that took place. For example, questionnaire items may not distinguish between *breadth* and *depth* of knowledge gained during the search. Similarly, they may not capture whether the searcher can engage in simple *and* complex cognitive processes with the material learned during the task.

Second, self-report assessments have the potential for inaccuracy. As previously noted, some studies have found a correlation between perceptions of learning and more objective measures of learning (Abualsaud, 2017; Collins-Thompson et al., 2016). However, other studies have found the opposite (Pennycook, Ross, Koehler, & Fugelsang, 2017; Persky, Lee, & Schlesselman, 2020). Pennycook et al. (2017) investigated the impact of a phenomenon called the Dunning–Kruger effect (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Kruger & Dunning, 1999) (i.e., individuals who are less knowledgeable of a particular topic tend to overestimate their understanding compared with those who are more familiar with the topic). Pennycook et al. found that participants who performed the worst on an objective assessment *overestimated* their performance by a factor of more than three. Additionally, studies have found that individual factors can influence perceptions. For example, studies suggest that women tend to rate their performance lower than men (González-Betancor, Bolívar-Cruz, & Verano-Tacoronte, 2019) and lower than their *actual* performance (Torres-Guijarro & Bengoechea, 2017). Prior studies have even found that individual factors can have an *interaction* effect. For example, Colbert-Getz, Fleishman, Jung, and Shilkofski (2013) studied individuals with high levels of anxiety. In this case, females with high anxiety outperformed males with high anxiety *and* had more accurate perceptions of performance. The fact that individual factors can influence perceptions also makes it difficult to compare perceptions across participants (e.g., groups assigned to different system conditions).

To summarize, additional research is needed to fully understand how individual characteristics influence self-reported perceptions of learning. Therefore, if the goal is to measure *objective* learning, self-reported measures should be used with caution and in conjunction with other assessment types.

4.2. Implicit measure

Implicit measures of learning aim to detect knowledge gains from behavioral measures captured during the search session (i.e., without eliciting information directly from participants). Chi et al. (2016) experimented with two implicit measures of learning: query complexity and click complexity. Both measures capture the extent to which searchers issue queries and click on search results that were rarely issued/clicked by other participants in the study. Query complexity (QC) is defined as:

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

where N denotes the number of study participants 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.

By splitting search sessions into five-minute segments, Chi et al. (2016) examined the average query and click complexity at different stages of the search session. Results found higher levels of query and click complexity at later stages. The intuition behind both complexity measures is that rarely issued queries and clicked documents require more knowledge to be issued or clicked (Chi et al., 2016, p. 2).

Implicit measures of learning have four key benefits. First, implicit measures are generated from search interaction data and therefore do not require grading. Second, implicit measures are easily compared across participants because they are generated from interactions with the *same* system. Third, it is easy to collect the data. There is no additional test needed after the search session. Fourth, implicit measures allow for timely feedback. For example, behavioral measures such as query and click complexity can be captured "on the fly" and communicated to searchers as a form of ongoing feedback during the session.

Importantly, implicit measures of learning have two main drawbacks. First, they lack specificity. Alternative assessment methods (e.g., multiple choice) can be designed to measure learning from different perspectives (e.g., depth, breadth, etc.). 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 of learning lack validity. We were only able to identify one study that used implicit measures as the only type of learning assessment. Additional studies are needed in order to validate whether a specific implicit measure is capturing learning versus some other phenomenon (e.g., search expertise).

4.3. Multiple-choice

Multiple-choice assessment involves asking questions with a closed set of correct and incorrect options. Many studies in search-as-learning have used multiple-choice tests to measure learning.

Freund et al. (2016) used multiple-choice tests to measure reading comprehension. Study participants read three articles under different experimental conditions. Microstructural questionnaire items tested participants' ability to *recall* information using true or false questions and *understand* information using a sentence verification technique. The sentence verification technique asked

participants to select "yes" or "no" to indicate whether a sentence accurately conveyed the overall theme or position of an article. Macrostructural questionnaire items required participants to select three statements (out of six) that accurately conveyed themes present in *all three* articles. These questions allowed participants to exhibit their understanding of the connections between articles. Participants answered 26 questions after each reading comprehension task and were scored based on the percentage of correct responses.

Gadiraju et al. (2018) measured learning by having participants complete the same multiple-choice test before and after each search task. Each search task had a pre- and post-test that included 10–20 questions, depending on the task topic. Each question had the answer options of true, false, and "I don't know". "I don't know" responses were counted as incorrect. Most questions required participants to recall factual knowledge. Knowledge gains were calculated as the difference between pre- and post-test scores. Results found that the pre-test questions had a small priming effect—participants used "an increasing number of terms from the [pre-test] in their queries" (Gadiraju et al., 2018, p. 11).

Yu et al. (2018) used the same multiple-choice tests from Gadiraju et al. (2018). However, tests were refined with a pilot test of 100 participants in order to remove questions that were too easy (more than 80% correct responses) or too ambiguous (many more false than true responses). Similar to Gadiraju et al. (2018), participants completed the same multiple-choice test before and after each search task. Xu et al. (2020) also used the multiple-choice tests from Gadiraju et al. (2018). While a portion of participants searched collaboratively (vs. individually), tests were taken individually. Knowledge gains were calculated as the difference between pre- and post-test scores. Additionally, knowledge gains were *normalized* by calculating the ratio between the participant's absolute knowledge gain and the maximum knowledge gain possible:

$$\mathrm{KG} = \frac{\mathrm{postScore} - \mathrm{preScore}}{\mathrm{maxScorePossible} - \mathrm{preScore}} \tag{1}$$

Shi et al. (2019) captured learning by having participants complete the same multiple-choice test before and after each task. For each test item, participants were asked to select the correct choice(es) from a set of options (i.e., items could have multiple correct choices). To discourage guessing, wrong answers were penalized more than unanswered items.

Qiu et al. (2020) used multiple-choice tests with the answer options of true, false, and "I don't know". Participants completed the same 10-item test before and after each task. Additionally, to measure retention of learning, participants completed the same test a few days after searching. To measure learning, the authors compared test scores between the pre-test and the immediate post-test. To measure retention, the authors considered the number of questions correctly answered in the immediate post-test but *incorrectly* answered in the long-term memory test.

Kalyani and Gadiraju (2019) asked participants to complete learning-oriented search tasks associated with specific cognitive processes from A&K's taxonomy (i.e., remember, understand, apply, analyze, evaluate, and create). Participants completed different types of pre-/post-tests depending on the cognitive process of the search task. To measure learning during remember tasks, participants completed tests comprised of statements with the answer options of true, false, or "I don't know". To measure learning during understand tasks, participants answered questions that asked them to select answers from a predefined set of options. To measure learning during apply tasks, participants completed tests that required participants to organize a set of events in the correct order. To measure learning during analyze tasks, participants completed tests that asked them to classify items (e.g., classify the following nutrients as either a mineral or vitamin).

Heilman and Eskenazi (2006) and Heilman et al. (2010) used cloze multiple-choice questions to assess vocabulary learning. Cloze questions were generated automatically from passages encountered during the search session. For example, the following is a cloze question from Heilman and Eskenazi (2006):

He could never _____ the success he had enjoyed with his first record.

acknowledge *comprise* *induce* *reproduce*

These assessments were scored by summing correct answers. To measure prior knowledge, Heilman and Eskenazi (2006) used a pre-test that also consisted of cloze questions. Pre- and post-test questions targeted the same vocabulary terms, but varied the sentence text surrounding the target word. Heilman et al. (2010) used a pre-test with self-assessment items that asked participants to answer "yes" or "no" to the question: "Do you know the word 'X'?".

Nelson et al. (2009) measured learning using pre- and post-tests with true or false options. Pre- and post-tests each included 20 questions developed around the task domain (i.e., Enterprise 2.0 Mashups). Both tests included an even distribution of easy and difficult questions as rated by 100 Amazon Mechanical Turk workers. Pre- and post-tests did not have overlapping questions. Knowledge gains were measured based on the difference between pre- and post-test scores and normalized based on Eq. (1).

von Hoyer, Pardi, Kammerer, and Holtz (2019) used a 10-item multiple-choice test designed to capture factual knowledge gains (e.g., What is a cumulonimbus cloud?). Each question had four choices and one correct answer. Participants also indicated their confidence level for each answer on a 4-point scale ranging from "not confident" to "very confident". Participants completed the pre-test one week prior to the search session to help mitigate a priming effect. Knowledge gains were measured as the difference between the pre- and post-test scores.

Syed and Collins-Thompson (2017) used multiple-choice questions to assess vocabulary learning. Each question had four choices and one correct answer. The four choices included three possible definitions and one choice that was either "all of the above" or "none of the above". Participants completed the same pre- and post-test, and knowledge gains were measured as the number of questions answered incorrectly in the pre-test and correctly in the post-test.

Weingart and Eickhoff (2016) measured learning using multiple-choice tests administered only after each search task (no pre-test). The authors designed a 5-item test per search task. Each question had six choices (with 1–4 correct answer options). Knowledge

gains were calculated proportional to the number of correct choices (i.e., true positives + true negatives). Salmerón et al. (2020) used true-or-false questions to assess prior knowledge. The authors developed two multi-item tests (one per task). Using Cronbach's α, the authors reported low internal consistency between multiple-choice items for the same task. Therefore, scores were not used to assess prior knowledge.[6]

Davies et al. (2013) explored the impact of two note-taking tools on learning during search. A factual knowledge assessment with 33 multiple-choice questions was administered before and after each search tasks. Tests were scored based on the number of correct answers.

Multiple-choice assessments have four main benefits. First, multiple-choice items 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 generate their own novel responses. Third, because tests are comprised of identical question and answer options, test scores can be easily compared across participants. Finally, because they are quick to administer and easy to grade, multiple choice tests can be used to provide timely feedback to users during a search session.

Multiple-choice assessments, however, also have several drawbacks. First, because answer options are provided, participants may guess their responses which may lead to inaccurate assessment scores. Second, multiple-choice tests have limited coverage. In other words, they test specific knowledge and may not capture everything a participant learned during the search session. Third, depending on their design, multiple-choice items may not capture a participant's ability to engage in complex cognitive processes. Simply selecting an answer from a set of options does not allow participants to demonstrate higher-level cognitive processes (e.g., creating a novel solution to a problem). Finally, multiple-choice assessments can be difficult to develop. Multiple-choice tests that are reliable and valid tend to be carefully constructed. Answer options should include incorrect distractors that are rooted in common misconceptions. Developing good multiple-choice tests requires domain knowledge and may also benefit from pedagogical experience in the task domain.

In order to mitigate these drawbacks, researchers should consider *carefully* selecting or constructing questions in multiple-choice assessments. As shown in prior work, the design of multiple-choice questions can vary greatly. One can envision a wide spectrum of multiple-choice items to measure different depths of learning. To measure rote memorization, one might use true/false questions. To measure deeper understanding, one might use questions that ask participants to categorize elements. To measure the ability to think critically, one might use questions that ask participants to prioritize elements based on specific criteria. As an example, to test different levels of reading comprehension, Freund et al. (2016) included multiple-choice items that asked about information in a *single* article as well as items that asked about themes across *multiple* articles.

4.4. Short-answer

Short-answer assessments involve asking questions that are open-ended but have a relatively short and objectively correct answer. Short-answer questions do not ask participants to make one or more selections from a predefined set of options. Instead, participants are a required to generate a response completely on their own.

Hersh et al. (1995) used short-answer questions to guide search tasks and measure learning. Before each task, participants completed a 10-item short-answer assessment. Participants were then asked to identify their 5 most uncertain answers. After searching, participants were asked to (re-)answer the same five questions using information found during the search. Pre- and post-search assessments were scored based on predefined correct answers. Hornbæk and Frøkjær (2003) conducted a study in which participants read scientific articles in different experimental conditions. The authors measured incidental learning using six short-answer questions generated from the content of the articles given to participants. Each question had a predefined correct answer, resulting in scores ranging from 0 to 6.

Câmara et al. (2021), Moraes, Putra, and Hauff (2018), Roy et al. (2020), Roy et al. (2021) used the same type of assessment to measure conceptual learning. In all four studies, participants completed the same pre- and post-test. All studies used assessment questions that combined self-report and short-answer formats. Each question asked participants to rate their knowledge of a specific concept on a 4-point scale. Additionally, participants who rated their knowledge at the levels of 3 or 4 were also asked to define the concept in their own words. In all four studies, the authors evaluated the reliability of participants' self-rated knowledge. To this end, the authors manually graded a sample of definitions provided by participants who self-rated their knowledge at the levels of 3 or 4. In all four studies, participants' self-rated knowledge of specific concepts was found to be sufficiently reliable to be used as the primary measure of learning.

In addition to using multiple-choice questions, Davies et al. (2013) included short-answer questions to measure factual learning during searches supported by different note-taking tools. Participants completed search tasks that asked them to learn about plate tectonics. The short-answer questions involved diagrams with blank spaces for participants to label with the correct components or processes depicted in the picture. Participants were asked to generate 13 diagram labels. Each label was given a score from 0 to 1 (0.5 for a partially-correct label).

Collins-Thompson et al. (2016) and Abualsaud (2017) both developed assessment questions that varied in cognitive complexity according to the A&K taxonomy. While the majority of questions were summary or open-ended, a few questions (e.g., remember cognitive process level) were short-answer questions (i.e., had predefined correct answers).

[6] When multiple-choice items are intended to capture knowledge gains along the same dimension, Cronbach's α can be used to determine whether they are actually capturing the same learning construct.

There are several benefits to using short-answer assessments to measure learning. First, short-answer assessments have predefined correct answers and are therefore easy to grade. Second, assessment scores can be easily compared across participants. Finally, short-answer assessments minimize guessing because the assessments do not provide a pool of options to choose from.

There are three drawbacks of short-answer assessments. First, short-answer assessments have low coverage. Because short-answer assessments have targeted questions with preset answers, the assessment may not fully capture all that was learned during a search session. Second, short-answer questions require a lot of time and effort to develop. Questions and answers from these assessments must be created carefully to measure both breadth and depth of learning. Finally, short-answer assessments have the *potential* to be limited in measuring a participant's ability to engage in complex cognitive processes. For example, consider short-answer questions that ask participants to define a concept in their own words (Câmara et al., 2021; Moraes et al., 2018; Roy et al., 2020, 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).

4.5. Free recall

Free recall assessments involve asking participants to list as many important terms, phrases, or facts related to the topic of a search task. Typically, participants complete free recall assessments before and after the search task. Knowledge gains can be measured in different ways, for example, by counting the number of *new* terms, phrases, and/or facts included the post-task list.

Bhattacharya and Gwizdka (2019) measured learning using free recall assessments. Before and after each search task, participants were asked to list single words or phrases relevant to the search task domain. Knowledge gains were measured in two ways. First, a "simple" knowledge gain measure was calculated as the difference between the number of words/phrases listed before and after each search task. Second, the authors compared the semantic similarity between these lists and a gold-standard list of words/phrases curated by a domain expert. Wilson et al. (2008) evaluated different interfaces for browsing a music collection. As part of the evaluation, participants were asked to recall and write down facts they learned about the items in the collection. Interfaces were evaluated by comparing the number of recalled facts across conditions. Kammerer et al. (2009) evaluated a search system that displayed social tags in addition to search results. As part of the evaluation, after each search task, participants were asked to list as many domain-relevant keywords as possible. Knowledge gains were measured by coding and tallying the number of "reasonable keywords" listed. Keywords present in the task description were excluded.

There are several benefits to using free recall assessments to measure learning. Free recall assessments are simple to develop, quick to administer, and relatively easy to grade (e.g., counting items). Scores can be computed by simply counting the number of keywords, phrases, and/or facts recalled. Additionally, if participants are given clear instructions, test scores are directly comparable across participants. Finally, free recall assessments require participants to generate their own answers, which minimizes guessing.

Similar to multiple-choice assessments, free recall assessments have two main drawbacks. First, free recall assessments typically ask participants to write down terms, phrases, and/or facts associated with a specific topic. This constraint does not allow participants to demonstrate knowledge gains outside of the given topic. Second, free recall assessments test a participant's ability to *remember* terms, phrases, and/or facts. They do not reliably test whether a participant can engage in more complex cognitive processes. For example, they do not reliably determine whether a participant understands a concept, can differentiate between concepts, or can evaluate the usefulness of a concept to explain a phenomenon.

4.6. Sentence generation

Sentence generation is typically used to assess vocabulary learning. During this assessment type, participants are asked to generate sentences using specific vocabulary terms. Usually, participants are asked to generate sentences that: (1) are grammatically correct and (2) demonstrate that the meaning of the target vocabulary word is *fully* understood. *Transfer of learning* occurs when an individual is able to use or apply acquired knowledge is a novel scenario or domain. Prior studies have used sentence generation techniques to measure *transfer of learning* during vocabulary acquisition (Heilman et al., 2010; Heilman & Eskenazi, 2006). For example, if someone learned the meaning of the word "spectrum" in the context of color, can they use it properly in the context of sound?

Heilman and Eskenazi (2006) and Heilman et al. (2010) asked participants to generate novel sentences using target vocabulary terms to measure transfer of learning. Generated sentences were scored from zero to three using the following grading scheme. One point was given to sentences that were grammatically correct. A second point was given if the word fit semantically with the topic of the sentence. A third point was given if the word was used in a way that unambiguously conveys a full understanding of its meaning. For example, the following sentence would be given three points: "It is too dangerous to *abandon* your children in a busy street". Conversely, the following sentence would be given one point: "He *abandon* his work". In this case, the use of *abandon* is semantically correct, but the sentence is both ungrammatical and it is unclear whether the meaning of *abandon* is fully understood. In Heilman et al. (2010), sentences were graded by two instructors and a course supervisor. Each instructor graded half the sentences and the supervisor graded all the sentences. This way, each sentence received two grades and disagreements were resolved by averaging the scores. The correlation between the instructor and supervisor grades was 0.68. The authors stated that this low correlation "reflects the difficulty of assessing vocabulary knowledge based on this type of written output" (Heilman et al., 2010, p. 88).

There are two main benefits in using sentence generation assessments to measure learning. First, sentence generation tests are easy to develop. Second, participants must generate their own responses, which minimizes guessing.

There are several drawbacks of sentence generation assessments. First, sentence generation responses can be difficult to compare across participants. This can be potentially alleviated with careful coding that allows for specific criteria to be compared. Second, because grading is time-consuming, feedback may also take time to administer. Third, sentence generation is specific to certain words. It does not measure learning of new vocabulary that is not targeted on the assessment. Finally, sentence generation can be used to assess whether someone *understands* the meaning of a word (e.g., What does it mean "to *abandon*?"). However, sentence generation may not reliably measure whether an individual can engage in more complex cognitive processes with a target vocabulary word. For example, it does not determine whether someone can differentiate between *abandon* and *relinquish* (i.e., an *analyze*-level cognitive process).

4.7. Mind map

A mind map is a diagram used to visually organize information about a specific topic or system. Similar to a concept map, a mind map is comprised of nodes and links between nodes. There are, however, several differences between mind maps and concept maps. First, mind maps tend to focus on one central concept, system, or idea. Therefore, they tend to follow a hierarchical tree structure—main topics connect to the central topic, and subtopics connect to those main topics, etc. Second, while concept maps have labeled links describing the relationship between connected concepts, mind maps typically have unlabeled links. Third, while concept maps represent objective relationships between concepts, mind maps are subjective to the creator, without predefined "correct" relationships between nodes (Liu, Liu, & Belkin, 2019).

Mind maps are a relatively new type of learning assessment in search-as-learning studies. Liu, Liu, and Belkin (2019) used mind maps to better understand knowledge shifts during the search process. Given an assigned task, participants were asked to create a mind map based on their prior knowledge and modify the mind map while gathering information. The authors characterized different types of changes participants made to their mind maps throughout the search session. Mind map changes were characterized based on the type of change (e.g., adding vs. modifying vs. deleting nodes) and the location of the change (e.g., level 1–2 nodes vs. level 3 nodes and beyond). This taxonomy was used to analyze common and uncommon types of changes made during difference stages of the search task. Additionally, it was used to characterize search sessions based on different patterns (e.g., search sessions with frequent changes early vs. late). Similarly, Zhang and Liu (2020) asked participants to create mind maps based on their prior knowledge and modify their mind maps while gathering information. Pre-search mind maps were used to better understand the role of prior knowledge on querying behavior. For example, results found that participants' queries mostly included vocabulary from level 2 nodes (vs. level 1 or level 3 and beyond). Post-search mind maps were used to better understand how the search process influenced participants' knowledge. Results found that most changes happened on level 3–4 nodes.

Mind map assessments have four main benefits. First, mind map assessments are open-ended and therefore have high topical coverage. In other words, participants can convey relationships between any elements deemed meaningful in the task domain. Second, mind map assessments can be used to better understand the learning *process* during search (i.e., how knowledge structures shift during a search session). Specifically, mind maps allow researchers to investigate which nodes (i.e., concepts) and edges (i.e., relationships between concepts) were added, deleted, or edited and when. Third, mind maps can be easily analyzed to measure learning. Prior studies have analyzed mind maps by considering specific types of modifications (e.g., node additions, deletions, and edits). Finally, there is no development required. Mind map construction tools are readily available, and the mind maps themselves are developed by participants.

Mind map assessments have two main drawbacks. First, mind maps convey relationships between elements. From a cognitive process perspective, generating a mind map requires comparing, contrasting, and differentiating, which are analyze-level processes. It is unclear whether mind maps can be used to measure a learner's ability to engage with more complex cognitive processes (e.g., evaluate and create). Second, mind maps are not necessarily familiar to participants. In fact, Liu, Liu, and Belkin (2019) required mind-mapping experience while enrolling participants in the study.

4.8. Argumentative essay

Argumentative essays ask participants to write an essay containing arguments for and against a specific stance or proposition. In Demaree et al. (2020), participants used different types of devices to gather information about a controversial topic—the potential for nuclear power to help solve the climate crisis. After searching, participants were asked to write an argumentative essay listing arguments for and against the proposition that nuclear power can help solve the climate crisis. Essays were graded based on the number of correct pro and con arguments. Two independent coders graded a subset of essays and achieved 96% and 93% agreement in identifying correct pro and con arguments, respectively.

Argumentative essay assessments have four main benefits. First, depending on the grading criteria, argumentative essays can be relatively easy to grade. For example, counting the number of pro and con arguments simply requires enumerating the different arguments made by all participants. Second, the assessment allows for direct comparison across participants. Researchers can simply compare the number of pro and con arguments across essays. Third, guesswork is minimized because the essays are fully generated by participants. Fourth, the assessment materials are easy to develop. Participants simply need instructions about enumerating arguments for and against a specific stance or proposition (e.g., nuclear energy can alleviate climate change).

Argumentative essay assessments have two main drawbacks. First, the assessment focuses on recalling arguments for and against a specific stance or proposition. Therefore, the assessment may not capture everything learned during the search session (e.g., gaining a deeper understanding of certain concepts). Second, the grading approach can limit the types of learning that are being reliably captured. Simply counting the number of pro and con arguments included in an essay is only reliably capturing a participant's ability to *recall* arguments. The participant may not fully *understand* all the arguments listed nor be able to *evaluate* the validity of an argument or its relative importance. In theory, grading criteria could include other dimensions (e.g., the extent to which arguments are substantiated with evidence and examples or the extent to which arguments are critiqued). However, these more subjective dimensions have not been explored in prior work.

4.9. Summary & open-ended

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

Collins-Thompson et al. (2016) and Abualsaud (2017) measured learning during complex search tasks by developing assessments that included six questions per task. Three questions were designed to measure "lower-level learning" and asked participants to recall specific factual information. Three open-ended questions were designed to measure "higher-level learning" and asked participants to: (1) write an outline for a hypothetical paper on the subject of the task, (2) describe what they learned about the topic, and (3) enumerate unanswered questions about the task topic. Responses to each question were graded on a 7-point scale using a qualitative coding scheme. Criteria in the coding scheme checked for the inclusion of facts, themes, issues, concepts, and the relationships between concepts. Additionally, before searching, participants answered an open-ended question about their prior knowledge on the task topic. Participants' responses about their pre- and post-task topic knowledge (item #2 above) were also graded on a 3-point scale based on the type knowledge exhibited in the response—no knowledge (0), factual knowledge (1), and conceptual knowledge (2). Knowledge gains were measured based on the difference in scores before and after the task.

Kalyani and Gadiraju (2019) measured learning during tasks associated with specific cognitive processes from the A&K taxonomy. Each learning-oriented search task had its own learning assessment, which was completed before and after the task. Assessments for low-complexity tasks (remember, understand, apply, analyze) had closed-ended questions. Assessments for high-complexity tasks (evaluate, create) had 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 noted that responses were "manually checked and marked [...] as valid upon encountering complete and comprehensive submissions" (Kalyani & 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 independent assessors. Instead, they measure the extent to which scores from different assessors yield similar rankings. Put differently, rank correlation metrics measure the extent to which independent assessors agree in *relative* (versus absolute) terms.

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' 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 that suggest critical thinking.

O'Brien et al. (2020) measured learning by asking participants to summarize their knowledge of the task topic 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 min 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). As previously noted, these dimensions consider the extent to which the cognitive processes of understand, analyze, and evaluate are reflected in the summary, respectively. 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 outcomes during search tasks with two different types of objectives: (1) compare alternatives without making a recommendation (analyze) and (2) weigh the pros and cons of one alternative and make a recommendation (evaluate). Specifically, the study investigated the effects on learning outcomes from the types of sources visited during the session (i.e., encyclopedic vs. community Q&A sources). To measure learning, participants were asked to summarize their answers to the task before and after searching. Pre- and post-task summaries were scored along seven dimensions. First, summaries were scored based on: (1) number of facts, (2) number of dimensions considered when analyzing alternatives, (3) the percentage of dimensions considered compared to *all* dimensions associated with the task (determined in advance), and (4) the ratio between the number of facts and dimensions considered. Additionally, summaries were scored based on the inclusion of: (5) relevant information, (6) pros and cons, and (7) the participant's own opinions. Two assessors coded all summaries and achieve 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 for these natural phenomena 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 their relation 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 correct concepts included in the summary. The authors report an intercoder agreement of 73%.

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 express all that they have 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, the assessment can target varying levels of cognitive complexity per question. 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. A few prior studies have simply asked participants 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 the development of detailed qualitative coding guides. This process involves defining grading criteria and measuring intercoder agreement to ensure that the coding guide is reliable. Prior studies have scored summaries along dimensions such as the inclusion of facts, relationships between facts, and evaluative statements. 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 are likely to vary across participants. Some participants may not be able to effectively *communicate* everything they learned. In other words, writing ability may be a confounding factor that we need to be wary of.

4.10. Summary of learning assessments

Table 3 provides an overview of all the different learning assessment types used in search-as-learning studies to date. The table is organized with the 9 assessment types on the left and 8 categories of benefits and drawbacks along the top. The various categories of benefits and drawbacks might serve as practical checklists for researchers based on the constraints of a potential search-as-learning study. The constraints across assessments include: (1) difficulty of grading; (2) difficulty in comparing assessment scores across participants; (3) difficulty in developing the assessment; (4) the time required to *administer* the assessment; (5) the time required to *evaluate* the assessment and potentially deliver feedback; (6) the risk of inaccuracy; (7) the level of coverage in assessing all that was learned; and (8) the potential to target a wide range of cognitive processes from the A&K taxonomy. To illustrate, consider the benefits and drawbacks of multiple-choice assessments. Multiple-choice tests have predefined correct answers, which allow for easy grading and comparison across participants. Development, however, can be difficult. Multiple-choice questions must be carefully constructed with clear correct answer candidates and distractor candidates that are grounded in common misconceptions. Multiple-choice tests are relatively quick to administer and allow for quick feedback. In terms of guesswork, multiple-choice tests allow participants to guess from a set of options, which may lead to inaccurate reflections of learning. Also, multiple-choice tests are very focused and do not allow participants to convey things that were learned outside of the questions in the assessment. Finally, multiple-choice tests may not be able to reliably capture a participant's ability to engage in highly complex cognitive processes (e.g., create a new solution to a problem).

Table 4 summarizes the assessment types used by each search-as-learning study included in our review. Additionally, the last two columns indicate which studies have measured long-term retention and transfer of learning, two dimensions of learning that have been understudied in prior work. The table was generated to serve as a reference for researchers interested in a particular assessment type or combination of types. Across studies, the most common assessment types are multiple-choice (15 studies), summary & open-ended (16 studies), and self-report (12 studies).

In terms of common combinations, summary & open-ended assessments have been frequently used in combination with multiple-choice (3 studies) and short-answer assessments (5 studies). There are two reasons for this trend. First, studies have used closed-ended assessments to target simple cognitive processes (e.g., recalling factual knowledge) and more open-ended assessments to target complex cognitive processes (e.g., analyzing the relations between concepts). Second, summary & open-ended responses are often graded using rubrics that consider subjective criteria (e.g., the inclusion of opinionated statements that suggest critical thinking). Multiple-choice and short-answer questions have predefined correct answers and can therefore be graded more objectively. Therefore, studies have used closed-ended assessments to check the validity of scores assigned to summaries and open-ended responses. Additionally, summary & open-ended assessments have been frequently used in combination with self-report (5 studies). These studies were mainly interested in the relationship between perceived learning and actual learning.

The majority of studies in Table 4 implemented a prior knowledge test in order to identify a baseline of knowledge for a given learning task. These pre-search assessments are varied and have included: (1) self-assessed prior knowledge or topic familiarity on a Likert scale (Capra et al., 2018; Collins-Thompson et al., 2016; Freund et al., 2016; Ghosh et al., 2018; Kammerer et al., 2009; Liu et al., 2013; Liu, Liu, & Belkin, 2019; O'Brien et al., 2020; Zhang & Liu, 2020), (2) writing a summary of everything already known about the topic (Collins-Thompson et al., 2016; Demaree et al., 2020; O'Brien et al., 2020; Pardi et al., 2020; Wilson & Wilson, 2013), (3) taking the same post-search test *immediately* before searching (Gadiraju et al., 2018; Kalyani & Gadiraju, 2019; Moraes et al., 2018; Nelson et al., 2009; Syed & Collins-Thompson, 2017; Xu et al., 2020) or several days prior (Pardi et al., 2020; Roy et al., 2020), (4) taking a pre-test and identifying the most uncertain answers to address during the search task (Hersh et al., 1995), (5) listing all topically relevant terms immediately before searching (Bhattacharya & Gwizdka, 2019), and (6) constructing a mind map of a given domain using only prior knowledge (Liu, Liu, & Belkin, 2019; Zhang & Liu, 2020).

An important question is: How can we account for prior knowledge when measuring learning outcomes? Prior studies have adopted at least three general strategies: (1) focusing on knowledge *gains* by directly comparing pre- and post-test scores, (2) verifying that participants assigned to different experimental conditions had similar levels of prior knowledge, and (3) adding pre-task prior knowledge measures as covariates in the statistical analysis.

First, some studies have asked participants to complete *exactly* the same assessment before and after each search task (Gadiraju et al., 2018; Qiu et al., 2020; Shi et al., 2019; Xu et al., 2020; Yu et al., 2018). In these cases, knowledge *gains* have typically been measured based on a participant's increase from their pre-test score (before searching) to their post-test score (after searching). Some studies have used *normalized knowledge gain* as the main dependent variable (Eq. (1)) (Xu et al., 2020). To illustrate, suppose that scores for a specific test range from 0% to 100%. If a participant achieves a pre-task score of 60% and a post-task score of 80%, then their *raw* knowledge gain is 20% (i.e., (80%–60%)). However, their *normalized* knowledge gain is 50% (i.e., (80%–60%)/(100%–60%)). Normalization helps to account for the fact that participants may have greater levels of prior knowledge (i.e., greater pre-test

Table 3
Benefits and drawbacks of search-as-learning assessment types.[7]

Assessment type	Grading		Comparison of Participants		Development		Administration (completion time)		Feedback		Potential for Inaccuracy		Potential Coverage		Cognitive Complexity	
	Easy	Difficult	Easy	Difficult	Easy	Difficult	Slow	Fast	Slow	Fast	Low	High	Low	High	Low	High
Self-report	✓			✓	✓			✓		✓		✓	✓		✓	
Implicit measure	✓		✓		✓			✓		✓		✓	✓		✓	
Multiple-choice	✓		✓			✓		✓		✓		✓	✓		✓	
Short-answer	✓		✓			✓	✓			✓	✓		✓		✓	
Free recall	✓		✓		✓			✓		✓	✓		✓		✓	
Sentence generation		✓		✓	✓		✓		✓		✓		✓		✓	
Mind map	✓			✓	✓		✓		✓		✓			✓	✓	
Argumentative essay	✓		✓		✓		✓		✓		✓		✓		✓	
Summary & Open-ended		✓		✓	✓		✓		✓		✓			✓		✓

scores) for some search tasks more than others. In this respect, improving from 90% to 95% on a popular topic is *as good* as improving from 80% to 90% on a *less* popular topic. In educational research, normalization techniques have strengths and weaknesses (Nissen, Talbot, Nasim Thompson, & Van Dusen, 2018).

Prior studies have used two other approaches to combine pre- and post-test scores more *indirectly*. Demaree et al. (2020) compared the learning outcomes of participants searching on a smartphone versus laptop computer. Device type was a between-subjects factor—half the participants searched on a smartphone and half searched on a laptop. Learning was measured using *only* an argumentative essay post-test. However, participants were also asked to summarize their prior knowledge before each search task. Importantly, Demaree et al. (2020) reported that participants' scores on these pre-task summaries were *not* significantly different across device conditions.

Finally, similar to the previous strategy, studies have used dependent variables based on only post-test scores. However, to control for differences in prior knowledge, studies have included pre-task measures of prior knowledge as covariates in the statistical analysis. For example, Kammerer et al. (2009) investigated the effects of an experimental system on learning. Knowledge gains were measured using a free recall post-test. However, to account for differences in prior knowledge, the authors included participants' self-reported perceptions of prior knowledge as a covariate.

5. Learning assessments outside of search-as-learning

Search-as-learning studies have used a wide range of learning assessments. However, studies in the fields of psychology and education have used other types of assessments worth considering in future search-as-learning research. In this section, we review three additional assessment types: (1) task performance, (2) mental models, and (3) comparative judgment. Additionally, we describe a study (i.e., McNeil and Alibali (2000)) that used a combination of assessments different from combinations used in prior search-as-learning work. This combination of assessments enabled the researchers to measure breadth and depth of learning, as well as a learner's ability to transfer what was learned to solve a *new* type of problem (i.e., transfer of learning).

5.1. Task performance

People often engage in learning activities (e.g., information search) in order to accomplish a higher-level task. For example, a searcher may decide to learn about a new cooking technique in order to make a specific recipe. In such cases, the quantity or quality of learning during the search process could be measured based on the learner's performance on the higher-level task itself. Task performance can be measured based on the quality of the task *outcome* and/or the quality of the learner's *approach* to the task (e.g., the number of unnecessary steps avoided).

Singley (1990) investigated the effects of a specific system intervention added to a calculus tutoring system. While engaging with the tutoring system, participants solved calculus word problems. To measure learning, each problem had a predefined *best* path to the solution (i.e., an optimum sequence of moves). Learning was measured based on the number of unnecessary or illegal moves *avoided* by participants in their solutions. Similarly, Koedinger and Anderson (1993) explored the effectiveness of a cognitive tutor for mathematical proofs. After a series of sessions with the cognitive tutor, learning was measured by asking participants to complete a series of proofs. Proofs were graded using a rubric adopted from Senk and Usiskin (1983). Each proof was given a binary score. Proofs were given a score of 1 if they had all the *key* steps correct (determined in advance).

[7] These determinations are based on how assessments have been implemented and graded in prior work. For example, prior studies have used multiple-choice tests that can be completed in a relatively short period of time (e.g., 30 min) and are therefore classified as being fast to administer. Similarly, argumentative essays have been graded based on the inclusion of specific pro/con arguments versus more complex criteria (e.g., the learner's ability to critique arguments). Therefore, they are classified as easy to grade and have low coverage of cognitive processes.

Table 4
Search-as-learning studies categorized into assessment types, retention of learning, and transfer of learning.

Study	Assessment type									Learning retention	Transfer of learning
	Multiple-choice	Free recall	Short-answer	Argumentative essay	Summary & open-ended	Mind map	Sentence generation	Implicit measure	Self report		
Abualsaud (2017)			✓		✓				✓		
Bhattacharya and Gwizdka (2019)		✓									
Câmara et al. (2021)			✓								
Capra et al. (2018)									✓		
Chi et al. (2016)								✓			
Collins-Thompson et al. (2016)			✓		✓				✓		
Davies et al. (2013)	✓		✓		✓						
Demaree et al. (2020)				✓	✓						
Freund et al. (2016)	✓								✓		
Gadiraju et al. (2018)	✓										
Ghosh et al. (2018)									✓		
Heilman and Eskenazi (2006)	✓						✓			✓	✓
Heilman et al. (2010)	✓						✓		✓		✓
Hersh et al. (1995)			✓								
Hornbæk and Frøkjær (2003)			✓		✓						
Kalyani and Gadiraju (2019)	✓				✓						
Kammerer et al. (2009)		✓			✓				✓		
Lei et al. (2015)					✓						
Liu and Song (2018)					✓						
Liu et al. (2013)									✓		
Liu, Liu, and Belkin (2019)						✓			✓		
Moraes et al. (2018)			✓								
Nelson et al. (2009)	✓										
O'Brien et al. (2020)					✓				✓		
Palani et al. (2021)					✓						
Pardi et al. (2020)					✓						✓
Qiu et al. (2020)	✓									✓	
Roy et al. (2020)			✓								
Roy et al. (2021)			✓		✓						
Salmerón et al. (2020)	✓				✓						
Shi et al. (2019)	✓										
Syed and Collins-Thompson (2017)	✓										
von Hoyer et al. (2019)	✓										✓
Weingart and Eickhoff (2016)	✓										
Willoughby et al. (2009)					✓						
Wilson and Wilson (2013)					✓				✓		
Wilson et al. (2008)		✓									
Xu et al. (2020)	✓										
Yu et al. (2018)	✓										
Zhang and Liu (2020)						✓			✓		

Problem-based learning is a style of pedagogy (i.e., instruction and assessment) that uses real-world problems as a *vehicle* to teach facts, concepts, and principles. Problem-based learning usually involves assessments such as problem-solving vignettes, simulations, and role-playing tasks. Simulation-based assessments are frequently used to evaluate and license physicians (Boulet, 2008). To illustrate, Hawkins et al. (2004) developed a case-based simulation system to assess physicians. The system presents physicians with a scenario involving a specific setting (e.g., emergency room) and a patient's symptoms, vital signs and medical history. Physicians are then required to make a diagnosis, select a treatment plan, and schedule follow-up appointments. Scores are assigned based on the quality of a physician's inferences. For example, a medium score may indicate that the physician recommended exams that might lead to the right diagnosis. Conversely, a high score may indicate that the physician made the right diagnosis, selected the right treatment, and made the appropriate follow-up appointments.

5.2. Mental model assessment

Mental models are subjective, cognitive representations of external reality (Jones, Ross, Lynam, Perez, & Leitch, 2011). Prior studies in education and psychology have used mental models to measure learning. The underlying assumption is that individuals with greater knowledge are able to generate more accurate and complete mental models of external phenomena. Nersessian argues that "the nature and richness of models one can construct [...] develops with learning domain-specific content and techniques" (Nersessian, 2002, p. 140). Communicating mental models often involves drawing diagrammatic representations with pictures, words, and symbols (Jones et al., 2011). Mental model assessments have been used to better understand the level of conceptual knowledge a learner has acquired. Additionally, mental model assessments can illuminate gaps in an individual's understanding of a system or phenomenon.

Chi, Siler, Jeong, Yamauchi, and Hausmann (2001) used mental model assessment to measure learning during a tutoring session about the human body's circulatory system. Before and after the tutoring session, participants were asked to draw and explain the path of blood through the circulatory system on a sheet of paper with an outline of the human body. To analyze the drawings made by participants, Chi et al. developed seven different mental models. Six of these seven models had different degrees of errors. All seven models were ranked from the most naïve "No Loop" model to the most accurate and complete "Double Loop-2" model. Using this ranking of mental models, pre- and post-test drawings were analyzed in two distinct ways to measure the effectiveness of the tutoring session. First, the authors counted how many students had the most accurate "Double Loop-2" model before the tutoring session (0/11 students) and after the session (8/11 students). Second, the authors computed the average number of mental model shifts per individual student. For example, students who drew the most naïve "No Loop" model during the pre-test and the most complex "Double Loop-2" model during the post-test received a score of 6 (equal to the number of mental model "upshifts" from the pre-test to the post-test).

5.3. Comparative judgment

Assessing conceptual learning is challenging. One approach is to develop multiple-choice questions that evaluate a learner's understanding of a concept. However, developing such multiple-choice questions requires careful thought and validation. Another approach is to develop open-ended questions that are graded using detailed rubrics. However, this requires rubrics that result in acceptable levels of intercoder agreement. Jones, Inglis, Gilmore, and Hodgen (2013) proposed an alternative approach to conceptual learning assessment referred to as *comparative judgment*.

The comparative judgment approach proceeds as follows. First, learners are asked an open-ended question designed to assess their understanding of a particular concept. Second, multiple domain experts are asked to evaluate *pairs* of responses. Each expert is presented with pairs of responses and asked to determine which one conveys a deeper understanding of the concept (ties are not allowed). Importantly, domain experts are *not* given a rubric. Instead, they rely solely on their expert knowledge of the concept. Finally, a single ranking of responses (from 'best' to 'worst') is generated from these redundant pairwise judgments.

The comparative judgment approach has three main advantages. First, assessors produce *pairwise* rather absolute judgments. It has long been argued that humans are better at comparing pairs of objects against one another than they are at scoring objects in isolation (Thurstone, 1927). Second, there is no need for a grading rubric. Instead, experts rely solely on their own subjective judgment. Jones et al. argued that this process allows for the grading to consider important criteria that are "difficult if not impossible to specify comprehensively [in a rubric]" (Jones et al., 2013, p. 115). Finally, a set of n items is associated with $\frac{n(n-1)}{2}$ pairs, which is typically a prohibitive number of pairwise assessments. However, the comparative judgment approach can leverage algorithms that selectively choose a much smaller number of pairs to judge in order to output a stable ranking (e.g., Pollitt (2012)).

Jones et al. (2013) used the comparative judgment technique to measure conceptual understanding of fractions within a group of students. The open-ended assessment asked participants to order a set of fractions from smallest to largest and explain their method for doing this. Pairs of responses were judged by eight domain experts (i.e., current and former math educators). Results found the final ranking of responses to be highly stable. For example, results found very similar rankings by considering pairwise preferences from different *subsets* of experts. Bisson, Gilmore, Inglis, and Jones (2016) also validated the comparative judgment approach. Here, the authors tested conceptual understanding of three math concepts: p-values, derivatives, and letters in algebra. Participants' conceptual understanding was assessed using the comparative judgment approach and other well-established instruments (e.g., the Calculus Concept Inventory test (Epstein, 2007)). Results from the comparative judgment approach were largely consistent with those from other instruments.

5.4. Coordinating assessments to measure learning from different perspectives

A single type of learning assessment is often insufficient to capture everything a student has learned. Pellegrino argued that "no single test score can be considered a definitive measure of a student's competence" and that "multiple measures enhance the validity and fairness of the inferences drawn by giving students various ways and opportunities to demonstrate their competence" (Pellegrino, 2014, p. 246). Additionally, Pellegrino asserted that multiple assessment types offer more evidence that higher test scores represent learning gains versus a narrow understanding of specific test material.

As an exemplary multiple-assessment study, McNeil and Alibali (2000) used different types of assessments to measure math-related learning by children (i.e., 4th graders). Specifically, the authors implemented assessments to measure three types of learning as a result of an instructional session: (1) conceptual learning, (2) procedural learning, and (3) transfer of learning.

To measure conceptual learning, McNeil and Alibali used a combination of multiple-choice and short-answer questions. The conceptual knowledge assessment involved three types of tasks: (1) classify 15 equations into standard form (e.g., $7 - 4 = 3$) or non-standard form (e.g., $8 = 2 + 6$), (2) explain the meaning of the equal sign in your own words, and (3) solve 3 equivalence problems (e.g., $a + b + c = a + \underline{?}$). This assessment was administered immediately before, immediately after, and two weeks after the instructional session. Most questions had a single correct answer. Responses to the "equal sign question" were scored as correct if they demonstrated a *relational* understanding of the equal sign (e.g., "it means same as" versus "it means the answer").

To measure procedural learning, McNeil and Alibali asked participants to solve three equivalence problems (e.g., $a+b+c = a+\underline{?}$). After completing each problem, students were asked to explain how they arrived at their answer. This assessment was also administered immediately before, immediately after, and two weeks after the instructional session. This assessment was scored in two different ways. First, the equivalence problems had predefined correct answers that were scored accordingly. Second, students'

explanations were analyzed using qualitative techniques. The authors identified three correct and four incorrect strategies used by participants. To assess procedural learning, participants were scored based on their use of correct versus incorrect strategies.

Finally, to measure transfer of learning, McNeil and Alibali used an additional set of equivalence problems. These equivalence problems had subtle differences from the examples used during the instructional session. For example, the authors varied the position of the unknown value (e.g., $a+b+c = \underline{?}+a$). This assessment included six questions. Some of these were designed to be intentionally difficult because the problem-solving procedures taught during the instructional session could not be *directly* applied. Specifically, these questions introduced a new addend on the right (e.g., $a+b+c = d+\underline{?}$). The transfer of learning assessment was administered immediately after and two weeks after the instructional session. The assessment was scored in two different ways. First, an overall transfer score was calculated by summing the number of correct solutions from the six equivalence problems. Second, the authors analyzed problem-solving strategies implemented during the intentionally difficult problems. Participants were assessed based on their ability to generate a new and generalizable problem-solving procedure (e.g., $a+b+c = d+\underline{a+b+c-d}$).

5.5. Opportunities for search-as-learning

Studies outside of search-as-learning have used different types of assessment techniques: task performance, mental models, comparative judgment, and assessment combinations that target different *dimensions* of learning. Next, we propose three actionable steps for future research in search-as-learning.

First, search-as-learning studies have not yet considered assessments based on task performance and mental models. Importantly, these types of assessments can be used to measure learning with respect to specific knowledge types—task performance can be used to measure procedural knowledge gains and mental models can be used to measure conceptual knowledge gains. Second, many search-as-learning studies have used open-ended assessments to measure learning. Open-ended assessments have many benefits, but one critical challenge—the grading rubric needs to be validated by measuring agreement between independent assessors. To help alleviate this challenge, future studies should borrow ideas from the comparative judgment technique (e.g., ranking assessments though pairwise comparisons from redundant assessors). Finally, search-as-learning studies should explore combinations of assessments that target different *dimensions* of learning. For example, McNeil and Alibali (2000) used a combination of assessments that targeted conceptual knowledge gains, procedural knowledge gains, and transfer of learning. Search-as-learning studies have combined assessment types such as multiple-choice and open-ended assessments. However, studies have mostly varied the assessment type to account for certain task manipulations (e.g., using multiple-choice for simple tasks and open-ended for complex tasks). Future research should consider using different assessment combinations to capture different dimensions of learning during the *same* task. For example, if a study is considering a novel tool to support learning, the researchers might consider whether the tool supports both procedural and conceptual knowledge gains through different assessments.

6. Reflections & recommendations for future research

Given the wide variety of learning assessments that exist, developing learning assessments for a search-as-learning study may seem like a daunting task. Pellegrino (2014) outlines three factors that should guide the development of learning assessments: (1) purpose (e.g., to assess individual learning or to evaluate an educational program); (2) context (e.g., classroom or large scale), and (3) practical constraints (e.g., resources and time). Although these factors *should* guide the assessment development process, assessments are often crafted based on a "[vague] description of what students are supposed to know and what they should be able to do" (Pellegrino, 2014, p. 239). Precise learning objectives are useful in developing assessments that are focused, targeted, and able to serve as a valid measure of intended learning. In this section, we discuss learning assessment guidelines. Additionally, we discuss practical factors to consider and strategies for mitigating drawbacks from specific types of assessments.

Developing a learning assessment begs the question: What do we mean by *learning*? Answering this question can be difficult because learning is inherently multidimensional. In this section, we propose four dimensions to consider and provide methods to develop assessments for each dimension. The first two dimensions (cognitive process and knowledge type) can be defined using the A&K taxonomy (Anderson et al., 2001) (Section 2.1). The second two dimensions involve retention and transfer of learning.

6.1. Learning assessment development—Identifying cognitive process and knowledge type

The first step in learning assessment development is to precisely define the learning objective(s) study participants are being asked to accomplish. What precisely is the participant being asked to learn during a specific search task? As described in Section 2.1, the A&K taxonomy situates learning objectives at the intersection of two orthogonal dimensions: cognitive process and knowledge type.

The cognitive process dimension defines the types of cognitive activities associated with the objective. In other words, the cognitive process dimension defines the types of mental activities the searcher will be able to perform if the learning objective is met. Cognitive processes range from simple to complex: remember, understand, apply, analyze, evaluate, and create. If a remember objective is met, the learner will be able to recall information verbatim. If an understand objective is met, the learner will be able to explain information in their own words or identify examples of a construct. If an apply objective is met, the learner will be able to execute a process. If an analyze objective is met, the learner will be able to explain relations (e.g., similarities and differences) between elements. If an evaluate objective is met, the learner will be able to critique or prioritize elements. Finally, if a create objective is met, the learner will be able to generate a new solution to a problem or organize information using a new representation.

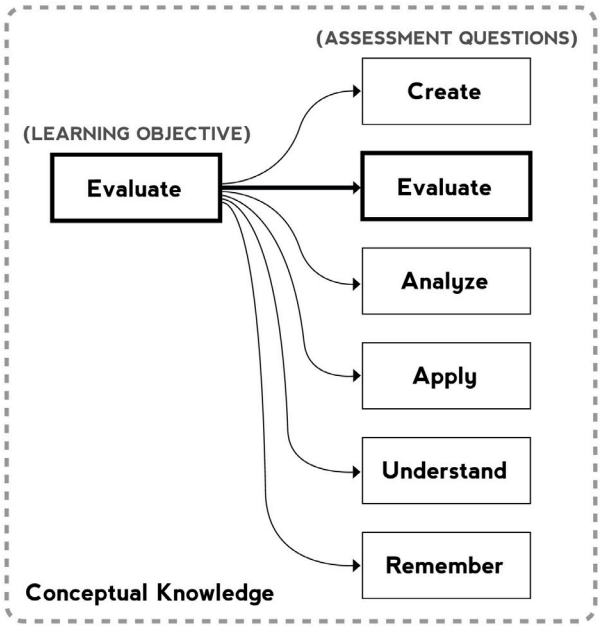

Fig. 1. The A&K taxonomy can be used to develop multiple types of assessment questions for a single learning objective. The learning objective and all assessment questions share the same knowledge type.

The knowledge type dimension defines the type of knowledge associated with the objective: factual, conceptual, procedural, and metacognitive knowledge. The first three knowledge types relate to external knowledge (i.e., knowledge about the outside world). Factual knowledge relates to isolated bits of objective information. Conceptual knowledge relates to concepts, categories, theories, principles, schemas, and models. Procedural knowledge relates to knowledge about how to perform a task. The last knowledge type (metacognitive) looks inward and relates to self-knowledge about one's own cognition.

Many studies have leveraged the A&K taxonomy to develop learning-oriented search tasks (Capra et al., 2015; Ghosh et al., 2018; Jansen et al., 2009; Kalyani & Gadiraju, 2019; Kelly et al., 2015; Urgo et al., 2020; Wu et al., 2012). Most of these studies have leveraged the cognitive process dimension and ignored the knowledge type dimension. As one exception, Urgo et al. (2020) developed search tasks with objectives that varied along three cognitive processes (apply, evaluate, create) *and* three knowledge types (factual, conceptual, procedural).

Anderson and Krathwohl argue that learning objectives can be defined as a verb–noun combination. The verb defines the cognitive process and the noun defines the knowledge type of the objective. For example, consider the following learning objective: "The learner will be able to *judge* which *principle* best explains lift acting on an airplane's wing: Bernoulli's principle or Newton's third law of motion". Using the A&K taxonomy, the verb of the objective, *judge*, can be directly mapped to the cognitive process of evaluate. Similarly, the noun of the objective, *principle*, can be directly mapped to conceptual knowledge. Another example objective might be: "The learner will be able to *use* the HeapSort *algorithm* to sort a list of numbers". The verb *use* can be directly mapped to the cognitive process of apply and the noun *algorithm* can be directly mapped to procedural knowledge. Using the A&K taxonomy, learning objectives can be easily modified to become assessment questions. For example, after gathering information, a searcher could be asked: "*Judge* which *principle* (Bernoulli's Principle or Newton's third law of motion) best explains lift and provide a justification". or "Sort the numbers below *using* the Heapsort *algorithm* and show each step"..

In the previous examples, each assessment question aligns perfectly with the objective. Importantly, each question asks the searcher to *demonstrate* that they achieved the exact learning objective set forth in the task description. Given a specific objective, searchers are likely to engage in a variety of cognitive processes as they gather and engage with information. The A&K taxonomy can also be leveraged to assess the *breadth* of cognitive processes a searcher can successfully engage in after searching. This assessment approach measures whether a learner can engage in cognitive processes of *varying* complexity within the *same* knowledge type (shown in Fig. 1). For example, consider the objective of judging whether Bernoulli's principle or Newton's third law of motion best explains lift. Given this 'evaluate/conceptual' objective, one could also assess whether a searcher can successfully engage in simpler or more complex cognitive processes within the same (conceptual) domain. For example, a remember question could ask the learner to identify the correct definition of Bernoulli's principle from a predefined set. An understand question could ask the learner to explain Bernoulli's principle in their own words. An apply question could ask the learner to use Bernoulli's principle to explain lift. An analyze question could ask the learner to differentiate between Bernoulli's principle and Newton's third law of motion. An evaluate question could ask the learner to explain which principle/law best explains lift (i.e., the primary objective). Finally, a create question could ask the learner to generate their own diagram of Bernoulli's principle applied to lift.

6.2. Learning assessment development—Capturing learning retention and transfer of learning

Beyond cognitive process and knowledge type, learning assessment methods can (and should) consider two additional dimensions: learning retention and transfer of learning. We believe researchers should capture learning retention and transfer of learning because they are both indicative of deeper understanding. Learning retention assessments indicate what information has been moved into a learner's long-term memory storage for future use. 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 fit the information into their existing knowledge structures) and meaning (i.e., the information is relevant to the learner). Therefore, learning retention shows which information has been deeply learned, having both sense and meaning to the learner. Transfer of learning assessments indicate the ability of the learner to use what has been learned in new situations. The ability to transfer knowledge is a core component of meaningful learning (Anderson et al., 2001). We argue that the goal of search-as-learning is to facilitate learning that goes beyond immediate recall and includes learning retention and transfer of learning, in which a learner will be able to (1) retain the information to be able to use it again in the future, and (2) be able to use the information in new situations. This section offers methods for developing and implementing retention and transfer of learning assessments in search-as-learning studies.

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 h after the search session before administering a learning assessment that is meant to capture learning retention. Relatively few studies in search-as-learning have administered assessments aimed at capturing learning retention of knowledge acquired during the search process (Heilman & Eskenazi, 2006; Qiu et al., 2020).[8] For example, in addition to an immediate post-task learning assessment, Qiu et al. (2020) measured retention of learning with a "long-term memory test" administered three days after the search session. The multiple-choice questions on the retention test were identical to those on the pre-test and immediate post-test.

Given the limited number of studies that have measured retention, many open questions remain. Namely, how is retention influenced by characteristics of the individual searcher, the task, or the system? For example, are levels of retention higher for users with more prior knowledge or during more cognitively complex tasks? And, how can we develop search tools to improve retention? Prior work in learning sciences has found several factors that influence learning retention that may be useful starting points for future search-as-learning work. Such factors include formative feedback (Shute, 2008), delayed feedback (Smith & Kimball, 2010), and spoken and written presentation of information (versus only spoken or only text), also known as "verbal redundancy" (Adesope & Nesbit, 2012).

Additionally, it is important for search-as-learning researchers to consider transfer of learning when developing learning assessments. 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. Similar to retention, only a few studies in search-as-learning 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. Outside of search-as-learning, McNeil and Alibali (2000) assessed transfer of learning by having participants complete equivalence math problems that had a different form than those used during the instructional session. During the instructional session, participants learned to solve problems of the form $a+b+c=a+\underline{?}$. During the assessment phase, participants were asked to solve problems of the form $a+b+c=d+\underline{?}$ (i.e., the problems introduced a new term on the right side of the equation). These transfer of learning questions required participants to go beyond the mathematical steps taught during the instructional session.

Transfer of learning has largely gone unexplored in search-as-learning. Anderson et al. (2001) argued that being able to transfer knowledge to new situations or problems is a core tenet of meaningful learning. Therefore, it is important for search-as-learning studies to include assessment questions or exercises that measure transfer of learning.

The A&K taxonomy can also be a useful framework to develop assessment materials to measure transfer of learning. Measuring transfer of learning may make more sense during search tasks with complex learning objectives. Let us consider learning-oriented search tasks with objectives associated with the cognitive processes of understand, apply, analyze, evaluate, and create (Fig. 2).

Imagine an objective that requires participants to explain how concept A is exemplified by (i.e., understand) example B. A transfer of learning question could then ask a participant to explain how concept A is exemplified by new example C. For instance, an objective could ask a participant why "The Temptation of Saint Anthony" by Dalí is an example of surrealism. Through searching, a participant might learn about features of surrealism that describe why the painting is an example of surrealism (e.g., the painting

[8] Outside of search-as-learning, prior studies have explored factors that influence a searcher's ability to remember specific functions of the search system (Liu, Wang, Mandal, & Shah, 2019; Moraveji, Russell, Bien, & Mease, 2011).

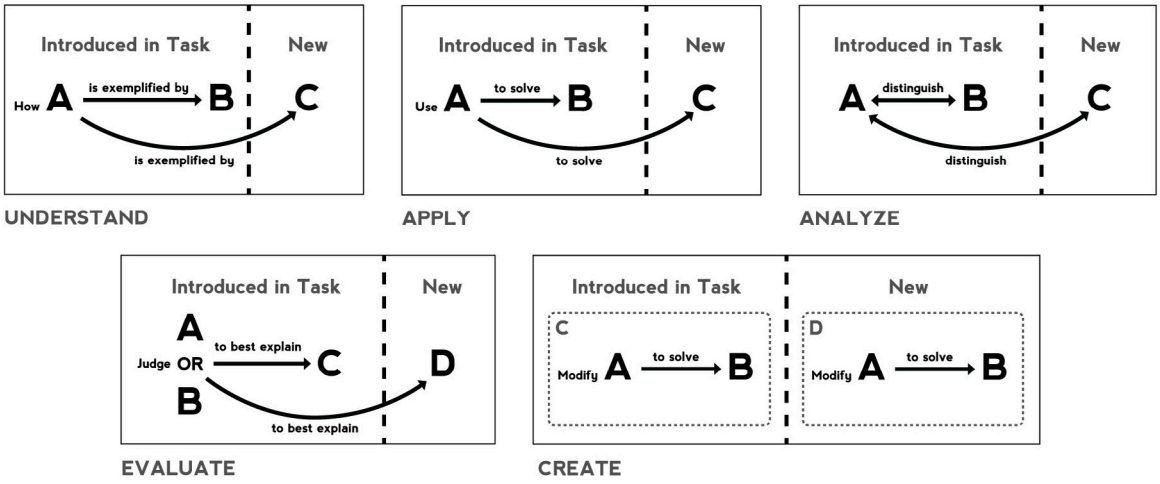

Fig. 2. Conceptualizing transfer of learning assessment items using the A&K taxonomy.

involves dream-like scenes and distorted figures). To test transfer of learning, a participant could be asked to describe why a painting they did *not* encounter during search, "Leonora in the Morning Light" by Ernst, exemplifies surrealism.

Imagine an objective that requires participants to use (i.e., apply) procedure A to solve problem B. A transfer of learning question might ask participants to use procedure A to solve problem C. For example, an objective could ask a participant to use factorization to solve the following problem, "Factor the following quadratic expression: $x^2 - x - 6$". By correctly applying factorization, the answer is $(x + 2)(x - 3)$. To test transfer of learning, a participant could be asked to use factorization to solve the problem, "Factor the following algebraic expression: $x^4 - 13x^2 + 36$". This type of problem requires the participant to apply factorization to arrive at the factors $(x^2 - 9)(x^2 - 4)$. At this stage, transfer of learning is necessary in the following two actions. First, the participant must recognize that the expression $(x^2 - 9)(x^2 - 4)$ can be further factored. Second, the participant must *again* apply factorization to find the factors $(x - 3)(x + 3)(x - 2)(x + 2)$.

Imagine an objective that requires participants to distinguish (i.e., analyze) between concepts A and B. A transfer of learning question might ask participants to distinguish between concept A and C. For example, a participant could be asked to distinguish between the artistic movements of surrealism and Dadaism. To test transfer of learning, the participant could be asked to distinguish between the artistic movement of surrealism and a different movement such as romanticism.

Imagine an objective that requires participants to decide (i.e., evaluate) whether concept A or B best explains phenomenon C. A transfer of learning question might ask participants to judge whether concept A or B best explains phenomenon D. For example, a participant could be asked to judge whether Bernoulli's principle or Newton's third law of motion best explains lift. To test transfer of learning, the participant could be asked to judge whether Bernoulli's principle or Newton's third law best explains a different phenomenon such as the thrust of a jet engine.

Finally, imagine a learning objective that requires participants to modify (i.e., create) procedure A to solve problem B under constraints C. A transfer of learning question might ask participants to modify A to solve B under constraints D. For example, a participant could be asked to modify a recipe for making crème brûlée under the constraints that it must be easy for kids to make. To test transfer of learning, the participant could be asked to modify the recipe to make crème brûlée under the constraints that it be made as quickly as possible.

In the above examples, the transfer of learning questions involve introducing a novel component (not directly part of the objective). One might then ask: What if participants lack prior knowledge of this novel component? To help address this, details about the novel component can be introduced as part of the assessment. For example, consider the above transfer of learning example involving analyzing. The participant is asked to distinguish between surrealism and Dadaism. To meet this objective, a searcher needs to gather information about both concepts. However, the transfer of learning question asks participants to distinguish between surrealism and romanticism. In order to remove prior knowledge as an obstacle, the question could include sufficient details about romanticism (e.g., romanticism artwork involves emotions, nature, and literal interpretations of the world) so that the participant can try to distinguish between concepts based on their acquired knowledge of surrealism.

Future search-as-learning studies should consider a searcher's ability to use what was learned to support new learning. Next, we propose five open questions for future research to explore. First, what are characteristics of the searcher that affect transfer of learning? Such factors may include domain knowledge, motivation, engagement, and effort. For example, a highly motivated searcher may be more likely to engage in strategies such as self-explanation, which deepen understanding. Strategies that deepen understanding are important components in successful transfer of learning (Chi & VanLehn, 2012). Second, how can we develop search tools to promote transfer of learning? Perhaps systems that promote transfer of learning *during* the search process are more likely to promote transfer of learning *after* the search process is complete. For example, suppose that a searcher is trying to understand how Bernoulli's principle applies to lift acting on an airplane's wing. A system could diversify examples used to

illustrate Bernoulli's principle (e.g., wings, sailboats, curveballs, etc.). Exposing searchers to different perspectives may lead to a greater ability to use what was learned to support new learning. Third, what is the relation between particular search behaviors and transfer of learning? As discussed, prior work in search-as-learning has investigated the relation between search behaviors and knowledge gains (Abualsaud, 2017; Bhattacharya & Gwizdka, 2019; Collins-Thompson et al., 2016; Gadiraju et al., 2018; Liu & Song, 2018; Lu & Hsiao, 2017; Palani et al., 2021; Yu et al., 2018). Investigating the relationship between search behaviors and transfer of learning may help researchers better understand signals of deeper, more meaningful learning. Fourth, what is the relation between serendipity and transfer of learning? During search, a serendipitous interaction involves investigating a direction that was not previously anticipated. This aspect of serendipity may promote transfer of learning. Specifically, during a serendipitous interaction, a searcher may be encouraged to use newly acquired knowledge in a *novel* and *unexpected* context. Prior work has investigated factors that may support serendipity during search, such as results diversification (Makri, Blandford, Woods, Sharples, & Maxwell, 2014; Taramigkou, Apostolou, & Mentzas, 2017). Future work in search-as-learning could explore the influence of such factors on transfer of learning during search. Finally, how is transfer of learning related to other types of learning outcomes? For example, is transfer of learning more closely related to outcomes associated with reading comprehension versus factual recall?

6.3. Practical considerations

In Sections 4 and 5, we enumerate different types of learning assessment used in search-as-learning studies and beyond. After clearly defining the type of learning to be assessed after a specific search task, it is important to reflect on practical considerations that might impact the choice of learning assessment. Table 3 can help guide the selection process based on practical considerations. Table 3 characterizes assessment types along eight dimensions. Three dimensions consider how easy it is to develop, administer, and grade the assessment. A fourth dimension considers how quickly the assessment can be graded to provide learners with timely feedback. A fifth dimension considers the assessment's reliability. Assessments can produce unreliable results because they are prone to guesswork (e.g., multiple-choice), are based on subjective perceptions (e.g., self-report), or are crude measures of learning (e.g., implicit behavioral measures). A sixth dimension considers how easy it is to compare test scores across participants. Multiple-choice test scores are easy to compare. Conversely, summary scores are more difficult to compare because grading is more subjective (i.e., involves qualitative coding) and writing skills may vary widely across participants. The final two dimensions consider whether the assessment is likely to capture *everything* the participant has learned in terms of breadth and depth. Open-ended assessments (e.g., summaries) typically have more coverage than closed-ended assessments (e.g., multiple-choice).

Next, we discuss three factors that are important to consider when choosing the type of assessment: time, setting, and cognitive complexity.

Time: Consider a researcher who is designing a lab study to investigate the impact of ongoing feedback from assessments administered during the search process. In this study, it seems important for the feedback to be administered quickly. Given this time constraint, multiple-choice, short-answer, and free recall assessments seem like viable options. These assessments can be graded quickly (even automatically) and can ask about material that is central to the assigned search task (known in advance).

Setting: Consider a researcher who is designing a study to investigate learning during genuine, longitudinal search tasks. In a naturalistic setting, learning assessments need to be open-ended because the search task is not known in advance. Given this constraint, summaries and open-ended questions seem like viable options. These types of assessments allow searchers to convey what they have learned across topics and levels of cognitive complexity. While grading requires more time and effort, it can be done after the data collection phase.

Cognitive Complexity: Consider a researcher who is designing a study to investigate the effects of task complexity on learning outcomes. In this case, the type of assessment may need to vary depending on the complexity of the task's objective. A simple objective (remember) may involve rote memorization. Multiple-choice, short answer, or free-recall questions can be used to test a participant's ability to recall information. A more complex objective (analyze) may involve decomposing a system and understanding the relations between its components. Mind maps or mental model diagrams can be used to test a participant's ability to convey accurate and nuanced relations between elements. Finally, a highly complex task (create) may involve generating a new solution to a problem. Open-ended questions can be used to test a participant's ability to describe a new solution to a problem or a novel representation of a domain. In general, open-ended (vs. closed-ended) assessments may be more appropriate during more complex tasks. Prior work has found that searchers with more complex objectives have more *divergent* search behaviors (e.g., issue different queries) when compared to each other (Kelly et al., 2015). In other words, during complex tasks, searchers tend to go in different directions. In this case, open-ended assessments may give searchers the opportunity to convey *unique* things they learned during the search process.

6.4. Mitigating drawbacks

Every type of assessment has drawbacks. However, strategies can be implemented to mitigate their impact. Here, we discuss strategies to mitigate the drawbacks of: (1) priming effects, (2) limited topical range, and (3) limited cognitive complexity range.

First, priming effects are a potential concern for certain types of assessments. For example, studies have used multiple-choice tests to measure learning before and after a search task. Pre-task multiple-choice tests risk revealing keywords that participants can use in their searches. In other words, pre-task assessments (particularly closed-ended ones) can give searchers an unintended "head start". This drawback can be mitigated by administering the pre-task assessment much earlier than the search session. Pardi et al. (2020) administered their multiple-choice pre-test one week before the search session.

Second, topical coverage can be quite limited with certain assessment types. For example, multiple-choice tests inquire about specific items learned, but may not capture all that was learned during the search session. This drawback can be mitigated by incorporating additional assessment types in order to capture breadth of learning. Pardi et al. (2020) addressed this problem by administering a multiple-choice test and an essay assessment. The essay assessment asked participants to explain how thunderstorms and lightning form. This allowed for a broader understanding of what was learned during the session. Similarly, Collins-Thompson et al. (2016) administered a short-answer assessment along with open-ended assessment questions that asked participants to write a summary of what was learned, generate an outline for a hypothetical paper, and enumerate questions they still had about the task topic.

Finally, some assessment types, such as multiple-choice tests, can be limited in the range of cognitive complexity that is measured. In Section 6.3, we advocated that open-ended assessments be used to measure a searcher's ability to engage in complex cognitive processes. Additionally, this drawback of closed-ended assessments can be mitigated by carefully constructing questions that require more complex cognitive processes to answer. For example, Freund et al. (2016) developed multiple-choice items that assessed both microstructural and macrostructural aspects of reading comprehension. Microstructural questions asked about information found in a *single* article and therefore tested the participant's ability to remember/understand. Macrostructural questions asked about common themes *across* articles and therefore tested the participant's ability to analyze (i.e., infer relations).

7. Conclusion

Learning assessment is a complex and challenging component of search-as-learning research. In order to address the challenge of learning assessment selection and development, we reviewed the learning assessments used in search-as-learning studies, investigated new approaches from outside search-as-learning, and offered guidelines and practical considerations for researchers when developing learning assessments. First, we categorized learning assessments from prior work in search-as-learning into nine learning assessment types. Within each assessment type, we explored how assessments were administered and detailed how assessments were scored. We also discussed the potential benefits and drawbacks of each assessment type. Categorizing and detailing learning assessments from search-as-learning studies has shed light on the many components that make up a learning outcome measurement. Given the variation observed in reporting, search-as-learning work would benefit from detailed descriptions of *all parts* of the assessment process. Specifically, it is important to elucidate—(1) the measurement of prior knowledge; (2) the complete assessment or example questions from the assessment; (3) a detailed guide of the grading process; and (4) the calculation process of the final learning metric. Each of these factors are important to communicate because they impact the interpretation of results in search-as-learning studies.

Apart from search-as-learning studies, we discussed three assessment techniques used in education and psychology (i.e., task performance, mental models, and comparative judgment). These assessment techniques may help search-as-learning researchers to alleviate grading challenges, capture more specific or subtle aspects of learning, and provide greater coverage of learning. Additionally, this review provides search-as-learning researchers with practical guidelines for developing learning assessments. First, we showed how the A&K taxonomy can be used to precisely define learning objectives and develop assessment materials that align with the objective(s). Next, we discussed learning retention and transfer of learning, and provided methods for capturing each. Although Anderson & Krathwohl highlight their importance, learning retention and transfer of learning remain largely unexplored in search-as-learning. We argue that *both* are important areas of inquiry for future work. We also provided guidance for researchers to select a type of assessment based on research goals and practical considerations. For example, consider a search-as-learning study that involves understand-level learning by crowdsourced workers as participants. In this case, a closed-ended multiple-choice test may be appropriate for several reasons. First, multiple-choice tests can be carefully crafted to measure understand-level learning. The questions can ask about definitions and examples, and include distractor options that are rooted in common misconceptions. Second, while developing the assessment may require substantial effort, the grading of the assessment is quick and easy. It could even be automated. This may be appealing if the study involves many participants. Finally, crowdsourced workers may be less inclined to satisfice on an assessment that has predefined correct answers versus an open-ended assessment. All assessment types have pros and cons. Therefore, we concluded our review by outlining ways to mitigate the drawbacks of certain assessment types.

Acknowledgments

This work was supported by National Science Foundation (NSF), United States of America grant IIS-1718295. Any opinions, findings, conclusions, and recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the sponsor.

References

Abualsaud, Mustafa (2017). Learning factors and determining document-level satisfaction in search-as-learning (MA thesis), Waterloo, Ontario, Canada: University of Waterloo.

Adesope, Olusola O., & Nesbit, John C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology*, [ISSN: 0022-0663] *104*(1), 250–263. http://dx.doi.org/10.1037/a0026147, URL: https://auth.lib.unc.edu/ezproxy_auth.php?url=https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2011-25165-001&site=ehost-live&scope=site.

Allan, James, Croft, Bruce, Moffat, Alistair, & Sanderson, Mark (2012). Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in lorne. *ACM SIGIR Forum*, [ISSN: 01635840] *46*(1), 2. http://dx.doi.org/10.1145/2215676.2215678, URL: http://dl.acm.org/citation.cfm?doid=2215676.2215678.

Anderson, Lorin W., Krathwohl, David R., Airasian, Peter W., Cruikshank, Kathleen A., Mayer, Richard E., Pintrich, Paul R., et al. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives, Abridged Edition* (1st ed.). New York: Pearson, ISBN: 978-0-8013-1903-7.

Bhattacharya, Nilavra, & Gwizdka, Jacek (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* (pp. 63–71). Glasgow Scotland UK: ACM, ISBN: 978-1-4503-6025-8, http://dx.doi.org/10.1145/3295750.3298926, URL: https://dl.acm.org/doi/10.1145/3295750.3298926.

Biggs, John B., & Collis, Kevin F. (Eds.), (1982). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. Australia: Academic Press, ISBN: 978-0-12-097552-5, http://dx.doi.org/10.1016/B978-0-12-097552-5.50001-6, URL: https://www.sciencedirect.com/science/article/pii/B9780120975525500016.

Bisson, Marie-Josée, Gilmore, Camilla, Inglis, Matthew, & Jones, Ian (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, [ISSN: 2198-9753] *2*(2), 141–164. http://dx.doi.org/10.1007/s40753-016-0024-3.

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

Boulet, John R. (2008). Summative assessment in medicine: The promise of simulation for high-stakes evaluation. *Academic Emergency Medicine*, [ISSN: 1553-2712] *15*(11), 1017–1024. http://dx.doi.org/10.1111/j.1553-2712.2008.00228.x, URL: http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1553-2712.2008.00228.x.

Câmara, Arthur, Roy, Nirmal, Maxwell, David, & Hauff, Claudia (2021). Searching to learn with instructional scaffolding. In *Proceedings of the 2021 conference on human information interaction and retrieval* (pp. 209–218). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-8055-3, http://dx.doi.org/10.1145/3406522.3446012.

Capra, Robert, Arguello, Jaime, Crescenzi, Anita, & Vardell, Emily (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* (pp. 23–32). New York, NY, USA: ACM, ISBN: 978-1-4503-3621-5, http://dx.doi.org/10.1145/2766462.2767741, URL: http://doi.acm.org/10.1145/2766462.2767741.

Capra, Robert, Arguello, Jaime, O'Brien, Heather, Li, Yuan, & Choi, Bogeum (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* (pp. 445–454). New York, NY, USA: ACM, ISBN: 978-1-4503-5657-2, http://dx.doi.org/10.1145/3209978.3210047, URL: http://doi.acm.org/10.1145/3209978.3210047.

Chi, Yu, Han, Shuguang, He, Daqing, & Meng, Rui (2016). Exploring knowledge learning in collaborative information seeking process. In *CEUR workshop proceedings: Vol. 1647*. (p. 5).

Chi, Michelene T. H., Siler, Stephanie A., Jeong, Heisawn, Yamauchi, Takashi, & Hausmann, Robert G. (2001). Learning from human tutoring. *Cognitive Science*, [ISSN: 1551-6709] *25*(4), 471–533. http://dx.doi.org/10.1207/s15516709cog2504_1, URL: http://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog2504_1, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog2504_1.

Chi, Michelene T. H., & VanLehn, Kurt A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist*, [ISSN: 0046-1520] *47*(3), 177–188. http://dx.doi.org/10.1080/00461520.2012.695709.

Colbert-Getz, Jorie M., Fleishman, Carol, Jung, Julianna, & Shilkofski, Nicole (2013). How do gender and anxiety affect students' self-assessment and actual performance on a high-stakes clinical skills examination? *Academic Medicine*, [ISSN: 1040-2446] *88*(1), 44–48. http://dx.doi.org/10.1097/ACM.0b013e318276bcc4, URL: https://journals.lww.com/academicmedicine/FullText/2013/01000/How_Do_Gender_and_Anxiety_Affect_Students_.18.aspx.

Collins-Thompson, Kevyn, Hansen, Preben, & Hauff, Claudia (2017). *Search as learning (Dagstuhl seminar 17092)*. http://dx.doi.org/10.4230/dagrep.7.2.135,

Collins-Thompson, Kevyn, Rieh, Soo Young, Haynes, Carl C., & Syed, Rohail (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* (pp. 163–172). New York, NY, USA: ACM, ISBN: 978-1-4503-3751-9, http://dx.doi.org/10.1145/2854946.2854972, URL: http://doi.acm.org/10.1145/2854946.2854972.

Cooper, Harris M. (1998). *Synthesizing research: a guide for literature reviews* (3rd ed.). Calif.: Thousand Oaks, URL: http://hdl.handle.net/2027/mdp.39015040152228.

Davies, Sarah, Butcher, Kirsten R., & Stevens, Corey (2013). Self-regulated learning with graphical overviews: When spatial information detracts from learning. In *Proceedings of the annual meeting of the cognitive science society: Vol. 35* (p. 7).

Demaree, Diego, Jarodzka, Halszka, Brand-Gruwel, Saskia, & Kammerer, Yvonne (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* (pp. 373–377). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-6892-6, http://dx.doi.org/10.1145/3343413.3378000.

Dunning, David, Johnson, Kerri, Ehrlinger, Joyce, & Kruger, Justin (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, [ISSN: 0963-7214] *12*(3), 83–87. http://dx.doi.org/10.1111/1467-8721.01235.

Eickhoff, Carsten, Gwizdka, Jacek, Hauff, Claudia, & He, Jiyin (2017). Introduction to the special issue on search as learning. *Information Retrieval Journal*, [ISSN: 1573-7659] *20*(5), 399–402. http://dx.doi.org/10.1007/s10791-017-9315-9.

Epstein, Jerome (2007). Development and validation of the calculus concept inventory. In *Proceedings of the ninth international conference on mathematics education in a global community: Vol. 9* (p. 6).

Fink, L. Dee (2013). *Jossey-bass higher and adult education series*: *Revised and updated edition*, *Creating significant learning experiences : an integrated approach to designing college courses*. San Francisco: Jossey-Bass, ISBN: 978-1-118-12425-3, URL: https://auth.lib.unc.edu/ezproxy_auth.php?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=632420&site=ehost-live.

Freund, Luanne, He, Jiyin, Gwizdka, Jacek, Kando, Noriko, Hansen, Preben, & Rieh, Soo Young (2014). Searching as learning (SAL) workshop 2014. In *Proceedings of the 5th information interaction in context symposium* (p. 7). New York, NY, USA: ACM, ISBN: 978-1-4503-2976-7, http://dx.doi.org/10.1145/2637002.2643203, URL: http://doi.acm.org/10.1145/2637002.2643203.

Freund, Luanne, Kopak, Rick, & O'Brien, Heather (2016). The effects of textual environment on reading comprehension: Implications for searching as learning. *Journal of Information Science*, [ISSN: 0165-5515] *42*(1), 79–93. http://dx.doi.org/10.1177/0165551515614472.

Gadiraju, Ujwal, Yu, Ran, Dietze, Stefan, & Holtz, Peter (2018). Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 2–11). New York, NY, USA: ACM, ISBN: 978-1-4503-4925-3, http://dx.doi.org/10.1145/3176349.3176381, URL: http://doi.acm.org/10.1145/3176349.3176381.

Ghosh, Souvick, Rath, Manasa, & Shah, Chirag (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* (pp. 22–31). New York, NY, USA: ACM, ISBN: 978-1-4503-4925-3, http://dx.doi.org/10.1145/3176349.3176386, URL: http://doi.acm.org/10.1145/3176349.3176386.

González-Betancor, Sara M, Bolívar-Cruz, Alicia, & Verano-Tacoronte, Domingo (2019). Self-assessment accuracy in higher education: The influence of gender and performance of university students. *Active Learning in Higher Education*, [ISSN: 1469-7874] *20*(2), 101–114. http://dx.doi.org/10.1177/1469787417735604.

Gwizdka, Jacek, Hansen, Preben, Hauff, Claudia, He, Jiyin, & Kando, Noriko (2016). Search as learning (SAL) workshop 2016. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 1249–1250). New York, NY, USA: ACM, ISBN: 978-1-4503-4069-4, http://dx.doi.org/10.1145/2911451.2917766, URL: http://doi.acm.org/10.1145/2911451.2917766.

Hansen, Preben, & Rieh, Soo Young (2016). Editorial: Recent advances on searching as learning: An introduction to the special issue. *Journal of Information Science*, [ISSN: 0165-5515, 1741-6485] *42*(1), 3–6. http://dx.doi.org/10.1177/0165551515614473, URL: http://journals.sagepub.com/doi/10.1177/0165551515614473.

Haskell, Robert E. (2001). *Transfer of learning: cognition, instruction, and reasoning*. San Diego, CA, US: Academic Press, ISBN: 978-0-12-330595-4, http://dx.doi.org/10.1016/B978-012330595-4/50003-2.

Hawkins, Richard, Gaglione, Margaret MacKrell, LaDuca, Tony, Leung, Cynthia, Sample, Laurel, Gliva-McConvey, Gayle, et al. (2004). Assessment of patient management skills and clinical skills of practising doctors using computer-based case simulations and standardised patients. *Medical Education*, [ISSN: 1365-2923] *38*(9), 958–968. http://dx.doi.org/10.1111/j.1365-2929.2004.01907.x, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2929.2004.01907.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2929.2004.01907.x.

Heilman, Michael, Collins-thompson, Kevyn, Callan, Jamie, & Eskenazi, Maxine (2010). Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 73–98.

Heilman, Michael, & Eskenazi, Maxine (2006). Language learning: Challenges for intelligent tutoring systems. In *Proceedings of the workshop on intelligent tutoring systems for ill-defined domains*.

Hersh, William R., Elliot, Diane L., Hickam, David H., Wolf, Stephanie L., & Molnar, Anna (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* (pp. 164–170). Seattle, Washington, United States: ACM Press, ISBN: 978-0-89791-714-8, http://dx.doi.org/10.1145/215206.215355, URL: http://portal.acm.org/citation.cfm?doid=215206.215355.

Hornbæk, Kasper, & Frøkjær, Erik (2003). Reading patterns and usability in visualizations of electronic documents. *ACM Transactions on Computer-Human Interaction*, [ISSN: 1073-0516] *10*(2), 119–149. http://dx.doi.org/10.1145/772047.772050.

von Hoyer, Johannes, Pardi, Georg, Kammerer, Yvonne, & Holtz, Peter (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* (pp. 3–10). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-6919-0, http://dx.doi.org/10.1145/3347451.3356730.

Jansen, Bernard J., Booth, Danielle, & Smith, Brian (2009). Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, [ISSN: 0306-4573] *45*(6), 643–663. http://dx.doi.org/10.1016/j.ipm.2009.05.004, URL: http://www.sciencedirect.com/science/article/pii/S030645730900051X.

Jones, Ian, Inglis, Matthew, Gilmore, Camilla, & Hodgen, Jeremy (2013). Measuring conceptual understanding: the case of fractions. URL:.

Jones, Natalie A., Ross, Helen, Lynam, Timothy, Perez, Pascal, & Leitch, Anne (2011). Mental models: An interdisciplinary synthesis of theory and methods. *Ecology and Society*, [ISSN: 1708-3087] *16*(1), URL: http://www.jstor.org/stable/26268859.

Kalyani, Rishita, & Gadiraju, Ujwal (2019). Understanding user search behavior across varying cognitive levels. In *Proceedings of the 30th ACM conference on hypertext and social media* (pp. 123–132). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-6885-8, http://dx.doi.org/10.1145/3342220.3343643.

Kammerer, Yvonne, Nairn, Rowan, Pirolli, Peter, & Chi, Ed H. (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* (pp. 625–634). Boston, MA, USA: Association for Computing Machinery, ISBN: 978-1-60558-246-7, http://dx.doi.org/10.1145/1518701.1518797.

Kelly, Diane, Arguello, Jaime, Edwards, Ashlee, & Wu, Wan-ching (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* (pp. 101–110). New York, NY, USA: ACM, ISBN: 978-1-4503-3833-2, http://dx.doi.org/10.1145/2808194.2809465, URL: http://doi.acm.org/10.1145/2808194.2809465.

Kelly, Diane, & Sugimoto, Cassidy R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, [ISSN: 1532-2890] *64*(4), 745–770. http://dx.doi.org/10.1002/asi.22799, URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.22799.

Kim, Kibum, Turner, Scott A., & Pérez-Quiñones, Manuel A. (2009). Requirements for electronic note taking systems: A field study of note taking in university classrooms. *Education and Information Technologies*, [ISSN: 1573-7608] *14*(3), 255–283. http://dx.doi.org/10.1007/s10639-009-9086-z.

Koedinger, K. R., & Anderson, J. R. (1993). Effective use of intelligent software in high school math classrooms. http://dx.doi.org/10.1184/R1/6614585.v1, URL:.

Kruger, Justin, & Dunning, David (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, [ISSN: 1939-1315] *77*(6), 1121–1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121.

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

Lei, Pei-Lan, Sun, Chuen-Tsai, Lin, Sunny S. J., & Huang, Tsung-Kuan (2015). Effect of metacognitive strategies and verbal-imagery cognitive style on biology-based video search and learning performance. *Computers & Education*, [ISSN: 0360-1315] *87*, 326–339. http://dx.doi.org/10.1016/j.compedu.2015.07.004, URL: https://www.sciencedirect.com/science/article/pii/S0360131515300099.

Liu, Jiqun (2021). Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management*, [ISSN: 0306-4573] *58*(3), Article 102522. http://dx.doi.org/10.1016/j.ipm.2021.102522, URL: https://www.sciencedirect.com/science/article/pii/S0306457321000315.

Liu, Jingjing, Belkin, Nicholas J., Zhang, Xiangmin, & Yuan, Xiaojun (2013). Examining users' knowledge change in the task completion process. *Information Processing & Management*, [ISSN: 0306-4573] *49*(5), 1058–1074. http://dx.doi.org/10.1016/j.ipm.2012.08.006, URL: http://www.sciencedirect.com/science/article/pii/S0306457312001136.

Liu, Hanrui, Liu, Chang, & Belkin, Nicholas J. (2019). Investigation of users' knowledge change process in learning-related search tasks. *Proceedings of the Association for Information Science and Technology*, [ISSN: 2373-9231] *56*(1), 166–175. http://dx.doi.org/10.1002/pra2.63, URL: http://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.63.

Liu, Chang, & Song, Xiaoxuan (2018). How do information source selection strategies influence users' learning outcomes'. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 257–260). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-4925-3, http://dx.doi.org/10.1145/3176349.3176876.

Liu, Jiqun, Wang, Yiwei, Mandal, Soumik, & Shah, Chirag (2019). Exploring the immediate and short-term effects of peer advice and cognitive authority on Web search behavior. *Information Processing & Management*, [ISSN: 03064573] *56*(3), 1010–1025. http://dx.doi.org/10.1016/j.ipm.2019.02.011, URL: https://linkinghub.elsevier.com/retrieve/pii/S030645731830236X.

Lu, Yihan, & Hsiao, I-Han (2017). Personalized information seeking assistant (PiSA): from programming information seeking to learning. *Information Retrieval Journal*, *20*(5), 433–455. http://dx.doi.org/10.1007/s10791-017-9305-y, [ISSN: 1386-4564, 1573-7659] URL: http://link.springer.com/10.1007/s10791-017-9305-y.

Makri, Stephann, Blandford, Ann, Woods, Mel, Sharples, Sarah, & Maxwell, Deborah (2014). "Making my own luck": Serendipity strategies and how to support them in digital information environments. *Journal of the Association for Information Science and Technology*, [ISSN: 2330-1643] *65*(11), 2179–2194. http://dx.doi.org/10.1002/asi.23200, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23200, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23200.

McNeil, Nicole M., & Alibali, Martha W. (2000). Learning mathematics from procedural instruction: Externally imposed goals influence what is learned. *Journal of Educational Psychology*, [ISSN: 0022-0663] *92*(4), 734–744. http://dx.doi.org/10.1037/0022-0663.92.4.734, URL: https://auth.lib.unc.edu/ezproxy_auth.php?url=http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2000-16403-013&site=ehost-live&scope=site.

Moraes, Felipe, Putra, Sindunuraga Rikarno, & Hauff, Claudia (2018). Contrasting search as a learning activity with instructor-designed learning. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 167–176). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-6014-2, http://dx.doi.org/10.1145/3269206.3271676.

Moraveji, Neema, Russell, Daniel, Bien, Jacob, & Mease, David (2011). Measuring improvement in user search performance resulting from optimal search tips. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information* (p. 355). Beijing, China: ACM Press, ISBN: 978-1-4503-0757-4, http://dx.doi.org/10.1145/2009916.2009966, URL: http://portal.acm.org/citation.cfm?doid=2009916.2009966.

Moreno-Marcos, P. M., Pong, T., Muñoz Merino, P. J., & Kloos, C. Delgado (2020). Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access*, [ISSN: 2169-3536] *8*, 5264–5282. http://dx.doi.org/10.1109/ACCESS.2019.2963503, Conference Name: IEEE Access.

Nelson, Les, Held, Christoph, Pirolli, Peter, Hong, Lichan, Schiano, Diane, & Chi, Ed H (2009). With a little help from my friends: Examining the impact of social annotations in sensemaking tasks. (p. 4).

Nersessian, Nancy J. (2002). The cognitive basis of model-based reasoning in science. In *The cognitive basis of science* (pp. 133–153). New York, NY, US: Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511613517.008, ISBN: 978-0-521-81229-0 978-0-521-01177-8.

Nissen, Jayson M., Talbot, Robert M., Nasim Thompson, Amreen, & Van Dusen, Ben (2018). Comparison of normalized gain and cohen's *d* for analyzing gains on concept inventories. *Physical Review Physics Education Research*, *14*, Article 010115. http://dx.doi.org/10.1103/PhysRevPhysEducRes.14.010115, URL: https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.14.010115.

O'Brien, Heather L., Kampen, Andrea, Cole, Amelia W., & Brennan, Kathleen (2020). The role of domain knowledge in search as learning. In *Proceedings of the 2020 conference on human information interaction and retrieval* (pp. 313–317). Vancouver BC, Canada: Association for Computing Machinery, ISBN: 978-1-4503-6892-6, http://dx.doi.org/10.1145/3343413.3377989.

Palani, Srishti, Ding, Zijian, MacNeil, Stephen, & Dow, Steven P. (2021). The "active search" hypothesis: How search strategies relate to creative learning. In *Proceedings of the 2021 conference on human information interaction and retrieval* (pp. 325–329). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-8055-3, http://dx.doi.org/10.1145/3406522.3446046.

Pardi, Georg, von Hoyer, Johannes, Holtz, Peter, & Kammerer, Yvonne (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* (pp. 378–382). Vancouver BC Canada: ACM, ISBN: 978-1-4503-6892-6, http://dx.doi.org/10.1145/3343413.3378001, URL: https://dl.acm.org/doi/10.1145/3343413.3378001.

Pellegrino, James W. (2014). A learning sciences perspective on the design and use of assessment in education. In R. Keith Sawyer (Ed.), *Cambridge handbooks in psychology*, *The cambridge handbook of the learning sciences* (2 ed.). (pp. 233–252). Cambridge: Cambridge University Press, ISBN: 978-1-107-62657-7, http://dx.doi.org/10.1017/CBO9781139519526.015, URL: https://www.cambridge.org/core/books/cambridge-handbook-of-the-learning-sciences/learning-sciences-perspective-on-the-design-and-use-of-assessment-in-education/CAF1F3B99E672BB41B95DE564400B1B0.

Pennycook, Gordon, Ross, Robert M., Koehler, Derek J., & Fugelsang, Jonathan A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review*, [ISSN: 1531-5320] *24*(6), 1774–1784. http://dx.doi.org/10.3758/s13423-017-1242-7.

Persky, Adam M., Lee, Edward, & Schlesselman, Lauren S. (2020). Perception of learning versus performance as outcome measures of educational research. *American Journal of Pharmaceutical Education*, *84*(7), ajpe7782. http://dx.doi.org/10.5688/ajpe7782, [ISSN: 0002-9459, 1553-6467]. URL: http://www.ajpe.org/lookup/doi/10.5688/ajpe7782.

Pollitt, Alastair (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300.

Qiu, Sihang, Gadiraju, Ujwal, & Bozzon, Alessandro (2020). Towards memorable information retrieval. In *Proceedings of the 2020 ACM SIGIR on international conference on theory of information retrieval* (pp. 69–76). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-8067-6, http://dx.doi.org/10.1145/3409256.3409830.

Roy, Nirmal, Moraes, Felipe, & Hauff, Claudia (2020). Exploring users' learning gains within search sessions. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (pp. 432–436). Vancouver BC, Canada: Association for Computing Machinery, ISBN: 978-1-4503-6892-6, http://dx.doi.org/10.1145/3343413.3378012.

Roy, Nirmal, Torre, Manuel Valle, Gadiraju, Ujwal, Maxwell, David, & Hauff, Claudia (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* (pp. 229–238). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-8055-3, http://dx.doi.org/10.1145/3406522.3446025.

Salmerón, Ladislao, Delgado, Pablo, & Mason, Lucia (2020). Using eye-movement modelling examples to improve critical reading of multiple webpages on a conflicting topic. *Journal of Computer Assisted Learning*, [ISSN: 1365-2729] *36*(6), 1038–1051. http://dx.doi.org/10.1111/jcal.12458, URL: http://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12458.

Senk, Sharon, & Usiskin, Zalman (1983). Geometry proof writing: A new view of sex differences in mathematics ability. *American Journal of Education*, [ISSN: 0195-6744] *91*(2), 187–201, URL: http://www.jstor.org/stable/1085041.

Shi, Jianwei, Otto, Christian, Hoppe, Anett, Holtz, Peter, & Ewerth, Ralph (2019). Investigating correlations of automatically extracted multimodal features and lecture video quality. In *Proceedings of the 1st international workshop on search as learning with multimedia information* (pp. 11–19). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-6919-0, http://dx.doi.org/10.1145/3347451.3356731.

Shute, Valerie J. (2008). Focus on formative feedback. *Review of Educational Research*, [ISSN: 0034-6543] *78*(1), 153–189. http://dx.doi.org/10.3102/0034654307313795, Publisher: American Educational Research Association.

Singley, Mark K. (1990). The reification of goal structures in a calculus tutor: Effects on problem-solving performance. *Interactive Learning Environments*, [ISSN: 1049-4820] *1*(2), 102–123.

Smith, Troy A., & Kimball, Daniel R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, [ISSN: 0278-7393] *36*(1), 80–95. http://dx.doi.org/10.1037/a0017407, URL: https://auth.lib.unc.edu/ezproxy_auth.php?url=https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2009-24668-022&site=ehost-live&scope=site.

Sousa, David A. (2017). *How the brain learns* (5th ed.). Corwin Press, ISBN: 978-1-5063-4632-8.

Syed, Rohail, & Collins-Thompson, Kevyn (2017). Optimizing search results for human learning goals. *Information Retrieval Journal*, *20*(5), 506–523. http://dx.doi.org/10.1007/s10791-017-9303-0, [ISSN: 1386-4564, 1573-7659]. URL: http://link.springer.com/10.1007/s10791-017-9303-0.

Taramigkou, Maria, Apostolou, Dimitris, & Mentzas, Gregoris (2017). Supporting creativity through the interactive exploratory search paradigm. *International Journal of Human–Computer Interaction*, *33*(2), 94–114. http://dx.doi.org/10.1080/10447318.2016.1220104, [ISSN: 1044-7318, 1532-7590]. URL: https://www.tandfonline.com/doi/full/10.1080/10447318.2016.1220104.

Thurstone, Louis L. (1927). A law of comparative judgment.. *Psychological Review*, *34*(4), 273.

Torres-Guijarro, Soledad, & Bengoechea, Mercedes (2017). Gender differential in self-assessment: a fact neglected in higher education peer and self-assessment techniques. *Higher Education Research & Development*, [ISSN: 0729-4360] *36*(5), 1072–1084. http://dx.doi.org/10.1080/07294360.2016.1264372.

Urgo, Kelsey, Arguello, Jaime, & Capra, Robert (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* (pp. 77–84). Virtual Event Norway: ACM, ISBN: 978-1-4503-8067-6, http://dx.doi.org/10.1145/3409256.3409815, URL: https://dl.acm.org/doi/10.1145/3409256.3409815.

Weingart, Nino, & Eickhoff, Carsten (2016). Retrieval techniques for contextual learning. In *SAL @ SIGIR* (p. 5).

Wildemuth, Barbara M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, [ISSN: 1532-2890] *55*(3), 246–258. http://dx.doi.org/10.1002/asi.10367, URL: https://www.onlinelibrary.wiley.com/doi/abs/10.1002/asi.10367.

Willoughby, Teena, Anderson, S. Alexandria, Wood, Eileen, Mueller, Julie, & Ross, Craig (2009). Fast searching for information on the internet to use in a learning context: The impact of domain knowledge. *Computers & Education*, [ISSN: 0360-1315] *52*(3), 640–648. http://dx.doi.org/10.1016/j.compedu.2008.11.009, URL: https://www.sciencedirect.com/science/article/pii/S0360131508001802.

Wilson, Max L., André, Paul, & schraefel, mc (2008). Backward highlighting: enhancing faceted search. In *Proceedings of the 21st Annual ACM symposium on user interface software and technology* (pp. 235–238). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-59593-975-3, http://dx.doi.org/10.1145/1449715.1449754.

Wilson, Mathew J., & Wilson, Max L. (2013). A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology*, [ISSN: 1532-2890] *64*(2), 291–306. http://dx.doi.org/10.1002/asi.22758, URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22758.

Wu, Wan-Ching, Kelly, Diane, Edwards, Ashlee, & Arguello, Jaime (2012). Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *IIIX '12, Proceedings of the 4th information interaction in context symposium* (pp. 254–257). New York, NY, USA: ACM, ISBN: 978-1-4503-1282-0, http://dx.doi.org/10.1145/2362724.2362768, URL: http://doi.acm.org/10.1145/2362724.2362768.

Xu, Luyan, Zhou, Xuan, & Gadiraju, Ujwal (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* (pp. 91–100). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-7098-1, http://dx.doi.org/10.1145/3372923.3404784.

Yu, Ran, Gadiraju, Ujwal, Holtz, Peter, Rokicki, Markus, Kemkes, Philipp, & Dietze, Stefan (2018). Predicting user knowledge gain in informational search sessions. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 75–84). New York, NY, USA: ACM, ISBN: 978-1-4503-5657-2, http://dx.doi.org/10.1145/3209978.3210064, URL: http://doi.acm.org/10.1145/3209978.3210064.

Zhang, Yao, & Liu, Chang (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* (pp. 47–56). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-7585-6, http://dx.doi.org/10.1145/3383583.3398532.