Tenison and Sparks
Large-scale Assessments in Education (2023) 11:28
https://doi.org/10.1186/s40536-023-00164-w

Large-scale Assessments in Education

RESEARCH Open Access

Combining cognitive theory and data driven approaches to examine students' search behaviors in simulated digital environments

Caitlin Tenison[1*] and Jesse R. Sparks[1]

*Correspondence: ctenison@ets.org

[1] Educational Testing Service, Princeton, NJ, USA

Abstract

Background: Digital Information Literacy (DIL) refers to the ability to obtain, understand, evaluate, and use information in digital contexts. To accurately capture various dimensions of DIL, assessment designers have increasingly looked toward complex, interactive simulation-based environments that afford more authentic learner performances. These rich assessment environments can capture process data produced by students' goal driven interactions with digital sources but linking this data to inferences about the target constructs introduces significant measurement challenges which cognitive theory can help us address.

Methods: In this paper, we analyzed data generated from a simulated web search tool embedded within a theoretically-grounded virtual world assessment of multiple-source inquiry skills. We describe a multi-step clustering approach to identify patterns in student's search processes by bringing together theory-informed process data indicators and sequence clustering methods.

Results: We identified four distinct search behaviors captured in students' process data. We found that these search behaviors differed both in their contribution to the web search tool subscores as well as correlations with task level multiple-source inquiry subconstructs such as locating, evaluating, and synthesizing information. We argue that the search behaviors reflect differences in how students generate and update their task goals.

Conclusion: The data-driven approach we describe affords a qualitative understanding of student strategy use in a complex, dynamic simulation- and scenario-based environment. We discuss some of the strengths and challenges of using a theoretical understanding of multiple-source inquiry to inform how we processed, analyzed, and interpreted the data produced from this assessment tool and the implications of this approach for future research and development.

Introduction

The societal importance of digital literacy competencies for academic and professional life (Organization for Economic Co-operation and Development [OECD], 2013) has necessitated the development of digital assessments that measure what individuals and groups know and can do with respect to these crucial skills. Large-scale international

assessments of related skills have been developed and administered across multiple age levels and populations (e.g., Mullis & Martin, 2019; PIAAC Expert Group, 2009; OECD, 2019a, 2019b, 2021). To reflect changing notions of digital information literacy (DIL), researchers and assessment developers have developed complex, interactive simulation-based assessments which capture authentic learner performances in rich environments, introducing measurement challenges (Scheerder et al., 2017). These digitally-based assessments afford opportunities to collect multiple streams of evidence of students' DIL skills, from item responses to moment-to-moment actions (i.e., process data captured in log files). This data can be used to investigate the relationships among digital literacy processes and final products and the complexities of the fine-grained processes and behaviors that make up those performances (see Coiro et al., 2018; Sparks et al., 2016). While these interactive assessment tasks can provide a rich, continuous stream of data capturing students' interactions with the task environment, these data also present a significant challenge for measurement and interpretation. Just as task developers have used frameworks like Evidence-Centered Design (ECD; Mislevy et al., 2003) to engage in principled approaches to the design of assessment tasks, analysis of data resulting from such tasks calls for a similarly principled approach in terms of using theory to inform our methods and models.

In this paper we apply recent recommendations for identifying process level features from large scale assessments (Arslan et al., 2023; Goldhammer et al., 2021; Kroehne & Goldhammer, 2018) to guide modeling of distinct strategies captured within process data generated from students' interactions with a simulated web search tool embedded within a Virtual World for English Language Arts (ELA), a theoretically-grounded virtual world environment and scenario-based assessment task designed to measure multiple aspects of students' DIL skills in the context of conducting online inquiry from multiple sources within an engaging, interactive technology-based environment (Sparks et al., 2018; see also Coiro et al., 2018). In our analyses of this dynamic web search tool, which was specifically designed to collect robust moment-by-moment process data, we consider the content and timing of student actions as well as the context in which the actions occurred from the perspective of models of multiple-document comprehension (i.e., Rouet & Britt, 2011) and new literacies perspectives on online reading (e.g., Coiro & Dobler, 2007; Leu et al., 2008, 2013). Using sequence clustering methods, we identify distinct behaviors captured in students' process data and explore how these behaviors relate to task performance. This approach affords a qualitative understanding of students' search behaviors that reflects the complexity of their interactions within increasingly rich DIL tasks and contributes to a growing research literature documenting the skills and strategies students use during online reading (see Afflerbach & Cho, 2009; Cho, Afflerbach, & Han, 2018; Coiro & Dobler, 2007; Coiro, 2020; Rouet, 2006; Salmerón et al., 2018).

Theoretical background

Assessing digital information literacy with interactive digital tasks

Contemporary notions of digital literacy reflect the complex interplay of knowledge, skills, and abilities (KSAs) needed to support individuals as they obtain, understand, evaluate, and use information within digital environments (Sparks et al., 2016). Thus,

DIL includes KSAs related to defining, accessing, evaluating, managing, integrating, and creating information through critical and effective use of digital tools and technologies to solve information-based problems. Large-scale international assessments that measure such competencies include Progress in International Reading Literacy Study (PIRLS and ePIRLS; Mullis & Martin, 2015, 2019), the International Computer and Information Literacy Study (ICILS; Fraillon et al., 2019); the Programme for International Student Assessment (PISA) literacy assessment (OECD, 2013; 2019a); and the Programme for International Assessment of Adult Competencies (PIAAC) Problem-Solving in Technology Rich Environments (PS-TRE) assessment (PIAAC Expert Group, 2009; OECD, 2019b). These assessments incorporate interactive, simulated digital environments (e.g., simulated web browsers, search engine results, and/or webpages) and ask test takers to interact directly with the interface to achieve certain tasks (e.g., interacting with a list of links in a simulated search results page to identify a website that meets certain criteria; see OECD, 2019a). While the scoring models of these assessments are often focused on the outcomes of students' interactions in such tasks, the collection of process data in digital log files enables finer-grained investigation of students' inquiry processes (e.g., Hinostroza et al., 2018; Van Deursen & Van Dijk, 2009).

Prior research using process data from these large-scale assessments to model DIL skills has primarily focused on identifying patterns in how individuals locate information in digital environments (e.g., Hahnel et al., 2019). These patterns can provide valuable descriptive information about individuals' search behaviors or can be further linked to the outcomes (i.e., products) of their search processes. A wide variety of data-driven methodologies have been explored for identifying patterns in student's problem-solving sequences (e.g., Gao et al., 2022; Hao et al., 2015; He & Von Davier, 2016; He et al., 2019, 2021; Tang et al., 2020; Ulitzsch et al., 2021). The primarily assumption across these methods is that strategic differences in test takers' problem solving is manifested in the variation in the order in which actions occur; that is, students who use similar strategies will produce sequences of similar actions, occurring in similar orders, of similar lengths. Analyses identifying similarities in student problem-solving processes have been applied to PS-TRE process data to understand strategy use on both single item (He & Von Davier, 2016; He et al., 2019; Ulitzsch et al., 2021) and multi-item data sets (He et al., 2021; Wang et al., 2020). These differences are frequently mapped to differences in outcome measures from the PS-TRE assessment or tied to certain background characteristics of the test takers. The link between process data differences and outcome variables (i.e., nomothetic span) is viewed as an important aspect of a validity argument (Embretson, 1983; Keehner et al., 2017; Goldhammer et al., 2021).

Capturing complexities of online inquiry with the virtual world for ELA task

Current large-scale assessments provide a limited view into students' DIL skills, especially in terms of how these skills are leveraged in the context of conducting extended inquiries from multiple sources in digital environments. The Virtual World for English Language Arts (ELA) project (Sparks et al., 2018) aimed to develop a digital platform for measuring information gathering, processing, and integration skills, considering growing emphasis on DIL skills across all aspects of the curriculum and especially their role in supporting tasks requiring research and inquiry (Coiro, 2011; Goldman et al., 2010;

NGA & CCSSO, 2010; Sparks & Deane, 2015; Zhang & Quintana, 2012). Specifically, we explored students' interactions in inquiry-based scenarios that situated them in a simulated social context to motivate engagement in locating, evaluating, reading, and writing synthetic arguments from multiple sources, specifically defining our target construct and designing assessment activities in ways that are consistent with contemporary models of multiple-text comprehension (e.g., Britt & Rouet, 2012; Lawless et al., 2012) and new literacies (e.g., Coiro, 2020; Leu et al., 2008, 2013).

Theories underlying the ELA task construct definition

In defining the target construct of multiple-source inquiry for the virtual world task, we conducted literature reviews synthesizing several topics under the DIL umbrella, including information problem solving (Brand-Gruwel & Stadtler, 2011; Brand-Gruwel et al., 2005; Walraven et al., 2008), multiple-text comprehension (Britt & Rouet, 2012; Goldman et al., 2010, 2011, 2012, 2013, 2018; Graesser et al., 2007; Wiley et al., 2009), and new literacies perspectives on online reading (Coiro, 2011, 2020; Leu et al., 2008, 2013) (see Sparks & Deane, 2015; Sparks et al., 2016; Coiro et al., 2018a,b for reviews). Thus, our construct definition requires students to define problems and information needs; locate potentially relevant information sources; evaluate those sources for relevance and reliability; process, analyze, and synthesize their contents; and communicate what one has learned (Coiro et al., 2018), in ways that are consistent with multiple-document comprehension models like the Multiple-Document Task-based Relevance Assessment and Content Extraction model (MD-TRACE; Rouet & Britt, 2011) and new literacies perspectives on online inquiry and reading comprehension (e.g., Afflerbach & Cho, 2009; Coiro, 2011, 2020).

Models of multiple-document comprehension

Broadly, models of multiple document comprehension have emphasized the additional complexities and skills required to adequately comprehend multiple sources (e.g., Britt & Rouet, 2012), as opposed to the cognitive processes required for comprehension of single texts (e.g., Kintsch, 1998). At minimum, distinct texts may be written by different authors with their own aims, purposes, and perspectives, which necessitates attending to information about the source of the documents in order to account for and resolve potential discrepancies in their contents (e.g., Braasch et al., 2012) which themselves may not be explicit and require knowledge-based and cross-text inferences (Britt & Aglinskas, 2002; Britt & Rouet, 2012; Goldman, 2004; Perfetti et al., 1999; Wineburg, 1991). The goal of multiple document comprehension is a coherent mental model of the relationships among the documents and their sources, or a *documents model* representation (Perfetti et al., 1999).

The MD-TRACE model (Rouet & Britt, 2011) elaborated on this product model to account for the unfolding cognitive processes that occur during multiple document comprehension, particularly in the context of inquiry and information-based problem-solving tasks that are common to many academic domains (Britt & Rouet, 2012). This model is a descriptive model that captures the interplay among cognitive processes, internal cognitive resources, and external task-based resources that individuals may experience as they undertake tasks involving multiple documents (e.g.,

using information from multiple web pages to write an essay discussing whether human activity contributes to global climate change, as part of a science assignment). MD-TRACE describes multiple document comprehension as a task-oriented process involving the following five steps, which are iteratively cycled through until the student satisfies their information needs and task goals (Rouet & Britt, 2011):

1. Construct or revise a mental model of the task, goals, and success criteria
2. Assess information needs, based on prior knowledge and questions that must be answered to satisfy task goals
3. Document use, which involves (a) locating and selecting sources that are relevant to the task, (b) reading and comprehending the sources to build a mental model of their content, (c) constructing or revising a documents model reflecting relationships among the sources
4. Apply information from the mental model to construct or revise the task product
5. Evaluate the task product in terms of task, goals, and success criteria

Relevance of information with respect to task goals is a central consideration of this model, consistent with cognitive research indicating that goals or purposes for reading affect readers' attention with consequences for recall and comprehension (Britt et al., 2018; McCrudden et al., 2010, 2011; Pichert & Anderson, 1977; van den Broek et al., 2001). This iterative cycle is consistent with research emphasizing the metacognitive and self-regulation skills required for successful inquiry (e.g., Azevedo & Cromley, 2004; Edelson, 2002; Zhang & Quintana, 2012).

New literacies perspectives on online reading comprehension

As seen from a new literacies perspective, successful participation in situations where readers are tasked with gathering and comprehending information from multiple sources (as described by MD-TRACE) in online contexts requires advanced reading skills beyond required in single or multiple-text comprehension in offline settings, including: (a) searching and navigating to identify relevant links in hypertext environments; (b) integrating multiple formats and distinct text content; and (c) critically evaluating information, including for trustworthiness (Afflerbach & Cho, 2009; Coiro, 2011; Coiro & Dobler, 2007; Leu et al., 2013; Salmerón et al., 2018). These complex online inquiry practices require distinct skills and strategies that are not explicitly accounted for in more general models of multiple-document comprehension (see Coiro et al., 2018). Using Internet search engines, constructing appropriate search terms, navigating websites, and monitoring one's progress toward information needs all reflect unique aspects of online inquiry. In search engine contexts, students make predictions about document relevance and reliability based on available cues (Metzger et al., 2010; Rieh, 2002), including keywords that indicate relevant semantic overlap (Rouet & Britt, 2011) or source features (e.g., type of website, publisher) that provide information about trustworthiness (Rieh, 2002; Rieh & Hilligoss, 2008). These predictions in turn shape individual's decisions about search

and navigation (Afflerbach & Cho, 2009; Cho, Afflerbach & Han, 2018; Salmerón et al., 2018).

ELA virtual world task and digital tools

The Virtual World for ELA was designed as a platform to measure multiple aspects of digital inquiry (Sparks et al., 2018; see also Coiro et al., 2018). We used a scenario-based assessment design to meaningfully situate and motivate digital tool use within a goal driven context (e.g., Coiro, 2011; Sabatini et al., 2018; Sparks & Deane, 2015; Sparks et al., 2021). The scenario-based task developed for the ELA Virtual World featured an overarching narrative context, a goal-driven scenario, and assessment activities that challenged students to locate task-relevant information by interacting with virtual characters and reading "print" and digital text resources housed in different locations within a virtual town (Sparks et al., 2018).

Specifically, students were asked to conduct research to evaluate the accuracy of 10 key claims contained in an artifact describing one character's (purported) participation in a historical event; students then use this analysis to determine whether the artifact should be placed into the virtual town's history museum, writing a source-based argument to defend their conclusions as a culminating performance task (Sparks et al., 2018). Of the 10 key claims, eight focus on the event itself, while two focus on a secondary topic related to the event (i.e., a narrower subtopic). This task had three phases: *Setup*, *Free Roam*, and *Conclusion*. These phases establish the scenario, context, and task goals; enable free exploration and navigation among available sources, including opportunities to read and evaluate those sources; and prompt integration and synthesis of collected sources toward an overall response to the inquiry task, respectively. Each phase of the task is associated with specific digital tools that are designed to elicit and capture students' responses and processes with respect to the student model, in a coherent and principled fashion. These digital tools allow for students to deepen their understandings as they interact within and across tools and texts, consistent with the iterative cycles and revisions to mental models and task products assumed in MD-TRACE (Rouet & Britt, 2011). Figure 1 illustrates two tools emphasized in the current analyses—the simulated web search and the evidence manager.

Simulated web search tool

Students could access the web search tool ("Toogle" search) by accessing the Internet Café location. This tool was designed to simulate searches on the open web, using a Google-like search engine interface with a single textbox in which keyword strings can be entered. This tool used a split-screen layout, with search engine results appearing on the left panel, and full-text sources (i.e., websites or image results) appearing on the right panel once clicked (see Fig. 1A). The tool contained approximately 25 sources that could potentially be retrieved; sources varied in topic relevance and reliability (i.e., document type, source expertise, bias). Information that addressed all ten key claims within the task was distributed across three key websites; two key websites addressed the historical event directly, and a third key website addressed two claims about a secondary topic.

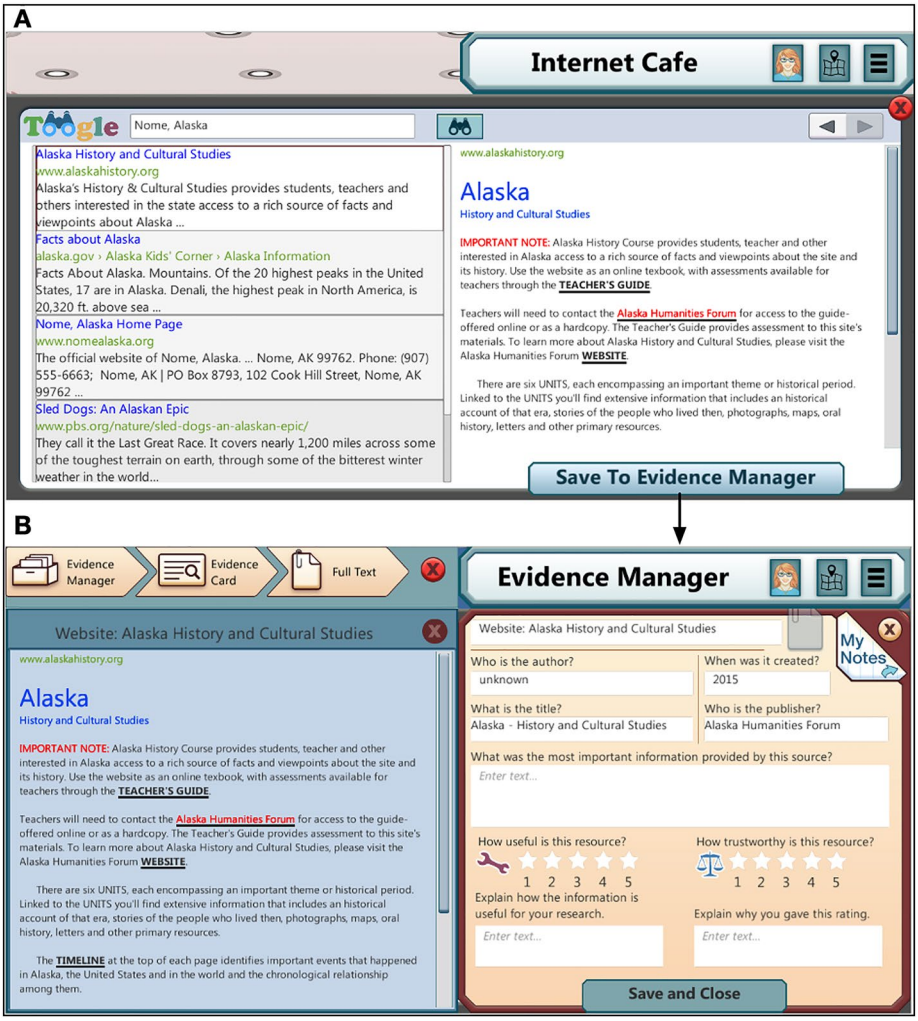

Fig. 1 Simulated Web Search and Evidence Manager Tools. **A** Screenshot of the simulated web search tool illustrating a search engine results page on the left panel and a selected source (i.e., website) on the right panel. The search results and website displayed reflect information with low relevance to the inquiry task. **B** Screenshot of the Evidence Manager tool, which is triggered by pressing the "Save to Evidence Manager" button available on any active source. The screenshot illustrates that the full text of the source is presented to students on the left panel, while prompts to evaluate the relevance, usefulness, and trustworthiness of the information appear on the right panel. After completing the fields on the right, the "Save and Close" button becomes active, enabling students to save their evaluations to the Evidence Manager

When students enter a set of search terms in natural language and click the search button, the interface displays one of several predefined sets of results, based on regular expressions (Jurafsky & Martin, 2020). Each set of results includes a list of five websites with varying degrees of relevance, reliability, and usefulness to the inquiry task. Two sets of results (retrieved via different keywords) each contained the same two key websites about the historical event, a third set of results contained a third key website (about the secondary topic), and a fourth and fifth set of results contained only tangentially relevant information. If students enter entirely irrelevant (off-topic) keywords, they are presented with a non-interactive (i.e., null) set of search results and a prompt from the virtual partner to align their search terms with the topic (i.e., a weak hint). Students can

complete an unlimited number of searches but following five irrelevant searches, they will be explicitly prompted by the virtual partner to type in a specific phrase that will yield relevant results (i.e., strong hint). It is then up to the student to act on the hint and to identify which of the websites in the set of results are most relevant and useful to the task. In cases when 29 min of free roam elapsed without the student finding all resources needed to complete the task, the system provided students with the resources needed for the final phase of the task (i.e., key resource resetting). These hints were incorporated following cognitive labs revealing that students sometimes had difficulty crafting search terms and subsequently locating the key websites (Sparks et al., 2018).

Evidence manager tool

The Evidence Manager tool provides students with an opportunity to collect, makes sense of, and evaluate resources they encounter in the course of their inquiry. When students locate a source they think is useful, they are instructed (by the virtual partner) to click the "Save to Evidence Manager" button which appeared at the bottom of every active source (see Fig. 1A). Clicking this button triggers the Evidence Manager to open. An *evidence card* containing prompts to evaluate the source to appear on the right side of the screen, while the full text of the source (i.e., website, library book summary, or conversation transcript) appeared on the left side, with scroll bars displaying as necessary (see Fig. 1B). This split-screen interface offered the affordance of being able to review the information contents and be asked to respond to comprehension and evaluation questions about the information simultaneously. For example, students are prompted to evaluate the relevance, usefulness, and trustworthiness of the source, while providing explanations for their ratings, thus providing evidence of their KSAs related to evaluating source relevance and reliability. Unlike the web search tool, which is tied to a specific "location" in the virtual world, the Evidence Manager tool can be accessed at any point throughout the task via a menu in the upper right of the screen. Thus, students can review the resources and evaluations saved to the Evidence Manager at any time; they can also revise their evaluations of each source.

Within the ELA Virtual World, we deliberately allowed for repeated cycles of locating, gathering, and evaluation, in addition to enabling just-in-time reference to resources supporting development of the task model, revisions to document evaluations in the evidence manager as new information was uncovered, ability to return to information gathering after beginning the synthesis tasks in the Conclusion phase, among other features, capturing students' dynamic interactions with these various features using process data.

ELA evidence model and scoring rules

Students' actions and responses to constructed-response questions embedded in the task were evaluated and scored to assess multiple-source inquiry subconstructs of planning, locating, evaluating, and synthesizing. The ELA evidence model defines scoring rules used to evaluate students' tool, the scoring rules, and associated points. These scoring rules illustrate how students' performances within the complex digital task environment—including both traditional item responses and actions taken within digital tools—are evaluated and transformed into quantitative measures for analysis (Table 1). In addition to selected actions being evaluated and scored, the task environment

Table 1 Example Evidence Model for Simulated Web Search Tool and Evidence Manager

Tool	Subconstruct	Description	Relevant tool actions	Scoring rules	Points
Web Search	Locate Sources (use effective search keywords)	Gather potentially relevant information by using appropriate search terms	Enter keywords into web search to retrieve relevant sets of results	Students receive up to 2 points for successfully retrieving high-relevance results (of two possible sets). Students can receive 1 point for successfully retrieving medium-relevance results. Points reflect the best set of results seen over all searches completed	0–2
Web Search	Locate Sources (access the key Websites)	Evaluate the relevance of sources, based on a cursory view of the content	Click to view key Websites from search results based on their descriptions	Students receive 1 point for successfully retrieving (i.e., "click to view" action) each of the three key Websites (i.e., most useful for the inquiry task). Points reflect the total number of key websites accessed	0–3
Web Search	Locate Sources (gather useful information for later application)	Save resources which may provide useful information	Click to save key Websites to Evidence Manager tool	Students receive 1 point for successfully gathering (i.e., "Save to Evidence Manager" action) each of the three key Websites (i.e., most useful for the inquiry task). Points reflect the total number of key websites saved	0–3
Evidence Manager	Process, analyze, and synthesize sources (Identify key task-relevant information)	Identify key information (task-relevant information) within a source	Write a description of the key task-relevant information from the source	Human scored. Students receive 1 point for accurately describing the critical content of the source for each of the three key websites (partial credit possible)	0–3
Evidence Manager	Evaluate information sources (Evaluate usefulness)	Identify source usefulness	Click to rate key Websites as Useful in Evidence Manager tool	Students receive 1 point for rating each of the three key websites as "Useful" (i.e., 4 or 5 stars)	0–3
Evidence Manager	Evaluate information sources (Evaluate usefulness)	Explain why collected source is useful	Write an explanation for why the information is useful for the research task	Human scored. Students earn 1 point for accurately evaluating how the source/information is useful for the research task for each of the three key websites (partial credit possible)	0–3
Evidence Manager	Evaluate information sources (Evaluate trustworthiness)	Identify source trustworthiness	Click to rate key Websites as Trustworthy in Evidence Manager tool	Students receive 1 point for rating each of the three key websites as "Trustworthy" (i.e., 4 or 5 stars)	0–3
Evidence Manager	Evaluate information sources (Evaluate trustworthiness)	Explain why collected source is trustworthy	Write an explanation for the trustworthiness rating	Human scored. Students earn 1 point for accurately evaluating the trustworthiness of the source for each of the three key websites (partial credit possible)	0–3

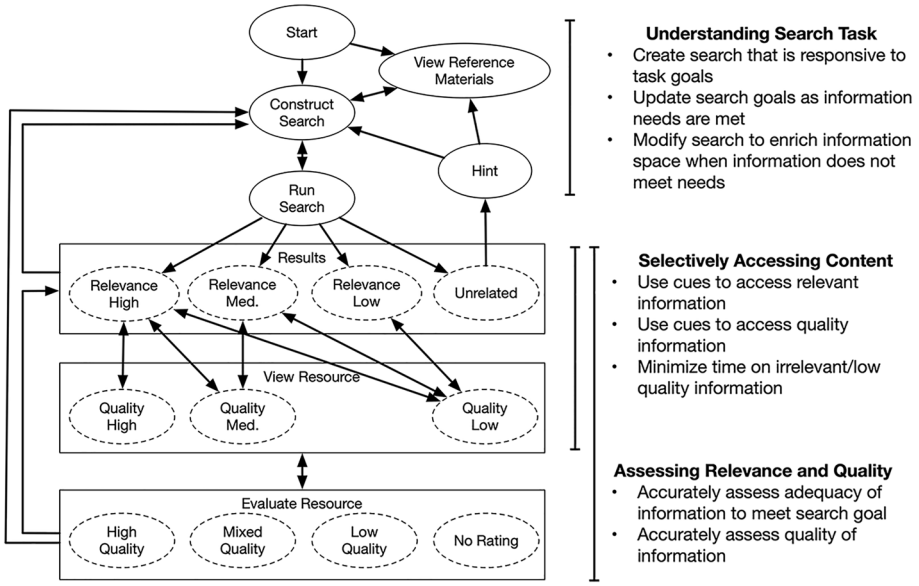

Fig. 2 Cognitive Task Model of the Simulated Web Search and Evidence Manager Tools. Cognitive task model of the simulated web search and evidence manager tools (left) and corresponding behaviors we expect based on the ELA Virtual World construct definition (right). Vertical bars between the cognitive task model and each set of behaviors indicate where in the task we expect to observe these behaviors

recorded rich process data designed to capture students' dynamic interactions with tasks and tools. These interactions can be examined using alternative analytic approaches to obtain additional evidence of students' KSAs—including potential to investigate strategy use.

The current study

The aim of the current study is to leverage the dynamic process data collected within the ELA Virtual World (Sparks et al., 2018) to provide deeper insights into the KSAs and strategies students engage in when conducting online inquiry tasks within a simulated web search tool. We focus on students' use of the simulated web search tool and associated use of the evidence manager tool to evaluate website contents (Fig. 2). These tools could be visited multiple times throughout the task, allowing for observation of iterative cycles as described in MD-TRACE (Rouet & Britt, 2011), as well as the potential to observe differences in the strategies students brought to bear when using these tools. Our aims in modeling this dataset were to (1) characterize differences in students' search behaviors and (2) identify whether there are systematic differences in between the task outcome measures and the behaviors students exhibit within the task.

Aims and research questions

In the current study we modeled students' search behaviors as they interacted with a simulated web search tool to locate and evaluate websites relevant to the overall goals of a scenario-based multiple-source inquiry task as part of a tryout study. We used an unsupervised clustering approach to identify differences in search behavior across

sessions and students. While there have been numerous recent studies using sequence mining approaches to analyze PS-TRE process data (e.g., Gao et al., 2022; He et al., 2019; Ulitzsch et al., 2021), one of our objectives was to explore how our modeling approach needed to be adapted to better handle the additional complexity of this ELA Virtual World task compared to existing assessments. We applied recent frameworks for analyzing process data (Arslan et al., 2023; Goldhammer et al., 2021; Kroehne & Goldhammer, 2018) to our dataset to pursue three research questions (RQs):

- RQ1. What are the primary strategies students use when constructing and modifying searches in this task?
- RQ2. How does students' use of different search behaviors relate to their performance on the ELA virtual world assessment?
- RQ3. How do students' search behaviors change throughout the task?

With respect to RQ1, in modeling students' interactions with the web search tool, we expected to observe differences in how students constructed searches, interacted with search engine results and sources, and made decisions on what sources are most useful for completing the task and should be saved for later use. In Fig. 2 we illustrate a cognitive task model for the simulated web search and evidence manager tools, the focus of our current analysis. This node and arrow representation captures the flow between different activities from search construction to location of relevant information, to source evaluation within the evidence manager. The.decisions that students make at these different stages of the search process have downstream effects on what choices are available to them as they interact within these tools—for example, the information that students can extract and evaluate in the evidence manager is linked to the relevance of sources they access and save, which are in turn constrained by the relevance of search engine results they are able to access based on the quality of the search terms they enter (Fig. 2). For this reason, we modeled search behaviors in terms of students' entire sequence of search processes within a single visit to the Internet Café location—which may include multiple iterations of constructing and executing searches, scrolling results, and accessing and saving sources. We expected to observe differences in students' understanding of their search goals and ability to generate queries that return relevant results. We may see adaptive strategies such as viewing reference materials, which would reflect attention to task goals or use of task-relevant information to guide search behaviors. We expect to also see variation in how students interact with search results and sources based on both their relevance to the task and their reliability (i.e., trustworthiness). For students whose searches produce results with low relevance to the task (e.g., students who receive weak or strong hints from the system) we investigated whether students adapted their search behaviors and what subsequent actions they took.

The evidence model and scoring rules (Table 1) delineate key indicators we would expect to be produced by students who are proficient with the different decision-making processes and behaviors described in Fig. 2. These scoring rules however do not consider the process by which students arrive to these different choices; RQ2 explores this relationship between processes and scores. According to MD-TRACE, an individual's behavior when searching for information balances tradeoffs between the satisfying the

individual's goals around finding relevant, reliable, and useful information and minimizing the cost (e.g., time, effort, energy, monetary) of accessing and processing information (Rouet et al., 2021). As predicted by MD-TRACE, individuals who can balance these needs can identify relevant information from a list of sources, and quickly abandon searches that render irrelevant or unreliable sources. In contrast, students who struggle to determine information relevance may fall back on the use of simple heuristics such as choosing the first option in the list (Gerjets & Kammerer, 2010; Metzger et al., 2010; Rouet et al., 2021) or employing exhaustive strategies (Gao et al., 2022). We expect that these differences in students' awareness of the relevance and reliability of information sources and ability to recognize when they need to change their behavior will result in differences in both student's process and outcome data.

RQ3 considers the role of task contexts in how students' search behaviors may change over the course of the task as they gain experience with the virtual world, gather, and accumulate additional information, and change their goals as appropriate for their progress through the extended inquiry task. This research question is primarily exploratory. Because students can use the web search tool at multiple points in the task and can submit an unlimited number of searches, it is likely that their goals will change from session to session as they accumulate information gathered from other virtual world locations and tools within their evidence manager. Within MD-TRACE, the student's task model plays an important role in guiding their understanding of their information needs, formation of goals, and construction of search terms. Variability in students' underlying reading comprehension and self-regulation skills have been shown to influence their ability to understand, articulate, and remember their search task goals (for a review, see Rouet et al., 2021). We expect that students who use strategies that are well aligned to the goals of the scenario-based task will meet their information needs more quickly and show greater sensitivity to relevant information.

Methods

Participants

Eighth-grade students (N = 130; 67 females, 60 males) from two schools ($n_{Urban} = 91$, $n_{Rural} = 39$) participated in the scenario-based task during a tryout study, completing the task and a post-survey in one 90-min session, with each student working individually. Of the 130 students, 127 (98%) visited the simulated web search tool at least once during the Free Roam phase. However, only 109 students proceeded from the Free Roam phase to the Conclusion phase, with 104 students completing the claim evaluation task outcome measure (i.e., selected-responses and short constructed-responses), and 98 students completing the subsequent source-based argument task outcome measure (i.e., extended constructed-response). Thus, approximately 24% of participating students did not complete the Conclusion phase measures, and 31% completed only the first of these measures, mainly due to an inability to complete the task within the 90-min time constraint.

Materials and scoring

Students completed the scenario-based task previously described. Responses and actions were evaluated based on predefined scoring rules, with a total of 100 points possible for the total task. Points were assigned for individual actions and responses and

were evaluated by a combination of automated scoring and human ratings. Scores could be evaluated both at the task level (e.g., summing scores by phase or by location/tool) and by underlying subconstruct (i.e., planning, locating, evaluating, or synthesis). Students' behaviors in the simulated web search and evidence manager tools were evaluated by applying the scoring rules presented in Table 1. Altogether 23 points were possible within the simulated web search tool, with up to 8 points based on students' actions in locating, retrieving, and saving key websites, and up to 15 points related to students' critical evaluations of those sources in response to the prompts embedded in the Evidence Manager tool (i.e., up to 5 points for each of three key sites).

Process data representation

We took special consideration in identifying the event types and granularity of event sampling that would serve as input for our models. Our approach for selecting this data representation considers the task design, the cognitive basis of the search strategies we aim to model, and the assumptions implicit within the unsupervised modeling approach we are using. These decisions reflect general recommendations for selecting process data features from assessment data (Arslan et al., 2023; Goldhammer et al., 2021; Kroehne & Goldhammer, 2018). The output of this effort is a temporally ordered series of events for each internet search session.

Event types

Actions

We first considered *interface actions* and *student-initiated actions* that result in meaningful changes to the task state. These events provide new information or action-options that may cause the student to alter their goals. These events include running searches, navigating to new pages, and help events. Unlike the student actions around planning, locating, and evaluating information, help events can be solicited (e.g., the student requests a hint by clicking the icon of their virtual partner on the menu bar; see Fig. 2) or unsolicited (e.g., the system provides help after some duration of unproductive behavior, such as the web search hints described previously). Finally, we included a start and end state to contextualize student's behaviors between the opening of the search environment (i.e., the start of the search session) to any action that resulted in the student leaving that environment (i.e., the session end). Altogether, this analysis resulted in a list of 24 actions which will be used for analysis.

Information relevance

MD-TRACE describes search behavior as a tradeoff between the cost of gathering information and the value of information for satisfying the task goals (i.e., usefulness). To capture the usefulness of information for modeling purposes, we coded the search results for their relevance (e.g., the degree to which they provided information relevant to and useful for student's overall inquiry task) and reliability (e.g., the degree to which the source could be trusted to provide trustworthy information), based on their intended designation in the underlying task design. This coding allows our model to test our hypotheses that strategic variation in behavior would be related to information quality. For each source we considered both its reliability and its relevance in creating

a 'quality' score. Sources with both high relevance and reliability were coded as 'high quality' (e.g., an article written by the Center for Disease Control that described key historical details around the event). Items with low reliability or relevance were coded as 'low quality' (e.g., a conspiracy blog questioning whether the event even happened while relevant to the topic comes from an unreliable source). The 'medium quality' items were partially relevant and of mixed reliability (e.g., an on-topic Wikipedia entry). This coding aligned with the classification of the sources in the underlying task design while simplifying the dimensions of reliability and relevance for the present analysis. Finally, we coded students' evaluations of sources in terms of their judgments of information usefulness and trustworthiness in the evidence manager. This provides our model with a means to evaluate whether students can differentiate among sources that are (or are not) useful for completing the overall inquiry task.

Pausing behavior

An action sequence representation captures the sequence in which actions occur; however, it does not capture specific temporal information about when these events occur. According to cognitive theory, the cost of gathering information in web search environments is typically quantified in terms of time (Pirolli & Card, 1999; Pirolli & Fu, 2003; Rouet et al, 2021). These pauses can reflect a variety of activities, ranging from students processing the information within a given page, to making decisions about next steps in their search, or executing strategic actions (Arslan et al., 2023). Prior research suggests that the length of a pause can contribute meaningful information to models of the problem-solving process and is an important indicator of student's problem-solving proficiency (Paquette et al., 2014; Tenison & Anderson, 2016; Tenison & Arslan, 2020; Xu et al., 2020).

We assigned *pauses between actions* to ordinal categories to provide our model with the ability to distinguish actions that are completed in quick succession from those which require more interpretation and planning prior to execution (Fig. 3). To code variation in pauses within student's action sequences we considered the cognitive activities preceding planning, locating and evaluating activities and identified cut points that provided distinction within these different categories. We did not include pauses under 250 ms, the average time it takes for humans to prepare and execute a motor action (Anderson, 2009). These pauses were unlikely to reflect meaningful cognitive activity and our preliminary analysis of these pauses suggested they primarily captured navigational behaviors such as double-clicks. These fast navigation pauses primarily occurred preceding actions made when students were locating information (Fig. 3b). Through further analysis of pauses, we found that other navigational behaviors such as scrolling through search results were of longer duration but did not appear to offer additional evidence about the student's current goal. Based on this analysis of pauses preceding planning, locating, and evaluating actions (see Fig. 3) we removed pauses shorter than 2.5 s from our analysis. We coded pauses between 2.5 s and 10 s as *short*, pauses between 10 and 30 s as *medium*, and pauses above 30 s as *long*. While these categories provide our model with a general distinction between pauses of different duration, the sequence mining approach (described later in this

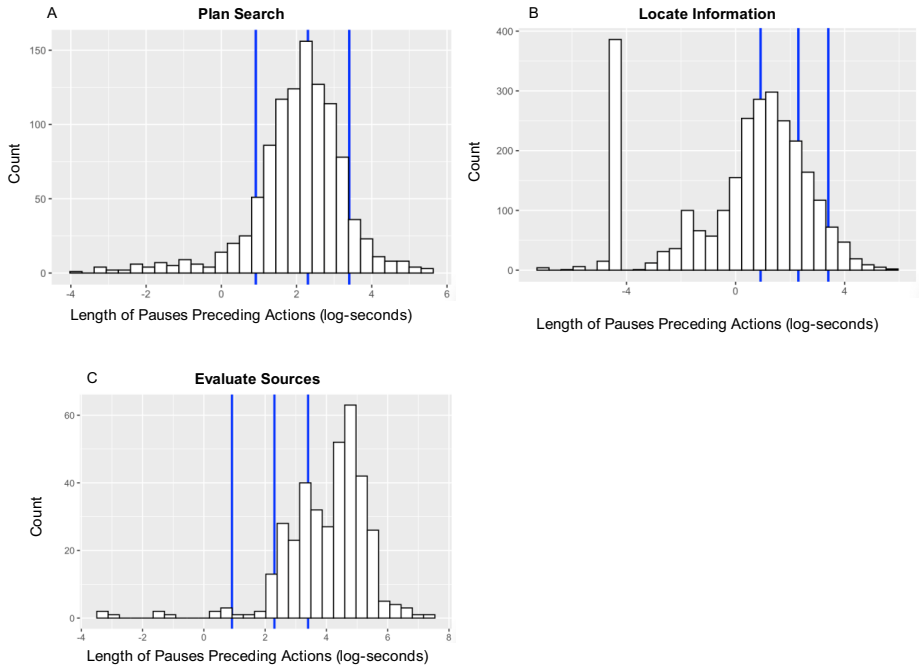

Fig. 3 Histograms of the Duration of Pauses Preceding Actions by Multiple-Source Inquiry Subconstruct. **a** Distribution of pauses preceding plan search actions (M 14.4 s, SD 22.7 s); **b** Distribution of pauses preceding locate information actions (M 7.4 s, SD 17.7 s); **c** Distribution of pauses preceding evaluation of source actions (M 103.6 s, SD 134.0 s). Duration is reported in log-seconds. Blue vertical bars indicate inclusion cutoff criteria. Pauses less than 2.5 s (the leftmost vertical line) are not coded in our action sequences. These pauses often capture time to navigate the interface

section) uses both the pause label and the context of the pause within student's action sequences to estimate the cognitive state that generated the pause.

Event granularity

Once we identified the events of significance, we then reviewed the degree of granularity with which we sampled these events. The raw log file data produced by the scenario-based task recorded detailed, time-stamped information about specific student actions and system events. Our aim in choosing a representation for this process data is to capture the high-level goal states that individuals experience when locating information (Rouet & Britt, 2011). Taking a rational analysis approach, we reviewed how the log files captured how students planned their search, located information, and evaluated sources, and identified a set of key indicator events that could be taken as evidence of execution of these goals (Arslan et al., 2023). In certain cases, we flagged low-level events such as navigational behaviors like scrolling websites and filling in item responses, which captured multistep execution of a single goal. These low-level events reflect micro process-related paradata and low-level response input data (Kroehne & Goldhammer, 2018). These low-level events reflected activities generated by a single "behavior", such as different positions where the scroll bar was placed within a document as students scrolled through it, or keystrokes made as students answer constructed-response items. For these events we recognized the first

occurrence of one of these events as the start of that activity. This decision reflects the inferential approach for identifying theoretically meaningful states given the availability of process data produced in an assessment discussed described by Kroehne and Goldhammer (2018) and Goldhammer et al. (2021) and played an important role in determining the duration of pauses between actions.

Data preparation

We removed some sessions from our dataset to improve the descriptiveness of our models. We removed 46 sessions in which students visited the Internet Café but did not attempt to conduct a search. In these sessions, it appeared that students directly returned to the main town map screen without performing any actions within the simulated web search tool; this could indicate exploratory behavior (e.g., clicking to see what is in the location) or a change in goals (i.e., deciding to visit the location but then changing one's mind and returning to the map to select a different location). Next, we removed three sessions that contained over 125 actions in the search tool. These sessions appeared to include bugs within the logging process that made it difficult to discern what actions were made by the student and what were system-generated. In removing these sequences, we removed an additional student from our final dataset. In the remainder of our analyses in this paper we consider a set of 319 unique sessions generated by 126 students. Because students could return to the web search at multiple points in the task, some have multiple search sessions.

Sequence clustering and modeling

Our aim in clustering students' search processes is to identify distinct behavioral patterns that help us characterize students' ability to plan their search, locate websites, and evaluate those websites. We used mixture Hidden Markov Models (mHMMs) as an approach for modeling clustered timeseries data to model the strategies students used in the ELA task. Following a procedure outlined by Helske and colleagues (2016), we took a multi-step approach that combined edit distance clustering approach with the use of hidden Markov Models (HMMs) to initialize the mHMM model we used to recluster the data. We illustrate our complete approach in Fig. 4, from identifying our process data representation (Fig. 4a; see section Process Data Representation), selecting an initial set of clusters (Fig. 4b), identifying HMM priors for each cluster (Fig. 4c), and fitting final mHMM model (Fig. 4d).

Walkthrough of approach

We first translated our raw log files into sequences of temporally ordered events (Fig. 4a; described in Process Data Representation section). For example, a short session for a student who viewed some reference materials but left without attempting a search would be represented as a vector of events (e.g., 'Start Session, Short Pause, View Reference Materials, Medium Pause, Finish Session).

Our next step in identifying the number of clusters present within our sequence data (Fig. 4b), we used a normalized *optimal matching* (OM) metric to calculate the edit distance between all sequences (TraMineR package in R; Gabadinho et al., 2011). This

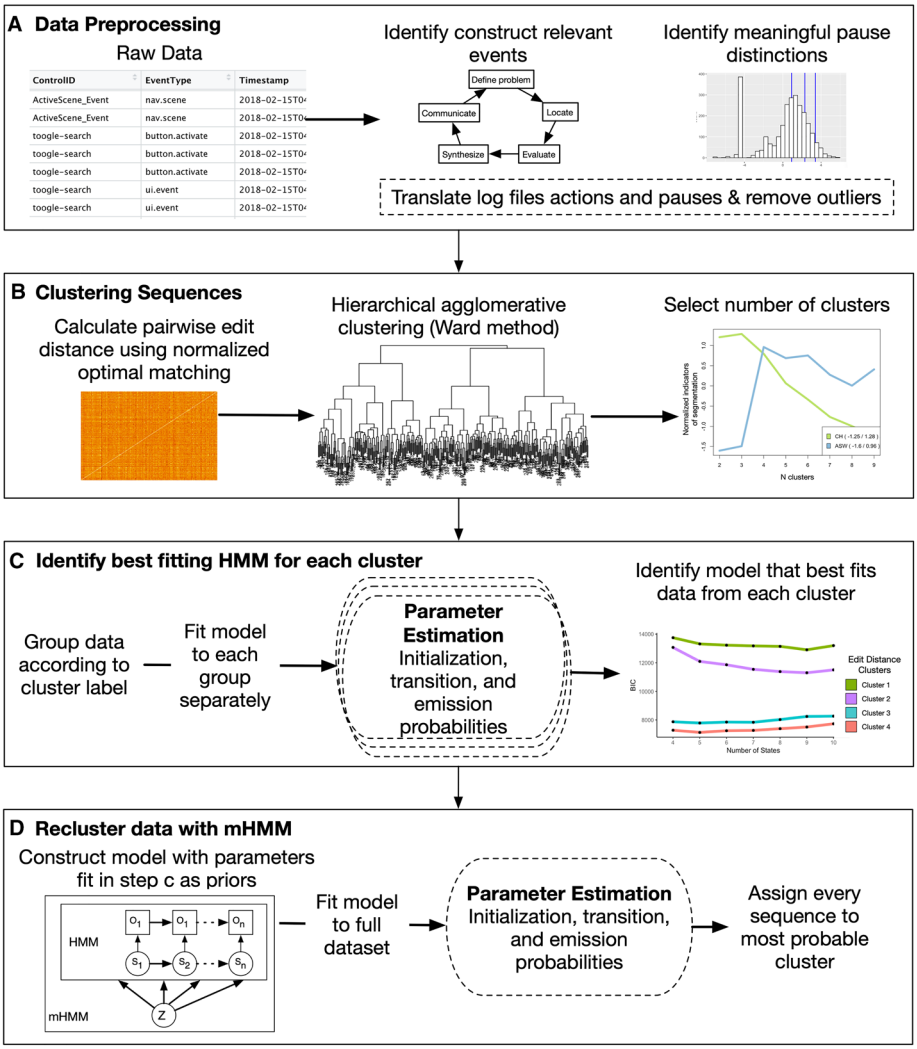

Fig. 4 Multi-Step Strategy Clustering Approach. The output of each step is as follows (**a**) each session represented as a temporally ordered sequence of events; **b** cluster labels for each sequence; **c** separate HMMs for each cluster provide priors for the transition and emission probabilities of the mHMM; (d) a final mHMM fit to the full dataset estimating

algorithmic approach determines the dissimilarity between two sequences by calculating the number of.substitutions, insertions and deletions (referred to as indels) necessary to transform one sequence into another (for introduction to OM see Martin & Wiggins, 2011, pg. 387–409). These transformations can be assigned different penalties and costs depending on certain properties of the sequences and goals of the analysis (Lesnard, 2009; Luu et al., 2020; Struder & Ritschard, 2016). The total cost to transform one sequence into another is referred to as the 'distance' between sequences. With

normalized OM we apply Abbott's normalization (Abbott & Forrest, 1986) to account for differences in sequence length by dividing the distance by the length of the longest of the two sequences. TraMineR then uses the dynamic programming solution proposed by Needleman and Wunsch (1970) to identify the 'cheapest' possible transformation between each pair of sequences given the cost of substitutions and indels and the normalization procedure (Gabadinho et al., 2011).

Prior research modeling PS-TRE process data has explored a variety of metrics, most popular of which is Longest Common Subsequence (LCS; Boroujeni & Dillenbourg, 2019; Hao et al., 2015; He et al., 2021). Like LCS, normalized OM uses a constant indel cost of 1. Unlike LCS, normalized OM derives substitution costs from the transition between events.[1] Accounting for variability in transition rates is suggested to be more appropriate for modeling this type of time-series data found in the social sciences (Lesnard, 2009).[2] The scenario-based task enables students to engage in repeated search cycles, such as iterating between scrolling the results and exploring or reading the sources; the number of search cycles the student engages in is less important to our construct definition than the relevance and reliability of the sources they interact with during the search process. Normalized OM allows us to measure differences in the content of students' sequences while minimizing the effect of variation in how co-occurring actions were ordered and influence of sequence length on cluster formation. We next applied hierarchical agglomerative clustering using the Ward method on the resulting pairwise OM distances to identify clusters of similar search sequences. We considered two different internal evaluation indexes, Calinski-Harabaz index and silhouette width, to determine the number of clusters that would maximize similarity between items within the cluster and differences between items across clusters (Batool & Henning, 2021). Once the number of clusters was chosen, we grouped sequences in terms of their cluster label.

While the normalized OM metric is sensitive to the temporal cooccurrence of events, it does not capture the fact that some actions may take on different meanings depending on the context in which they occur. This has implications for how pauses contribute to cluster formation. To account for this we use mHMMs to re-cluster the data. Using the grouping of sequences based on the edit-distance clusters, we fit separate HMMs, one for each cluster (Fig. 4c). We considered models with 4 to 10 states to describe the four clusters and used Bayesian Information Criterion (BIC) to determine best model fit. We used the seqHMM package in R (Helske & Helske, 2017) to estimate the model parameters using the expectation–maximization (EM) algorithm (Baum & Petrie, 1966; Rabiner, 1989) for both HMMs and mHMM. In fitting our HMMs, we used random priors to initialize all our emission and transition probabilities, except for a start state that was given a 1 probability of the model starting in that state and an end state that was given

[1] . As described by Gabadinho et al. (2009) substitution costs are determined from the estimated transition rates as $2 - p(s_i|s_j) - p(s_j|s_i)$ where $p(s_i|s_j)$ is the probability of observing event s_i at time $t+1$ given that event s_j has been observed at time t. As a result, when certain events frequently transition into each other substitution costs are lower than events with lower transition rates.

[2] Because of its popularity we did explore application of LCS to this dataset. This approach is equivalent to OM with a constant substitution cost of 2 rather than variable substitution rate (Elzinga, 2006). We found that LCS was overly sensitive to differences in the length of the sequence. This behavior is noted in prior work comparing dissimilarity measures (Studer & Ritschard, 2016).

a 0 starting probability and 0 probability of transitioning to any other state. Parameters were estimated using the EM algorithm. We fit five models with different randomized starting values to avoid fitting local optima.

The final transition and emission probabilities from the best fitting model for each cluster was then used to initialize the priors for the mHMM (Fig. 4d). The mHMM model contains two types of variables (Vermunt et al., 2008): time-varying discrete latent variables and a time-constant discrete latent variable. As a mixture of HMMs, the mHMM consist of varying sub-models that characterize the clusters that can have different parameters (e.g. initial, transition, and emission probabilities; see Vermunt et al., 2008; Helske & Helske, 2016; Helske et al., 2016). The time-varying discrete latent variables capture a Markovian transition structure (as a traditional HMM). We choose HMMs over a simpler Markov model to represent students' distinct strategies because we believe that students' actions provide an indirect observation of the cognitive state in which they are engaged. This is especially evident in our pause states, which contain pauses of different durations occurring within the same context. Prior research has used HMMs to model web search strategies from clickstream data captured in realistic settings (e.g., Liu et al., 2006; Luu et al., 2020; Roman et al., 2010). The time-constant discrete latent variables can be viewed as capturing clusters of distinct strategies in web environments. Using the mHMM structure, we estimate the probability that a sequence is generated by a given submodel (Fig. 4d).

While it is possible to use mHMMs to identify both the number of clusters and number of states within each cluster, this approach requires the use of information criteria to select the number of models and states that best capture the data (Dias et al., 2015). In our case, the small number of sequences we have in each cluster and limited number of observations within each sequence limits the reliability of information criteria to determine the numbers of clusters and of latent states within each cluster that best fit the data (Costa & Angelis, 2010). In fact, when we explored using the mHMMs initialized with random priors to cluster our data as well as estimate the number of states, we best fit a single model that captured all possible states and closely resembled our task model (Fig. 2). Prior research using such models suggests that these benefit from the selection of informed starting values (e.g., priors for the model transition and emission probabilities; Helske, 2021; Helske et al., 2016). We chose to initialize our mHMM such that sequences had an equal probability of being clustered in each of the clusters and computed both the posterior cluster probabilities and most probable path of hidden states for each sequence.

Multi-Step clustering of ELA data

For our multi-step clustering approach, we used the Ward method to cluster sequences based on their edit distance (Fig. 4b shows dendrogram of the projected clusters). In considering between 2 through 8 clusters, we found 3 clusters generated the highest Calinski-Harabasz index value (11.7) closely followed by 4 clusters (10.9). Selecting 4 clusters produced the highest average Silhouette width (0.83), this index was much lower for the 3-cluster solution (0.69). For the purposes of our analysis, we present a 4-cluster solution.

For each of the 4 clusters, we then fit separate HMMs with between 4 and 10 states to the sequences in that cluster (Fig. 4c). Using BIC to compare models for each cluster, we found a 9-state model produced the smallest BIC for Cluster 1 and 2, whereas a 5-state model produced the smallest BIC for Cluster 3 and 4. Using the transition and emission probabilities from these models as priors for our mHMM we refit the data. The state transition and emission probabilities for our final mHMM model did not change greatly, however we did find many of the sequences were reassigned. These four clusters of search behaviors differed in terms of the quality of search results and websites viewed, the occurrence of significant pauses within students' search processes, and the overall structure of their search processes.

Results

Simulated web search tool usage within the task

During the Free Roam phase of the task students could visit the simulated web search tool an unlimited number of times and could submit an unlimited number of search terms (i.e., distinct presses of the "search" button). Of the 130 students, 127 (98%) visited this location at least once, submitting on average 8 search terms over the course of completing the task (range: 0 to 59). The modal number of searches submitted was 2, but only 27% of participants submitted 2 or fewer search terms; a majority (52%) submitted 5 or fewer search terms and there was a long tail of outliers submitting more than 30 searches.

Average performance within the web search tool was 7 of 23 possible points, approximately 32%. Performance was low because students struggled with selecting relevant sources, on average viewing only one of the three key websites; only three students (2%) viewed all three key websites; 52 students saw one key website (41%), 34 saw two key websites (27%), and 38 students (30%) were unsuccessful at locating any of the three useful websites. Most students saved only one of these websites to the evidence manager (n = 75, 58%), while 49 students (38%) saved two, and again only three students (2%) saved all key websites to the evidence manager. Students who did save key websites (or who were provided key websites through system actions) showed difficulty evaluating the sources' relevance, usefulness, and trustworthiness, especially based on responses to constructed-response items, suggesting that.some students identified these sources as useful and trustworthy but had difficulty explaining or justifying their ratings.

Students varied in how they interacted with the simulated web search tool. Students had an average of 2.9 search sessions (SD = 2.3, Min = 1 session, Max = 11 sessions), and spent on average 4.3 min conducting each session (SD = 4.6 min, Min = 5 s, Max = 27.6 min). In Table 2 we report the average frequency with which each of the 24 possible actions appear in students' sequences. Considering these sources of information, we observed several patterns. First, most sequences only contain a small subset of the possible actions. Second, sequences can contain long runs of repetitive actions (such as when students construct and run successive cycles of searches). Third, sessions can vary widely in the total number of actions, with an average of 29.4 actions in each sequence and a long-tailed distribution (Min actions = 4, Max actions = 124).

Table 2 Simulated Web Search Tool Action Labels by Subconstruct Categories

Alignment to Inquiry Subconstructs	Description	Action Label	Mean Frequency (SD)
Plan Search	Student constructs or revises their search terms or consults reference materials to generate a query for the search task	Construct Search	8.4% (3.6%)
		View Reference Materials	9.5% (8.9%)
Locate Information	Student submits a search by pressing search button	Run Search	10% (4.2%)
		View Previous Search	5% (3.1%)
	Search results are displayed. Results were coded in terms of their task relevance. Completely off-topic results automatically triggered a search term hint	Results (High-relevance; i.e., two useful websites appear in list)	13.2% (6.7%)
		Results (Medium-relevance, i.e., one useful website appears in list)	9.6% (7.1%)
		Results (Low-relevance, i.e., no useful websites appear in list)	11.2% (7.1%)
		Off-topic Results	6.6% (4.7%)
	Sources (i.e., websites and image results) were coded in terms of their task relevance and reliability	View Source (High-quality, i.e., highly relevant and highly reliable)	4.7% (2.7%)
		View Source (Medium-quality, i.e., partially relevant and mixed reliability)	4.5% (2.8%)
		View Source (Low-quality, i.e., tangentially relevant, and low reliability)	7.4% (4%)
Evaluate Sources	Websites were evaluated by students in terms of Usefulness and Trustworthiness	Evaluate (High-quality, i.e., student rates source as high usefulness and trustworthiness)	5.7% (3.7%)
		Evaluate (Mixed-quality, i.e., student ratings of usefulness and trustworthiness are mixed)	5.0% (3.3%)
		Evaluate (Low-quality, i.e., student rates source as low usefulness and trustworthiness)	4.4% (2.9%)
		Evaluate (No Rating)	3.6% (3.3%)
Interface actions	Start and End Session; these states aid in interpretation of strategies	Start Session	6.4% (5.4%)
		Finish Session	5.9% (4.6%)
	Specific help actions triggered by the system if students appear to be off-task	Timer Alert (i.e., system warns students they have 5 min left to search)	4.7% (4.4%)
		Key Source Resetting (i.e., system provides students with relevant resource)	4.6% (3.8%)
		Search Term Hint (i.e., system provides students with suggested search terms)	6.2% (4.1%)
	Students can ask for help and receive instructions about using the interface	Help (i.e., students can press a button to access contextual help menu)	5.7% (5.6%)
Pauses between actions	Between 2.5 and 10 s	Short Pause	18.9% (7.6%)
	Between 10 and 30 s	Medium Pause	12.2% (5.5%)
	Greater than 30 s	Long Pause	8.6% (5.1%)

Process data logs were translated into 24 action labels alongside descriptions and alignment to the ELA inquiry construct. We also report the mean frequency with which these actions occur within students' search sequences (standard deviation in parentheses reflects variation of these frequencies across sequences)

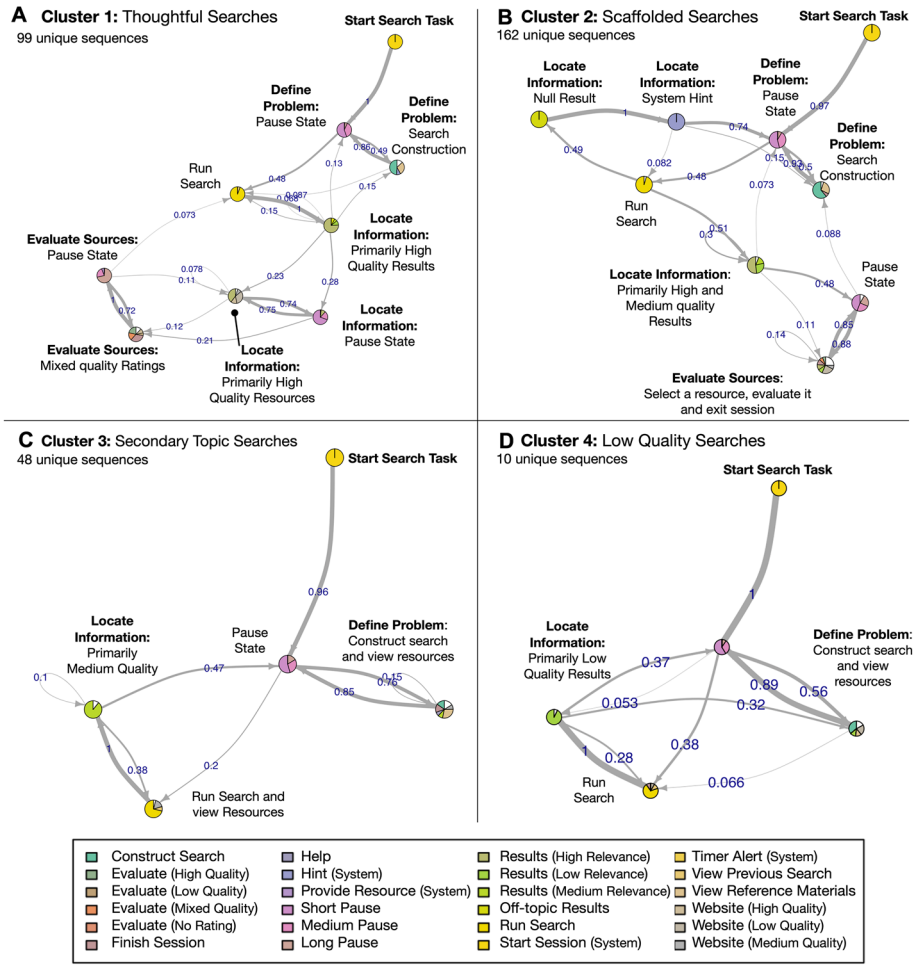

Fig. 5 Node and arrow representation of HMMs fit to sequences from each of the four identified Clusters. Nodes represent hidden states with the color reflecting the probability of the hidden state emitting action events (color coded in the legend along the bottom). Arrows represent the transition probabilities, with labels and density reflecting specific probabilities. For readability we do not display transition probabilities less than .05

Characterizing clustered search behaviors

To address RQ1 we consider the evidence that the four clusters our approach identified reflect distinct search behaviors. Figure 5 provides a visualization of the individual sequences and a node and arrow representation of the four clusters with a brief qualitative description of the different states that students experienced.

Cluster 1: Cluster 1 (n = 99 sequences) appears to reflect a relatively proficient strategy use with 89.9% sessions end with the student having saved one or more sources to the Evidence Manager. A closer look at the outcome of the search processes categorized in Cluster 1 (Table 3) indicates that 87.9% of these sequences contain instances of a student generating search terms that yield a highly-relevant set of results containing the two most useful websites. Of these sequences, 52.5% involve the student viewing a key source and 57.6% involve saving a key source. Many of these sequences also involve viewing (69.7%) and saving (47.5%) at least one low-quality (i.e., irrelevant and/or unreliable) resource as well. The search results available within the highly-relevant results

Table 3 Percent of sequences in each cluster that contained student search, viewing and saving actions of different qualities

Student action	Quality	Percent of sequences (%)			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Search results retrieved from search terms	High relevance	87.9	55.6	4.2	20.0
	Medium relevance	11.1	24.1	45.8	-
	Low relevance	14.1	40.7	-	100
	Off topic results	15.2	80.9	-	20.0
Sources accessed from results list (Click to View)	High quality resource	52.5	25.9	-	10.0
	Medium quality resource	28.3	38.9	37.5	10.0
	Low quality resource	69.7	61.1	14.6	90.0
Sources saved to Evidence Manager (Save to Evidence Manager)	High quality resource	57.6	28.4	4.2	10.0
	Medium quality resource	16.2	27.8	35.4	-
	Low quality resource	47.5	45.1	10.4	40.0
	Saved no resources	10.1	27.8	58.3	50.0
	Resource automatically saved by help function	14.1	13.0	4.2	-

page all mention the historical event that is the focus of the task, but the sources vary in reliability and the number of key task claims addressed. The unexpectedly high rate of viewing and saving some of the less useful websites in Cluster 1 suggest that it may not immediately be apparent to students which of the sources among the highly-relevant results meet their needs, or that students are considering relevance without also jointly considering reliability or applicability to the ultimate goals of the task.

Our HMM model of Cluster 1 consists of 9 latent states and provides a descriptive model of the process by which students interact with the web-search tool (Fig. 5). This model identifies three separate pause states that occur in different contexts, likely representing different processing. Initially, students transitioned between a pause state and constructing a search. This search construction state contains actions around viewing reference materials, typing new search terms, or revisiting previously generated searches. Once a search is issued, we observed potentially thoughtful behavior as students iterated between short pauses and review of results and websites. This pattern ends with students deciding to save a website, followed by being prompted to evaluate its importance, usefulness, and trustworthiness. The pauses increase in length as students complete their evaluations, and from that state they either finish the session or decide to run another search.

Cluster 2: Cluster 2 (n = 162 sequences) is characterized primarily by the presence of an off-topic search yielding off-topic results and system generated hints (Table 3). Unlike Cluster 1, sequences of this type are more likely to terminate without the student saving a resource (27.8%). Most of these sequences involve at least one null search yielding off-topic results (80.9%); however, over half of these sequences also involve a search that returns highly-relevant results (55.6%). This high probability of entering a high-quality search query is likely a result of the hints provided to students on running an off-topic search. Compared with Cluster 1, fewer of these sequences result in the decision to view (25.9%) and subsequently save (28.4%) a key resource. Nearly half (45.6%) of the saves occurred due to a system action triggered by the student running out of time on the Free Roam section of the task that automatically prompted students to save one of the key websites to Evidence Manager before moving on to the Conclusion phase.

Our HMM model captures fewer and shorter pause states in Cluster 2 than we see in Cluster 1 (Fig. 5). Similar to Cluster 1, the first pause state we observe in student's process captures time spent constructing their search terms. Students spend less time in this pause state but transition to comparable states (e.g., Construct search and Run search) with similar probability. Unlike Cluster 1, the large number of null searches justifies the addition of two states to capture student's high probability of running an invalid search. Additionally, the HMM fit to these sequences features a more direct path from the results page to the selection of a website and completion of the evaluation prompts in the Evidence Manager. Rather than estimate two pause states dedicated separately to locating and evaluating sources, our HMM fit to these.clusters identifies a single pause state that students enter after viewing the results page and as they select and evaluate a source. The low probability of transitioning from this state back to revisiting the list of results for more information or constructing a new search is consistent with flimsy behaviors observed in prior research (e.g., Gao et al, 2022; Juvina & Oostendorp, 2008).

Clusters 3: Clusters 3 (n = 48 sequences) capture search sequences that focus on search results related to a secondary topic to be investigated in the inquiry task (i.e., two of the key task claims students must answer are specific to one aspect of the historical event, which could be investigated using scientific, instead of historical sources). One set of results included the third key website that addressed both claims about this secondary topic, and a substantial portion of sessions in this cluster retrieved this set of results (45.8%). These results are considered "medium relevance" (i.e., partially useful) because this search does not yield useful information about the larger context of the event (i.e., the remaining 8 claims about other aspects of the event). Students using the Cluster 3 strategy selected medium (37.5%) or low-quality (14.6%) sources to view, but students frequently abandoned the search session without saving any sources (58.3%; see Table 3). The high rate of abandoning searches without saving accessed sources that do address the task suggests either that these students do not recognize that they need to collect information specific to the secondary topic to fully address the task, or that these students may be aware that the results generated by their searches do not meet all of their information needs but may struggle to construct searches that yield results more useful to the task.

Our HMM models of both Cluster 3 and Cluster 4 have a similar structure containing a single pause state that connects all action states (Fig. 5). The action states our models identify are less clearly separable, with multiple states emitting the same type of event (e.g., two states in Cluster 3 have some probability of emitting actions associated with viewing medium quality results) and the creation of states that collapse two conceptually different events (e.g., both models create a 'Run Search' state that has some probability of also emitting viewing resources and running new searches). While the models identify a single pause state, the hub-and-spoke structure of the model suggests that this pause state reflects a variety of different cognitive processing that occurs between actions. Without more observations of these behaviors or greater consistency in the expression of these behaviors across sessions, we are unable to separate the different pause states using the current approach.

Cluster 4: Sequences classified as Cluster 4 (n = 10 sequences), do not show that same awareness of relevance we observe in Cluster 3 sequences. While there are very few sequences in this cluster, they all include submission of search terms that yield

Table 4 Sample size and scoring of task when different information locating strategies are used

Search Behavior Cluster	Size	Percent of sequences in which any points were earned (%)	Average percent possible points earned (%)	
			Mean	SD
Cluster1	99	65.7	29.3	25.9
Cluster2	162	50.0	18.5	22.1
Cluster3	48	31.3	8.5	14.6
Cluster4	10	20.0	7.5	16.9

Average percent possible points earned reflects how many points a student earns in a search session out of the number of points they have yet to earn within the task. This accounts for the fact that the amount of information available to find diminishes as students locate and accumulate information

low-relevance results pages (Table 3). Students retrieving low-relevance results frequently view (90%) low quality sources and 40% save those sources while the other 50% (appropriately) leave the web search without saving anything. Unlike Cluster 1, where the less useful sources included in the highly-relevant results pages contain some mention of information relevant to the goals of the task, the sources linked within the low-relevance results sets are not useful for completing the goals of the task (i.e., they do not mention and/or are not about the primary or secondary topics and do not address any of the key task claims students must evaluate). The high number of views of irrelevant sources may be indicative of these students' struggles with constructing their search, recognizing relevant information, and allocating their time and attention within the task.

Relationship between search behaviors and performance

To address RQ2, we examined whether the clusters of search behaviors were related to performance on the scenario-based inquiry task. We considered this from two perspectives; how the behaviors contribute to our calculation of score points within the web search tool, and how students' search behaviors relates to their outcome scores on the entire scenario-based task (i.e., across all task phases, locations, and tools), both by task phase, and in terms of the constructs the task was designed to measure (planning, locating, evaluation, synthesis; see Coiro et al., 2018, 2019). For both analyses, we exclude Cluster 4 from our interpretations due to the limited number of observations within this cluster.

Students' actions within the simulated web search tool were scored as providing evidence of their ability to locate information (see Table 1). In Table 4 we report student's point earning across the different clusters. To understand the relationship between point earning and strategy use, we fit a binomial logistic regression using the lme4 package in R (Barr et al., 2013; Bates, 2007). We modeled whether students earned any points within a session as our dichotomous outcome variable. We included in our model the Cluster label and how many points the student had left to earn as fixed effects, along with a nested random effect to account for the fact that sessions are nested within students.[3] The coefficients estimated by this model are scaled in terms of logs, so to improve interpretability we report the coefficients of our model in terms of odds ratio (OR—the

[3] In lme4 syntax this is formulated as anyEarnedPoints ~ ClusterLabel + nPointsLeft + (1|Student/Session).

exponentiated coefficient). We find significant differences between Cluster 1 and 3 (OR 0.35, $z = -2.6$ $p < 0.05$) but not between Cluster 1 and 2 (OR 0.73, $z = -1.1$, $p = 0.26$). The odds of earning one or more points from a search is estimated to be 65% lower (95% CI of OR 0.16–0.77) for Cluster 3 than Cluster 1. We also observe that the number of points a student has left to earn has a significant impact on whether they will earn points with any strategy (OR 0.67, $z = -5.1$, $p < 0.001$). This suggests that sequences classified as Cluster 1, contribute to higher scores on this subtask than Cluster 3.

Non-parametric Spearman correlations examining the proportion of sessions within each cluster indicated distinct patterns of relationships with student-level performance in terms of task outcomes; Table 5 reports these correlations for the total task, as well as subscores by task phase and by construct. Having a higher proportion of sessions in Cluster 1 was associated with higher.performance overall and across all phases of the task, while a higher proportion of sessions in Cluster 2 was associated with lower overall and phase-level performance. Cluster 1 was especially associated with higher scores for searching, evaluation, and synthesis activities, while Cluster 2 was associated with lower scores for these constructs, as well as lower scores for planning within the Setup phase. The proportion of sessions in Cluster 3 was positively associated with planning and questioning subscores and negatively associated with searching subscores but was not related to overall or phase-level performance. The proportion of Cluster 4 sessions showed weak-to-no relationships to scores.

Search behaviors in task context

With our final research question (RQ3) we consider how search behaviors change throughout the task. One of the challenges in modeling search behaviors within the ELA virtual world is that students can engage with the simulated web search tool at various points throughout the task and can submit an unlimited number of searches within the time constraints of the assessment. This changing task context will likely influence the type of strategies students use when they engage with the web search tool. Due to the design of the task, students can only access the three key websites via the web search tool (unless they time out of the Free Roam section and are provided one of the sources in order to move on), so, when first entering and interacting with this tool, students should have the same underlying information need to locate one or more of those key websites. These information needs are needs demanded by the task. While initial needs may be similar, prior research suggest student's awareness and understanding of their task goals can vary greatly (Rouet et al, 2021). Over half of students' first sessions are classified in Cluster 2 (56.9%; see Fig. 6). Cluster 1 occurs less frequently (27% of first sessions) but among those individuals whose first sessions are classified as this behavior, 75% can locate key information within that first visit and never revisit the web search.

We can look at how students who visit the web search tool multiple times throughout the task transition between clusters across their search sessions. Table 6 presents the probability of students transitioning from one cluster to another throughout the task into a finished state (i.e., nomore searching is conducted). We see that students who conduct a search using the Cluster 1 session are most likely to transition into a finished state (52% probability). Students who have run a high-quality search and continue to engage

Table 5 Spearman Correlations (ρ) of Inquiry Task Performance Variables with Students' Proportion of Search Sessions Classified into Each Cluster

	Proportion sessions in cluster 1 (n = 99)	Proportion sessions in cluster 2 (n = 162)	Proportion sessions in cluster 3 (n = 48)	Proportion sessions in cluster 4 (n = 10)
Total task score				
Inquiry task total score (max: 100)	**0.310**	**− 0.325**	0.035	0.019
Task phase-level subscores				
Task phase: setup (max: 12)	**0.228**	**− 0.251**	0.066	− 0.017
Task phase: free roam (max: 51)	**0.209**	**− 0.260**	0.044	0.098
Task phase: conclusion (max: 37)	**0.203**	**− 0.259**	0.117	− 0.005
Construct subscores				
Subconstruct: planning (max: 6)	0.100	**− 0.240**	**0.207**	0.083
Subconstruct: locating (max: 22)	0.150	− 0.197	0.084	0.015
Locating: Questioning (max: 7)	− 0.053	− 0.089	**0.202**	0.083
Locating: Searching (max: 4)	**0.407**	**− 0.254**	**− 0.226**	− 0.041
Locating: Choosing Sources (max: 5)	0.095	− 0.112	0.070	− 0.053
Locating: Saving Sources (max: 6)	0.117	− 0.176	0.112	0.001
Subconstruct: Evaluating (max: 35)	**0.260**	**− 0.275**	− 0.019	0.099
Evaluating: Importance (max: 7)	**0.215**	**− 0.257**	0.029	0.095
Evaluating: Usefulness (max: 14)	**0.247**	**− 0.236**	− 0.038	0.052
Evaluating: Trustworthiness (max: 14)	**0.225**	**− 0.245**	− 0.025	0.118
Subconstruct: Synthesis (max: 37)	**0.203**	**− 0.259**	0.117	− 0.005

Values exceeding |.20| appear in boldface. Subscores for the Free Roam phase and the Locating and Evaluating subconstructs reflect students' performance on the simulated web search tool, in addition to a simulated library search tool and simulated conversations with virtual characters (Coiro et al., 2018, 2019)

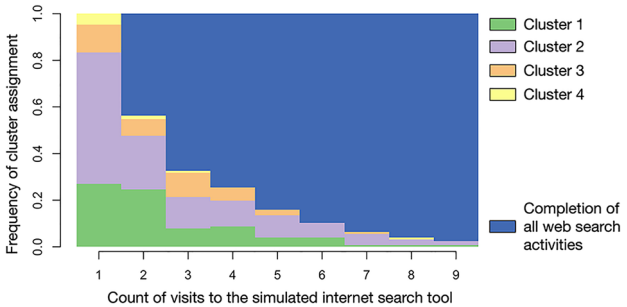

Fig. 6 Proportion of Sequences Classified as One of the Four Clusters

Table 6 Transition probability matrix of search strategies used throughout the task

	Cluster assignment of next web search session				
	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Done**
Cluster assignment of current web search session					
Cluster1	0.19	0.19	0.09	0	0.53
Cluster2	0.19	0.35	0.08	0.02	0.37
Cluster3	0.27	0.29	0.19	0	0.25
Cluster4	0.30	0.20	0.20	0.10	0.20

with the tool are equally likely to return and run another Cluster 1 or 2 search. Students that apply a search behavior classified as Cluster 2 are less likely to reach a finished state (37%) and their next search is likely to either be another Cluster 2 or 1 search behavior. Students use search behaviors classified as Cluster 3 and Cluster 4 are likely to continue using the web search tool to run more searches; from Cluster 3, it is more likely for their next search to be classified as Cluster 1 or 2 rather than another Cluster 3 type search.

We applied a mixed effects logistic regression to test whether students who have information left to find within the web search tool are more likely to return to conduct another search than students who have met their information needs. We are specifically interested in whether there is an interaction such that students who use the Cluster 1 behavior are more sensitive to the information needs than students who exhibit other search behaviors. Again, Cluster 4 was excluded from this analysis due to limited observations. We treated whether the student returned to conduct another search after completing the current search as our dependent variable. We included in our model the Cluster label of the current search behavior (1–3) and the number of remaining points (0–8) students could earn within the tool after completing their current search (see Table 1 for scoring rules). We also included in our model the interaction between current strategy cluster and points remaining, and a random intercept for students.[4] We fit the mixed effects logistic regression with nested random effects to capture sessions within students using the lme4 package in R (Bates, 2007).

In our model, we accounted for a significant amount of the variance predicting students' return to use the tool with a fixed effects for current strategy cluster and the number of points left to earn as well as significant interactions between points remaining and clusters (Table 7). This analysis shows that for all three clusters the probability of returning to use the tool is greatest when students have the most information remaining to locate (quantified in terms of points remaining to earn for searching, viewing, and saving actions based on the current task scoring rules). This effect was strongest for sequences classified as Cluster 1. This suggests that after executing a Cluster 1 search students are more sensitive to their individual, task-specific information needs than Cluster 2 or 3.

[4] In lme4 syntax this model is formulated as ReturnsToSearch ~ ClusterLabel * nPointsLeft + (1|Student/Session).

Table 7 Mixed effects logistic regression predicting the probability of returning to use web-search tool

Predictors	Odds Ratios	logit(p_{return}) CI	p
(Intercept)	0.01	0.01–0.01	< 0.001
ClusterLabel [Cluster 2]	12.63	12.55–12.70	< 0.001
ClusterLabel [Cluster 3]	57.72	57.37–58.07	< 0.001
nPointsLeft	2.46	2.45–2.48	< 0.001
ClusterLabel[Cluster 2]* nPointsLeft	0.61	0.60–0.61	< 0.001
ClusterLabel[Cluster 3]* nPointsLeft	0.50	0.50–0.51	< 0.001
$N_{Session}$	11		
N_{UserID}	124		
Observations	309		
Marginal R^2 /Conditional R^2	0.238/0.339		

Discussion

In this study we consider some of the limitations of current assessments of DIL and present an assessment task that seeks to create a realistic context in which students are tasked with locating, comprehending, evaluating, synthesizing, and applying information in response to a scenario-based multiple-source inquiry task in a richly interactive virtual world platform. The scoring rules we developed when creating this task (e.g., Table 1) capture key indicators we would expect to be produced by students who are proficient with the different decision-making processes and search strategies that would underlie successful demonstrations of multiple-source inquiry skills within the ELA task. These scoring rules however do not consider the processes by which students arrive to these different choices. In this study, we brought together theory-driven and data-driven elements to understand how process data may provide greater insight into the information search skills of students. Using theoretically-grounded approaches to analyze process data we aimed to characterize differences in students' search behaviors and identify whether there are systematic differences between the task outcome measures and patterns in student behavior within the task.

From search behaviors to information search strategies

With RQ1 we sought insight into the different information search strategies students used when engaging with the simulated web search tool. We used an unsupervised clustering approach that combines edit-distance clustering and mixture hidden Markov models (mHMMs) to identify groups of sequences that shared similarities in the quality of the materials interacted with, the time spent on different actions, and the context of those actions. These clusters provide a descriptive account of the different search behaviors in our task. We suggest descriptive labels for the types of search behaviors captured by each cluster: thoughtful searches, scaffolded searches, secondary topic searches, and low-quality searches. Students who construct well-targeted search terms on entering the search tool (Cluster 1; *Thoughtful Search*) are much more likely to view and save useful resources. Students who construct search terms that are off-topic receive strong feedback from the system, including a noninteractive list of completely irrelevant results and system generated hints designed to encourage more task-relevant search terms (Cluster

2; *Scaffolded Search*). Although these hints guide students to primarily high-relevance results, we continue to see differences between these two strategies in terms of the interactions with resources. Students using the *Thoughtful Search* behavior more frequently view and save useful resources than students using the *Scaffolded Search* behavior. In both search behaviors we observe students frequently viewing and saving less useful results as well. This type of behavior may indicate a lack of cognitive development or prior educational experiences affecting students' ability to distinguish the content's relevance and reliability from cues present within the search results descriptions and within the sources themselves (Rouet et al., 2021).

The other two search behaviors we identify capture students' searching primarily medium-relevance (Cluster 3; *Secondary Topic Searches*) and low-relevance results (Cluster 4; *Low Quality Searches*). Without the scaffolding support generated by the system in response to an off-topic search, students using these two behaviors are much more likely to abandon their search without viewing or saving any sources. Our *Secondary Topic* search behavior (Cluster 3) captures students who construct search queries that focus on a secondary topic within the task (i.e., information relevant to only two of the key task claims). This cluster may reflect either a *flimsy* navigation pattern or a *satisficing* search strategy. Students using flimsy navigation patterns only consider a few resources before ending their search prematurely. This is believed to be due to issues surrounding student's lack of motivation (Gao et al., 2022) or disorientation within the interface (Juvina & Oostendorp, 2008). The satisficing strategy, on the other hand, involves sequential evaluation of the sources against a specific goal and minimal exploration of available resources. These goals are often set prior to the search and if this acceptability threshold is not met by the set of results, students may draw conclusions about the general usefulness of the information space and choose to abandon that location to search elsewhere (List & Alexander, 2017; Wilkinson et al., 2012). While the satisficing strategy has been described as an effective navigation pattern (Gao et al., 2022), if students struggle to identify the primary topic of the task the poor alignment between their task goals and the information provided by the web search tool may cause students using this strategy to abandon their searches when it is inappropriate to do so. In our study, students using the Secondary Topic search behavior frequently abandon their searches without saving any sources (i.e., perhaps overlooking the relevance of the partially useful source to two of the key task claims). They also appear more selective in viewing resources, viewing, and saving low utility resources at much lower rates compared to the other behavior clusters. While the high rates of search abandonment are consistent with a flimsy navigation pattern, if these searches are abandoned because sources do not meet students' goals, this behavior is in line with a satisficing strategy. More research is needed to distinguish between these two behaviors. Such research would benefit from insight into student's goals and how students evaluate the individual items on the results pages.

The least frequently observed strategy, *Low Quality Searches* (Cluster 4), captures students who run searches that yield low relevance results and choose to view and save resources that are not useful, relevant, or reliable. These individuals seem to struggle across all dimensions of the task both in their creation of search terms, their sensitivity to cues about information reliability, and their ability to assess information relevance.

With so few sequences classified in this cluster it is difficult to know if this strategy would generalize to a different dataset.

Strategic behavior reflects deliberate decisions and actions taken to achieve specific goals. Within the space of multiple document comprehension and digital literacy, such strategies are often discussed in terms of the individual's effort, intentionality, and goal orientation (Afflerbach & Cho, 2009; Afflerbach et al., 2008; Cho, Afflerbach, & Han, 2018). The strategies students use are responsive to both the digital information environment and the continuously updated mental model of the student (Cho & Afflerbach, 2017; Rouet, 2006; Rouet & Britt, 2011), suggesting that student's search behaviors reflect larger processing chains of strategies that interact and influence each other as information is uncovered (Cho et al., 2018). Across the four search clusters we identify within this task we see evidence of strategic behaviors; however, it is unlikely that each of the clusters represents a singular strategy. A limitation of our current approach is that it is challenging to separate the individual strategies that students use throughout the search process from the larger processing chain which our model appears to capture. Additionally, without insight into the mental processes of the student (e.g., as captured by think-aloud protocols; Keehner et al., 2017), decoding how these behaviors reflect strategies observed in different task environments involves a degree of subjectivity. One promising direction for future research would be to generate example sequences within this task reflecting common search and navigation strategies and use an edit-distance based approach to identify if similar patterns exist within our observed data. This would extend the approach used by He and colleagues (2021) to calculate distance from the optimal solution to detect strategies.

Relationship between search behaviors and task performance

In addressing RQ2, we found evidence that while search behaviors differed in their contribution to students' total scores, correlations between the use of these behaviors and overall assessment scores were generally weak-to-moderate. Rouet et al. (2021) identify three critical demands of DIL tasks that pose significant challenges for young people: "(a) the need of information users to understand their task and to generate and update their search goals accordingly; (b) the need to use proximal and distal cues in order to access information of interest while minimizing the time spent processing irrelevant information; (c) the need to assess the adequacy and sufficiency of information with respect to the end goal and/or product" (p. 5). In piloting the ELA scenario-based task, we observed the presence of these challenges with some students struggling to locate the most useful sources and complete all activities within the 90-min administration window (Sparks et al., 2018). The hints and alerts we built into the task to support students in understanding the task goals and managing their time-on-task, are also many of the events that distinguish between the four search behaviors we identified in our analysis. Students were not penalized for receiving these nudges in the task scoring rules. A promising direction for future work is to explore joint modeling approaches for scoring ability using both response accuracy and hint use (e.g., Bolsinova et al., 2022). Based on our investigation of RQ2 we hypothesize that alternative scoring models such as assigning partial credit to sources found as a result of scaffolds and considering student's

sensitivity to scaffolding when information is missed may increase the sensitivity of the ELA scoring model to the challenges described by Rouet et al. (2021).

Search behaviors in context

Our third research question was focused on understanding how the use of search behaviors evolved within the changing context of the task. As students used the tools within the ELA virtual world environment, we expected to see changes in their web search strategies driven by the accumulation of useful information for the task. *Thoughtful Search* behaviors occurred primarily within the first few visits and had a high probability resulting in students finishing their use of the web search tool. Overall, we see that students are more likely to return to the web search tool when there was additional relevant information in the tool they had yet to view or save, however sessions classified as *Thoughtful Search* showed the greatest sensitivity to how much information was left to locate. Sessions classified as *Scaffolded Search* and *Secondary Topic* were less likely to conclude students' use of the web search tool. In comparison with *Thoughtful Searches*, students who ran one of these searches were more likely to return to the web search tool when they had already exhausted all relevant information or finish using the web search tool when there was information yet to find. This difficulty in identifying whether the information gained within a search is sufficient to complete a task is a well-documented challenge for students (Rouet et al., 2021). In future iterations of our task design, we could add additional supports that provide students who repeatedly use these search behaviors with explicit feedback concerning how much information they have left to find and where they should look. As discussed in the previous paragraph, not only could this type of support may help students complete the task within the administration window, but we could also use this information to extend our scoring model to account for the use of these triggers in our estimation of student skill.

A Cognition-centered approach

Throughout this study we aimed to employ a theory-based approach from assessment design to data analysis and interpretation (Arslan et al., 2023; Goldhammer et al., 2021; Kroehne & Goldhammer, 2018). Following recommendations for valid interpretation of our process data (Goldhammer et al., 2021), we outlined the connections between our target attribute (multiple-source inquiry), the design of our task and behaviors it elicits and the construction of process indicators that provide empirical evidence of these behaviors of interest. Reflecting on our approach there were several areas where theory was especially helpful in guiding data-driven analysis of process data as well as areas where it was unclear how theory should contribute to how we analyze, interpret, and use process data. In this final section we consider the challenges and highlight areas where future research can support how we incorporate theory in the use of data-driven methods.

The role of theory when applying data-driven methods

The development of complex, interactive assessments creates an opportunity to use a wide variety of data mining approaches to make sense of students' processes and performances. While these approaches have been used frequently to study instructional

data (e.g., intelligent tutoring systems; Pardos, 2017), using these approaches to support the design, use, and interpretation of large-scale assessments poses a unique set of challenges. In their recent paper on theory-driven construction of process data indicators, Goldhammer et al. (2021) outline some guidance for how assessment frameworks such as ECD could be extended to inform the creation of process indicators. We found building from this approach especially valuable when identifying how to represent the process we used to fit our models. Process data generated by the scenario-based ELA task capture a variety of behaviors from task specific problem solving to interactions with the interface. While we focus on theories of multi-document comparison and information search, our analysis approach provides a path by which we could capture evidence consistent with alternative theories (e.g., information foraging theory) and consider other types of behaviors. In future research, we could adapt the event representation to reflect other information we believe students use to drive their decisions such as the order in which resources are presented in the search results returned by the web search tool (Kammerer & Gerjets, 2012).

As we explore new data-driven approaches it is important to consider the role of theory in informing and evaluating the modeling choices we make. In the current study, there were several decisions we made in building this analysis where it was less clear how theory could be used to inform our approach. For example, we used an edit-distance approach originally developed to support natural language processing but increasingly used within large-scale assessments to identify strategies in process data (for review Goldhammer et al., 2020). Identifying an edit distance metric that was appropriate for our hypotheses about strategy use and sensitive to indicators of strategy within our log data was challenging. We found that LCS, a popular edit distance metric used in prior studies using PS-TRE data, was sensitive to small differences in navigation patterns and formed clusters that primarily captured variation in sequence length. In designing our task, we provided students with navigational freedom to choose between a variety of actions. Given this freedom we expected variability in how students searched for information. Applying a new literacies perspective, we recognized that while the specific order of actions might vary, the context in which actions occurred was an important reflection of variation in how students construct their knowledge. This led us to select an alternative approach, normalized OM, which was better suited for our dataset since it expressed greater sensitivity to patterns cooccurrence between these events.

Even so, the limitation of this edit-distance approach in capturing the structure of information search cycles that are present in models of multiple-document use like MD-TRACE (Rouet & Britt, 2011), led us to explore the use of mHMMs to detect meaningful between group differences in timeseries processes. Throughout this process of selecting and applying modeling methods we looked to MD-TRACE to guide our expectations; however, it was not always clear how best to modify our data-driven approaches to capture student's search processes. In the same way we consider how theory should inform the creation of process indicators (Goldhammer et al., 2021), we should also consider how theory can best inform the selection and use of data driven approaches. Not only would the field benefit from comparisons across methods and model parameterizations (e.g., Lesnard, 2009), but future research using these approaches should explicitly

discuss how modeling choices capture properties of the task, assessment instrument and cognitive processes underlying student behavior (e.g., Luu et al., 2020).

The role of theory when validating data-driven models

A traditional validation strategy, *nomothetic span* involves showing a relationship between the process indicator and a standardized measure measuring a similar construct (Embretson, 1983; Keehner et al., 2017; Goldhammer et al., 2021). Much of the prior research using unsupervised approaches to model strategies in large scale assessments consider nomothetic span by showing differences in the item score (e.g., Gao et al., 2022) or assessment score (e.g., He et al., 2021). The challenges we face in establishing nomothetic span in the current study is in part related to the challenge of assessing a complex set of competencies. While environments like the ELA Virtual World are designed to elicit behaviors and component abilities of the multiple-source inquiry construct, these abilities may come interact within an open task to create experiences that can differ widely between individuals. This interaction between subconstructs/components is an important aspect of contemporary literacy and DIL and is reflected in models of multiple-document use such as MD-TRACE. However, efforts to design performance-based assessments that attempt to simulate real world contexts should also explore building and validating more nuanced scoring models that capture the complex interactions between task variability and individual variability. Future research would benefit from measuring the lower-level skills hypothesized to drive individual differences in strategy use. For example, identifying whether measures such as individuals' awareness of cues to the usefulness of information for their task (e.g., relevance and reliability) and sensitivity to costs (e.g., time, effort), explain variance in student's use of different search behaviors would provide evidence of nomothetic span while maintaining a sensitivity to the cognitive processes underlying the construct.

An alternative approach for establishing construct validity, the *construct representation* approach, seeks to explain task variability through the theoretical cognitive mechanisms underlying the task (Embretson, 1983). Computational cognitive models offer a method for establishing a clear expectation for how differences in cognitive processing would result in different observable behaviors when completing the task. Establishing that the variance detected in process data reflects variation in cognitive processing becomes much more challenging when using data-driven approaches. In the current study, we found evidence that our clustered search behaviors capture differences in student behavior that are consistent with theories of how context and ability influence information search processes (Rouet et al., 2021). While this analysis provides a descriptive view of how students engaged with the web search tool, we are limited in the claims we can make with this type of evidence. One approach to establish construct representation validity evidence, is to introduce variation in task characteristics or experimental designs to create conditions under which differences should be present and detectable (Goldhammer et al., 2021). Applying this approach in complex tasks such as the current ELA task risks imposing constraints which change the realism of the task. A promising direction for future research would be to revisit Embretson's (1983) original suggestion and take advantage of computational cognitive models such as SNIF-ACT and ACT-R to build models of task performance (Anderson, 2009; Pirolli & Fu, 2003; Ritter et al.,

2019). Such models would provide us with a way to generate predictions that are sensitive to an individual's unique path through the task and capture how student behavior reflects their current task goals.

The ELA Virtual World was developed to explore what future large-scale assessments of DIL competencies could look like if they reflected the complexities of contemporary theory and practice. In this paper we discuss how the theory which informed this task design should similarly inform how we model the process data generated from that task. The primary contribution of the theoretically-grounded approach we describe is to provide a qualitative understanding of student strategy use in a complex, dynamic simulation-based and scenario-based environment. This reflects ongoing research on how these search behaviors can be used to complement and contextualize quantitative scores generated from complex task environments, yielding a more nuanced picture of students' DIL proficiency as estimated from integrated performances within online multiple-source inquiry tasks. Based on these results we identify future directions for modifying the scenario-based ELA task to better capture differences in student search strategies and to improve our measurement model to provide a more nuanced view of the constituent subconstructs required for skilled performance.

Acknowledgements
We gratefully acknowledge the contributions of Brian Young, Colleen Appel, Hilary Persky, Heather Nadelman, Irv Katz, Gary Feng, Madeleine Keehner, Julie Coiro, Jody Underwood, and Intelligent Automation, Inc. in supporting the design and development of the ELA Virtual World; Colleen Appel, Ted Kolwicz, Pavan Pilarisetti, Keith Kiser, Doug Stein, and Metacog for supporting the data collections; and Jie Gao, Mengxiao Zhu, Rafael Quintana, and Jonathan Steinberg for supporting tryout study analyses. We are especially thankful to the middle school students who participated in and contributed to our iterative development and pilot research. The authors would like to thank Dr. Madeline Keehner and Dr. Priya Kannan for providing valuable comments and suggestions on earlier versions of this manuscript.

Author contributions
CT lead the conceptualization of the paper, methodology, formal analysis, writing. J.R.S. led the conceptualization and design of the ELA Virtual World, scored and analyzed task responses, analyzed and interpreted the correlation results, and was a major contributor in writing the manuscript. Both authors read and approved the final manuscript.

Funding
The ELA Virtual World was developed, piloted, and scored through funding from the National Center for Education Statistics as part of the National Assessment of Educational Progress (NAEP) Survey Assessment Innovations Laboratory (SAIL) research initiative. The current analyses of process data were funded in part by Educational Testing Service. Views expressed reflect those of the authors and not of the funding institutions.

Availability of data and materials
The datasets generated and analyzed during the current study are available from the corresponding author on request.

Declarations

Competing interests
The authors declare they have no conflict of interest.

Received: 30 March 2022 Accepted: 23 April 2023
Published online: 24 July 2023

References

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History, 16*(3), 471–494.

Afflerbach, P., & Cho, B. (2009). Identifying and describing constructively responsive comprehension strategies in new and traditional forms of reading. In S. Israel & G. Duffy (Eds.), *Handbook of reading comprehension research* (pp. 69–90). Mahwah, NJ: Erlbaum Associates.

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher, 61*(5), 364–373.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Arslan, B., Tenison, C., & Finn, B. (2023). Going beyond observable actions: a cognition-centered approach to interpreting pauses represented in process data. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000756
Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*(3), 523–535.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.
Bates, M. D. (2007). The lme4 package. *R Package Version, 2*(1), 74.
Batool, F., & Hennig, C. (2021). Clustering with the average silhouette width. *Computational Statistics & Data Analysis, 158*, 107190.
Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 37*(6), 1554–1563.
Bolsinova, M., Deonovic, B., Arieli-Attali, M., Settles, B., Hagiwara, M., & Maris, G. (2022). Measurement of ability in adaptive learning and assessment systems when learners use on-demand hints. *Applied Psychological Measurement, 46*(3), 219–235.
Boroujeni, M. S., & Dillenbourg, P. (2019). Discovery and temporal analysis of MOOC study patterns. *Journal of Learning Analytics, 6*(1), 16–33.
Braasch, J. L. G., Rouet, J.-F., Vibert, N., & Britt, M. A. (2012). Readers' use of source information in comprehension. *Memory & Cognition, 40*(3), 450–465.
Brand-Gruwel, S., & Stadtler, M. (2011). Solving information-based problems: evaluating sources and information. *Learning and Instruction, 21*(2), 175–179.
Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Computers in Human Behavior, 21*(3), 487–508.
Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*, 485–522.
Britt, M. A., & Rouet, J.-F. (2012). Learning with multiple documents: component skills and their acquisition. In M. J. Lawson & J. R. Kirby (Eds.), *Enhancing the quality of learning: dispositions, instruction, and learning processes* (pp. 276–314). Cambridge University Press.
Britt, M. A., Rouet, J.-F., & Durik, A. M. (2018). *Literacy beyond text comprehension: a theory of purposeful reading*. Routledge.
Cho, B. Y., & Afflerbach, P. (2017). An evolving perspective of constructively responsive reading comprehension strategies in multilayered digital text environments.
Cho, B. Y., Afflerbach, P., & Han, H. (2018). Strategic processing in accessing, comprehending, and using multiple sources online. Handbook of multiple source use. Routledge.
Coiro, J. (2011). Predicting reading comprehension on the Internet: contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research, 43*, 352–392.
Coiro, J. (2020). Toward a multifaceted heuristic of digital reading to inform assessment, research, practice, and policy. *Reading Research Quarterly, 56*(1), 9–31. https://doi.org/10.1002/rrq.302
Coiro, J., & Dobler, E. (2007). Exploring the on-line reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the internet. *Reading Research Quarterly, 42*(2), 214–257.
Coiro, J., Sparks, J. R., & Kulikowich, J. M. (2018). Assessing online collaborative inquiry and social deliberation skills as learners navigate multiple sources and perspectives. In J. L. G. Braasch, I. Braten, & M. T. McCrudden (Eds.), *Handbook of Multiple Source Use* (pp. 485–501). Routledge.
Coiro, J., Sparks, J.R., Kiili, C., Castek, J., Lee, C-H., & Holland, B.R. (2019). Students engaging in multiple-source inquiry tasks: Capturing dimensions of collaborative online inquiry and social deliberation. *Literacy Research: Theory, Method, and Practice*. https://doi.org/10.1177/2F2381336919870285
Costa, M., & De Angelis, L. (2010). Model selection in hidden Markov models: a simulation study.
Dias, J. G., Vermunt, J. K., & Ramos, S. (2015). Clustering financial time series: new insights from an extended hidden Markov model. *European Journal of Operational Research, 243*(3), 852–864.
Edelson, D. C. (2002). Design research: what we learn when we engage in design. *The Journal of the Learning Sciences, 11*(1), 105–121.
Elzinga, C. H. (2006). Sequence analysis: Metric representations of categorical time series. Sociological methods and research.
Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). IEA international computer and information literacy study 2018 assessment framework. *Springer*. https://doi.org/10.1007/978-3-030-19389-8
Gabadinho, A., Ritschard, G., Mueller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1–37.
Gao, Y., Cui, Y., Bulut, O., Zhai, X., & Chen, F. (2022). Examining adults' web navigation patterns in multi-layered hypertext environments. *Computers in Human Behavior, 129*, 107142.
Gerjets, P., & Kammerer, Y. (2010). Topical relevance and information quality in cognitive models of Web search behavior: introducing epistemic scent into information foraging theory. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 32, No. 32).
Goldhammer, F., Hahnel, C., & Kroehne, U. (2020). *Analysing log file data from PIAAC large-scale cognitive assessment*. Cham: Springer.
Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: on validating the interpretation of process indicators based on log data. *Large-Scale Assessments in Education, 9*(1), 1–25.
Goldman, S. R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Ferris & D. M. Bloome (Eds.), *Uses of intertextuality in classroom and educational research*. Greenwich, CT: Information Age.
Goldman, S. R., Blair, A., & Burkett, C. M. (2018). Assessment of multiple resource comprehension and information problem solving. In J. L. G. Braasch, I. Bråten, & M. T. McCrudden (Eds.), *Handbook of multiple source use* (pp. 466–484). Routledge.

Goldman, S. R., Braasch, J. L. G., Wiley, J., Graesser, A. C., & Brodowinska, K. M. (2012). Comprehending and learning from Internet sources: processing patterns of better and poorer learners. *Reading Research Quarterly, 47*(4), 356–381.
Goldman, S. R., Lawless, K. A., Gomez, K. W., Braasch, J. L. G., MacLeod, S. M., & Manning, F. (2010). Literacy in the digital world: comprehending and learning from multiple sources. In M. G. McKeown & L. Kucan (Eds.), *Bringing reading research to life.* New York: Guilford.
Goldman, S. R., Lawless, K., & Manning, F. (2013). Research and development of multiple source comprehension assessment. In M. A. Britt, S. R. Goldman, & J.-F. Rouet (Eds.), *Reading from words to multiple texts* (pp. 180–199). New York: Routledge.
Goldman, S. R., Lawless, K., Pellegrino, J., Manning, F., Braasch, J., & Gomez, K. (2011). A technology for assessing multiple source comprehension: An essential skill of the 21st century. In M. C. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: theoretical and practical implications from modern research* (pp. 173–210). Charlotte: Information Age Publishing.
Graesser, A. C., Wiley, J., Goldman, S. R., O'Reilly, T., Jeon, M., & McDaniel, B. (2007). SEEK web tutor: fostering a critical stance while exploring the causes of volcanic eruption. *Metacognition Learning, 2*, 89–105.
Hahnel, C., Kroehne, U., Goldhammer, F., Schoor, C., Mahlow, N., & Artelt, C. (2019). Validating process variables of sourcing in an assessment of multiple document comprehension. *British Journal of Educational Psychology, 89*(3), 524–537.
Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: an edit distance approach. *Journal of Educational Data Mining, 7*(1), 33–50.
He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education, 166*, 104170.
He, Q., Liao, D., & Jiao, H. (2019). Clustering behavioral patterns using process data in piaac problem-solving items. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement*. Cham: Springer International Publishing.
He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development*. Pennsylvania: IGI Global.
Helske, S., Helske, J., & Eerola, M. (2016). Analysing complex life sequence data with hidden Markov modelling. In International Conference on sequence analysis and related methods. LIVES-Swiss national centre of competence in research; swiss national science foundation; Université de Genevè.
Helske, S. (2021). Examples and tips for estimating Markovian models with seqHMM.
Helske, S., & Helske, J. (2017). Mixture hidden Markov models for sequence data: the seqHMM package in R. *arXiv*. https://doi.org/10.4855/arXiv.1704.00543
Hinostroza, J. E., Ibieta, A., Labbé, C., & Soto, M. T. (2018). Browsing the Internet to solve information problems: a study of students' search actions and behaviours using a 'think aloud'protocol. *Education and Information Technologies, 23*(5), 1933–1953.
Jurafsky, D., & Martin, J. H. (2020). Speech and language processing: *an introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
Juvina, I., & van Oostendorp, H. (2008). Modeling semantic and structural knowledge in web navigation. *Discourse Processes, 45*(4–5), 346–364.
Kammerer, Y., & Gerjets, P. (2012). How search engine users evaluate and select Web search results: the impact of the search engine interface on credibility assessments. In D. Lewandowski (Ed.), *Web search engine research*. Bingley: Emerald Group Publishing Limited.
Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2017). Developing and validating cognitive models in assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 75–101). Hoboken: John Wiley & Sons.
Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika, 45*(2), 527–563.
Lawless, K. A., Goldman, S. R., Gomez, K., Manning, F., & Braasch, J. (2012). Assessing multiple source comprehension through evidence centered design. In J. P. Sabatini, T. O'Reilly, & E. R. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 3–17). Rowman & Littlefield.
Lesnard, L. (2009). Cost setting in optimal matching to uncover contemporaneous socio-temporal patterns. ffhalshs-00435428f
Leu, D. J., Jr., Coiro, J., Castek, J., Hartman, D. K., Henry, L. A., & Reinking, D. (2008). Research on instruction and assessment of the new literacies of online reading comprehension. In C. C. Block, S. Parris, & P. Afflerbach (Eds.), *Comprehension instruction: Research-based best practices*. New York: Guilford Press.
Leu, D. J., Jr., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (2013). New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment. In D. Alvermann (Ed.), *RB Ruddell Theoretical models and processes of reading. International Reading Association* (pp. 1150–1181). DE: Newark.
List, A., & Alexander, P. A. (2017). Cognitive affective engagement model of multiple source use. *Educational Psychologist, 52*(3), 182–199.
Liu, H., Janssen, J., & Milios, E. (2006). Using HMM to learn user browsing patterns for focused Web crawling. *Data & Knowledge Engineering, 59*(2), 270–291.
Luu, V. T., Forestier, G., Weber, J., Bourgeois, P., Djelil, F., & Muller, P. A. (2020). A review of alignment based similarity measures for web usage mining. *Artificial Intelligence Review, 53*(3), 1529–1551.
Martin, P., & Wiggins, R. D. (2011). Optimal matching analysis. The sage handbook of innovation in social research methods, 385–408.
McCrudden, M. T., Magliano, J. P., & Schraw, G. (2010). Exploring how relevance instructions affect personal reading intentions, reading goals and text processing: a mixed methods study. *Contemporary Educational Psychology, 35*(4), 229–241.
McCrudden, M. T., Magliano, J. P., & Schraw, G. (2011). *Text relevance and learning from text*. Greenwich CT: Information Age.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication, 60*, 413–439.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29.
Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). PIRLS 2016 Assessment Framework (2nd ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/pirls2016/framework.html
Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). PIRLS 2021 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2021/frameworks/
National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common Core State Standards for English Language Arts*. Washington, DC: Authors
Organisation for Economic Cooperation and Development. (2013). Skills outlook 2013: First results from the survey of adult skills. https://doi.org/10.1787/9789264204256-en
Organisation for Economic Cooperation and Development. (2019a). *PISA 2018* Assessment and Analytical Framework. PISA: Paris, France: OECD Publishing. https://doi.org/10.1787/b25efab8-en
Organisation for Economic Cooperation and Development. (2021). *21st-Century Readers: Developing Literacy Skills in a Digital World.* PISA. Paris, France: OECD Publishing. https://doi.org/10.1787/a83d84cb-en.
Organisation for Economic Cooperation and Development. (2019b). The survey of adult skills: reader's companion. *Third Edition, OECD Skills Studies, OECD Publishing, Paris.* https://doi.org/10.1787/f70238c7-en
Paquette, L., de Carvalho, A. M., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. In *CogSci*.
Pardos, Z. A. (2017). Big data in education and the models that love them. *Current Opinion in Behavioral Sciences, 18*, 107–113.
Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. V. Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 88–108). Lawrence Erlbaum.
PIAAC Expert Group in Problem Solving in Technology-Rich Environments. (2009). PIAAC Problem Solving in Technology-Rich Environments A Conceptual Framework OECD Education Working Papers, No. 36. OECD Publishing. https://doi.org/10.1787/220262483674
Pichert, J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. *Journal of Educational Psychology, 69*(4), 309–315.
Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106*(4), 643.
Pirolli, P., & Fu, W. T. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. In P. Brusilovsky, A. Corbett, & F. de Rosis (Eds.), *International Conference on User Modeling*. Berlin: Springer.
Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.
Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology, 53*(2), 145–161.
Rieh, S. Y., & Hilligoss, B. (2008). College students' credibility judgments in the information-seeking process. *Digital Media, Youth, and Credibility, 49*, 72.
Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 10*(3), e1488.
Román, P. E., L'Huillier, G., & Velásquez, J. D. (2010). Web usage mining. *Advanced Techniques in Web Intelligence-I*. https://doi.org/10.1007/978-3-642-14461-5_6
Rouet, J.-F. (2006). *The skills of document use: from text comprehension to web-based learning*. Mahwah: Lawrence Erlbaum Associates.
Rouet, J. F., Ayroles, J., Macedo-Rouet, M., & Potocki, A. (2021). *Children's acquisition of text search strategies: the role of task models and relevance processes* (pp. 185–212). Cham: Understanding and Improving Information Search Springer.
Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text*. Greenwich: Information Age.
Sabatini, J. P., O'Reilly, T., Wang, Z., & Dreier, K. (2018). Scenario-based assessment of multiple source use. In J. L. G. Braasch, I. Bråten, & M. T. McCrudden (Eds.), *Handbook of multiple source use* (pp. 447–465). New York: Routledge.
Salmerón, L., Strømsø, H. I., Kammerer, Y., Stadtler, M., & van den Broek, P. (2018). Comprehension processes in digital reading. In M. Barzillai, J. Thomson, S. Schroeder, & P. van den Broek (Eds.), *Learning to read in a digital world* (pp. 91–120). John Benjamins.
Scheerder, A., Van Deursen, A., & Van Dijk, J. (2017). Determinants of Internet skills, uses and outcomes. A systematic review of the second-and third-level digital divide. *Telematics and Informatics, 34*(8), 1607–1624.
Sparks, J. R., & Deane, P. (2015). Cognitively based assessment of research and inquiry skills: defining a key practice in the English language arts. *Educational Testing Service*. https://doi.org/10.1002/ets2.12082
Sparks, J. R., Katz, I. R., & Beile, P. M. (2016). Assessing digital information literacy in higher education: a review of existing frameworks and assessments with recommendations for next-generation assessment. *Educational Testing Service*. https://doi.org/10.1002/ets2.12118
Sparks, J. R., Appel, C., Gao, J., & Zhu, M. (2018). *NAEP SAIL Virtual World for Assessment of ELA Inquiry*. New York: Paper in coordinated symposium session presented at the annual meeting of the American Educational Research Association.
Sparks, J. R., van Rijn, P., & Deane, P. (2021). Assessing source evaluation skills of middle school students using learning progressions. *Educational Assessment*. https://doi.org/10.1080/10627197.2021.1966299
Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society Series A (statistics in Society), 179*(2), 481–511.
Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika, 85*(2), 378–397.

Tenison, C., & Arslan, B. (2020). Characterizing pause behaviors in a science inquiry task. In Proceedings of the 18th International Conference on Cognitive Modeling, Applied Cognitive Science Lab, Penn State, University Park, PA (pp. 283–298).
Tenison, C., & Anderson, J. R. (2016). Modeling the distinct phases of skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(5), 749.
Ulitzsch, E., He, Q., & Pohl, S. (2021). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986211010467
Van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*(8), 1081–1087.
Van Deursen, A. J., & Van Dijk, J. A. (2009). Using the Internet: skill related problems in users' online behavior. *Interacting with Computers, 21*(5–6), 393–402.
Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. Handbook of longitudinal research Design, measurement, and analysis, 373–385.
Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. A. (2008). Information-problem solving: a review of problems students encounter and instructional solutions. *Computers in Human Behavior, 24*, 623–648.
Wang, Z., Tang, X., Liu, J., & Ying, Z. (2020). Subtask analysis of process data through a predictive model. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12290
Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179.
Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal, 46*(4), 1060–1106.
Wilkinson, S. C., Reader, W., & Payne, S. J. (2012). Adaptive browsing: Sensitivity to time pressure and task difficulty. *International Journal of Human-Computer Studies, 70*(1), 14–25.
Wineburg, S. S. (1991). Historical problem solving: a study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83*, 73–87.
Xu, H., Fang, G., & Ying, Z. (2020). A latent topic model with Markov transition for process data. *British Journal of Mathematical and Statistical Psychology, 73*(3), 474–505.
Zhang, M., & Quintana, C. (2012). Scaffolding strategies for supporting middle school students' online inquiry processes. *Computers & Education, 58*(1), 181–196. https://doi.org/10.1016/j.compedu.2011.07.016

Publisher's Note