Hahnel, Carolin; Jung, Alexander J.; Goldhammer, Frank

Theory matters. An example of deriving process indicators from log data to assess decision-making processes in web search tasks.

European journal of psychological assessment 39 (2023) 4, S. 271-279, 10.1027/1015-5759/a000776

Special Issue: Process Data in Computer-Based Assessment: Opening the Black Box
Original Article

Theory Matters

An Example of Deriving Process Indicators From Log Data to Assess Decision-Making Processes in Web Search Tasks

Carolin Hahnel[1,2], Alexander J. Jung[3], and Frank Goldhammer[1,2]

[1]DIPF | Leibniz Institute for Research and Information in Education, Frankfurt a.M., Germany
[2]Centre for International Student Assessment (ZIB), Frankfurt a.M., Germany
[3]Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

Abstract. Following an extended perspective of evidence-centered design, this study provides a methodological exemplar of the theory-based construction of process indicators from log data. We investigated decision-making processes in web search as the target construct, assuming that individuals follow a heuristic search (focusing on search results vs. websites as a primary information source) and stopping rule (following a satisficing vs. sampling strategy). Drawing on these assumptions, we describe our reasoning for identifying the empirical evidence needed and selecting an assessment to obtain this evidence to derive process indicators that represent groups differentiated by search and stopping rule combinations. To evaluate our approach, we reanalyzed the processing behavior of 150 university students who were requested in four tasks to select a specific website from a list of five search results. We determined the process indicators per item and conducted multiple cluster analyses to investigate group recovery. For each item, we found three clusters, two of which matched our assumptions. Additionally, we explored the consistency of students' cluster membership across items and investigated their relationship with students' skills in evaluating online information. Based on the results, we discuss the tradeoff between construct breadth and process elaboration for deriving meaningful process indicators.

Keywords: process indicator construction, evidence identification, web search, log data, feature selection

OPEN MATERIALS

Process data, such as log data that record events of mouse clicks and keystrokes with timestamps, can enrich assessments in many ways (e.g., representing construct-related processes such as planning, Eichmann et al., 2019; ensuring data quality by detecting aberrant responses, van der Linden & Guo, 2008). Nevertheless, deriving meaningful process indicators from log data is challenging. It requires integrating available log events with characteristics of the respective task design and expectations about an individual's behavior in the task situation (Goldhammer et al., 2021). However, complex constructs tend to be underdefined in how exactly involved mental processes translate into observable behaviors (cf. sourcing in multiple document comprehension: Hahnel et al., 2019), or they are too versatile to be described in terms of a single indicator (cf. different patterns of exploration: Greiff et al., 2018).

Elaborate conceptual ideas about how a mental process manifests in a particular task situation are crucial for task design. They can motivate the derivation of process indicators tailored to represent attributes of the mental process at work. Uses for log data might also be discovered after the design phase, requiring the identification of new arguments that justify inferences about the mental process of interest from process data. This study seeks to provide a methodological exemplar of such a theory-based construction of process indicators from log data. For this purpose, we investigate the decision-making process of individuals in online information searches. Using log data from the computer-based test EVON (*Ev*aluation of *Onl*ine Information; Hahnel et al., 2020), we intend to derive a set of process indicators that maximize the differences between distinct decision-making approaches to selecting online information.

Reasoning From Evidence Provided by Log Data

Based on the idea that the nature of the construct of interest should guide the development or selection of tasks, scoring methods, and statistical models, evidence-centered design (ECD) suggests specifying the variables representing the construct (student model), the situation used to obtain

https://doi.org/10.1027/1015-5759/a000776

evidence about them (task model), and the rules of using the obtained evidence to draw inferences about characteristics of the construct (evidence model; Mislevy et al., 2003). Applying the ECD perspective to reason from the evidence provided by log data, Goldhammer and colleagues (2021) elaborate that a theory-driven construction of process indicators focuses on the attributes of a work process (student model). Accordingly, theoretical rationales about these attributes provide information about the kind of evidence needed for their identification (evidence model) and about the task design to make this evidence observable (task/activity model). By applying evidence identification rules ("scoring the work process"), process indicators can be derived, provided that the raw log events (e.g., a *LinkLogEntry* event at time *t*) can be interpreted as low-level features – that is meaningful process components (e.g., the *LinkLogEntry* event can be interpreted as visiting a specific page at time *t*). Several low-level features (e.g., multiple events indicating a page visit) can then be synthesized into a more complex process indicator (e.g., adding the event occurrences gives the number of page visits) which is assumed to indicate the presence (or absence) of the target attribute.

An assessment is ideally designed around a specific goal of log data use (e.g., Hahnel et al., 2019). However, situations exist where uses for log data are discovered post hoc. In this case, the validity argument that justifies inferences about the target attribute from a process indicator cannot rely on theoretical a priori arguments that drove the task design (Goldhammer et al., 2021). Instead, based on the theoretical rationales about the target attribute, new evidence identification rules need to be established that specify what kind of empirical evidence is required and by what degree of low-level feature aggregation it is obtained. Accordingly, the task/activity model of a selected task must be evaluated in terms of whether it provides sufficient interactivity to capture log data that allow the identification of the required low-level features. The present study illustrates this reasoning by investigating the decision-making process in web search as the work process, the use of certain heuristics shaping this process as the target attributes, and the EVON instrument as the situation-providing assessment to obtain the required empirical evidence to reason about the work process.

Decision-Making in Web Search

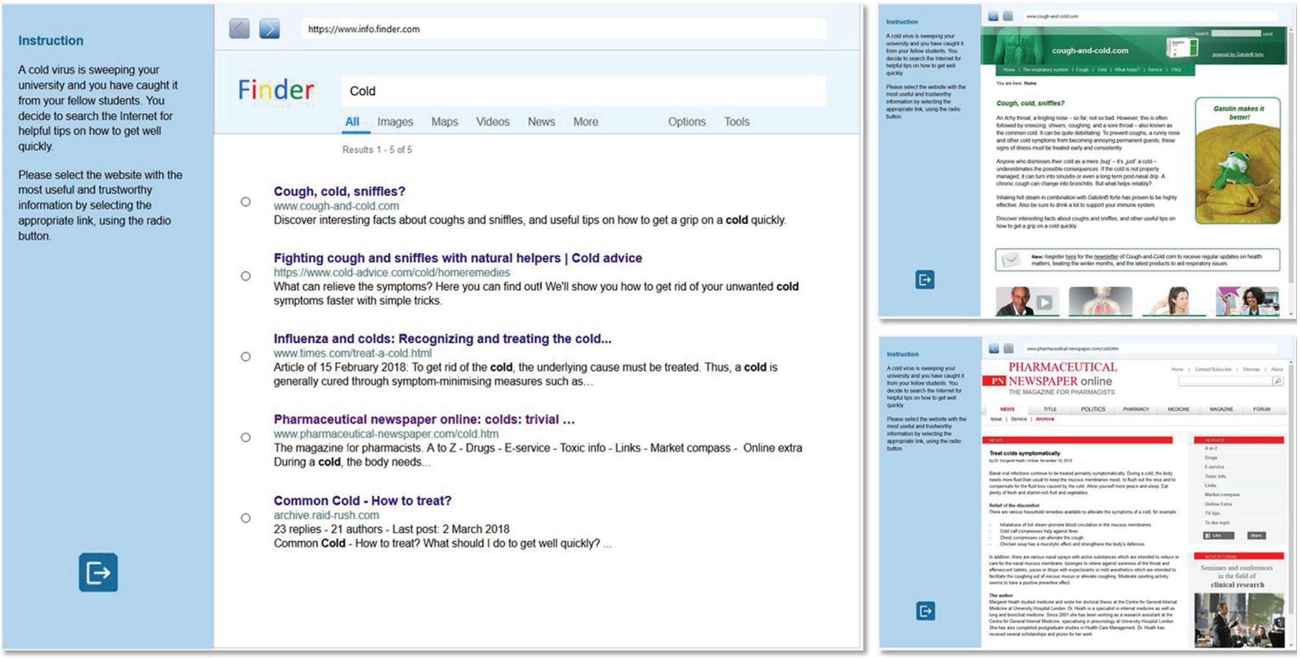
Web search is an information problem-solving process in which individuals must make the best possible decision under time constraints with limited cognitive resources (Brand-Gruwel et al., 2009; Pirolli, 2005). To cope with the abundance of available information efficiently, individuals will apply heuristics (Metzger et al., 2010). Heuristics are cognitive shortcuts meant to save mental effort. They are considered to share a set of rules for search and stopping as their common building blocks (e.g., Gigerenzer & Gaissmaier, 2011). We applied the concept of this "adaptive toolbox" to the context of web search and identified search and stopping rules individuals might apply when making web search decisions.

Search rules specify the direction in which a search extends. According to the idea of information foraging (Pirolli, 2005), individuals will engage in predicting the value of information for a task at hand using proximal cues in the immediate task environment. Consequently, individuals may focus predominantly on the information presented on a search result page (SERP) and refrain from an inspection of website content that did not pass their initial judgment ("SERP-focused" search rule). Empirical results suggest a second search rule focused on the evaluation of website content. Wirth and colleagues (2007) found that many individuals, especially those with above-average domain-specific knowledge, tended to visit a larger number of websites briefly before focusing on a specific website more thoroughly. Accordingly, these individuals might base their decision for a particular website less on the SERP information and more on skimming viable alternatives ("website-focused" search rule).

Stopping rules define sufficiency thresholds and express the moment at which individuals terminate their search efforts (Browne et al., 2007). Reader and Payne (2007) discuss satisficing and sampling as two strategies for deciding when to stop selecting texts for learning purposes. They describe satisficing as sifting through the available options until readers find the first option that exceeds their self-set threshold of the desired outcome, while sampling is characterized as going through and comparing the available options until the best option is found. While a sampling-based search will be more comprehensive, it also requires more cognitive effort and time. Based on these descriptions, we distinguish stopping rules that refer to the outcomes of having applied a satisficing ("early stop") or sampling strategy ("late stop").

Assuming that individuals follow a specific combination of the above-described search and stopping rules, we distinguish four mutually exclusive groups that represent our target attributes:

- A "SERP-focused/early stop" (SE) group that focuses predominantly on inspecting search results and ends their search process as soon as a sufficient match between task and information source has been found.
- A "SERP-focused/late stop" (SL) group that also uses the SERP as the primary search space, but ends the search after finding the supposedly best result.
- A "website-focused/early stop" (WE) group that focuses on the evaluation of website content and stops as soon as any matching website has been found.



Instruction

A cold virus is sweeping your university and you have caught it from your fellow students. You decide to search the Internet for helpful tips on how to get well quickly.

Please select the website with the most useful and trustworthy information by selecting the appropriate link, using the radio button.

https://www.info.finder.com

Finder Cold

All Images Maps Videos News More Options Tools

Results 1 - 5 of 5

○ **Cough, cold, sniffles?**
www.cough-and-cold.com
Discover interesting facts about coughs and sniffles, and useful tips on how to get a grip on a **cold** quickly.

○ **Fighting cough and sniffles with natural helpers | Cold advice**
https://www.cold-advice.com/cold/homeremedies
What can relieve the symptoms? Here you can find out! We'll show you how to get rid of your unwanted **cold** symptoms faster with simple tricks.

○ **Influenza and colds: Recognizing and treating the cold...**
www.times.com/treat-a-cold.html
Article of 15 February 2018: To get rid of the **cold**, the underlying cause must be treated. Thus, a **cold** is generally cured through symptom-minimising measures such as...

○ **Pharmaceutical newspaper online: colds: trivial ...**
www.pharmaceutical-newspaper.com/cold.htm
The magazine for pharmacists. A to Z - Drugs - E-service - Toxic info - Links - Market compass - Online extra During a **cold**, the body needs...

○ **Common Cold - How to treat?**
archive.raid-rush.com
23 replies - 21 authors - Last post: 2 March 2018
Common **Cold** - How to treat? What should I do to get well quickly? ...

Instruction

A cold virus is sweeping your university and you have caught it from your fellow students. You decide to search the Internet for helpful tips on how to get well quickly.

Please select the website with the most useful and trustworthy information by selecting the appropriate link, using the radio button.

www.cough-and-cold.com

search: send

cough-and-cold.com

powered by Gatolin® forte

Home | The respiratory system | Cough | Cold | What helps? | Service | FAQ

You are here: **Home**

Cough, cold, sniffles?

An itchy throat, a tingling nose – so far, not so bad. However, this is often followed by sneezing, shivers, coughing, and a sore throat – also known as the common cold. It can be quite debilitating. To prevent coughs, a runny nose and other cold symptoms from becoming annoying permanent guests, these signs of illness must be treated early and consistently.

Anyone who dismisses their cold as a mere ‚bug' – it's ‚just' a cold – underestimates the possible consequences. If the cold is not properly managed, it can turn into sinusitis or even a long term post-nasal drip. A chronic cough can change into bronchitis. But what helps reliably?

Inhaling hot steam in combination with Gatolin® forte has proven to be highly effective. Also be sure to drink a lot to support your immune system.

Discover interesting facts about coughs and sniffles, and other useful tips on how to get a grip on a cold quickly.

Gatolin makes it better!

New: Register here for the newsletter of Cough-and-Cold.com to receive regular updates on health matters, beating the winter months, and the latest products to aid respiratory issues.

Instruction

A cold virus is sweeping your university and you have caught it from your fellow students. You decide to search the Internet for helpful tips on how to get well quickly.

Please select the website with the most useful and trustworthy information by selecting the appropriate link, using the radio button.

www.pharmaceutical-newspaper.com/cold.htm

PN PHARMACEUTICAL NEWSPAPER online
THE MAGAZINE FOR PHARMACISTS

Home | Contact/Subscribe | Sitemap | About

NEWS TITLE POLITICS PHARMACY MEDICINE MAGAZINE FORUM

Issue | Service | Archive

NEWS

Treat colds symptomatically
by Dr. Margaret Heath | Article: November 10, 2018

Banal viral infections continue to be treated primarily symptomatically. During a cold, the body needs more fluid than usual to keep the mucous membranes moist, to flush out the virus and to compensate for the fluid loss caused by the cold. Allow yourself more peace and sleep. Eat plenty of fresh and vitamin-rich fruit and vegetables.

Relief of the discomfort
There are various household remedies available to alleviate the symptoms of a cold, for example:

- Inhalations of hot steam promote blood circulation in the mucous membranes.
- Cold calf compresses help against fever.
- Chest compresses can alleviate the cough.
- Chicken soup has a mucolytic effect and strengthens the body's defences.

In addition, there are various nasal sprays with active substances which are intended to reduce or care for the nasal mucous membrane, lozenges to relieve against soreness of the throat and effervescent tablets, juices or drops with expectorants or mild anesthetics which are intended to facilitate the coughing out of viscous mucus or alleviate coughing. Moderate sporting activity seems to have a positive preventive effect.

The author
Margaret Heath studied medicine and wrote her doctoral thesis at the Centre for General Internal Medicine at University Hospital London. Dr. Heath is a specialist in internal medicine as well as lung and bronchial medicine. Since 2001 she has been working as a research assistant at the Centre for General Internal Medicine, specialising in pneumology at University Hospital London. She has also completed postgraduate studies in Health Care Management. Dr. Heath has received several scholarships and prizes for her work.

SERVICE
A to Z
Drugs
E-service
Toxic info
Links
Market compass
Online Extra
TV tips
To the topic

ADVERTORIAL
Seminars and conferences in the field of **clinical research**

Figure 1. Example of an EVON item (from Hahnel et al., 2020).

- A “website-focused/late stop” (WL) group that also concentrates on website content but searches until the best match to the information problem has been identified.

Assessing Decision-Making Processes With the EVON

An assessment that makes evidence of our target attributes observable needs to provide individuals with situations where they are invited to make a decision between multiple available alternatives at their own pace with the option to explore different search spaces autonomously. Such a task/activity model is realized for the items of EVON (Hahnel et al., 2020). EVON was originally constructed to capture students' skill in using semantic cues and structural, message-based, and sponsor-based credibility features to evaluate information online. The items request students to identify the website with the most useful and trustworthy information, providing them with an information problem, a SERP, and websites (Figure 1). Low-level features that can be extracted from log data include actions and states that mark the beginning and end of item processing (item start and end events), accessing and staying on a specific page (SERP and websites), and submitting a response (radio button selection).

So, how do we link the target attributes to observable behavior within the EVON situations? We apply the theoretical assumptions that motivated the distinction of the four target attributes to the specific application context of the activity space of the task environment. That means, based on our specification of the distinguished search and stopping rules, we portray how individuals who apply specific rules should behave in the EVON items and what kind of traces they are likely to leave behind that can be extracted from the log data.

Concerning the search rules, individuals who apply a SERP-focused rule should spend the majority of their time on the search results and be less likely to inspect websites. In contrast, individuals who apply a website-focused rule will spend most of their time on one or more websites. Accordingly, process indicators reflecting individuals' time investment on a particular page type (indicator 1: *SERP-to-website time ratio*) and their access to different pages (indicator 2: *number of websites visited*) should be suitable to differentiate between different search rules.

Concerning the stopping rules, individuals who apply an early-stop rule should be less likely to spend time working on an information problem than those who look for the best result. They should also be less inclined to revisit already accessed websites, as these websites were considered insufficient if the search continues. Accordingly, individuals with an early-stop rule should be likely to select their last inspected option as their final response. In contrast, individuals with a late-stop rule should take time to review all available options and might even compare them multiple times. Accordingly, process indicators reflecting individuals' time investment for task processing (indicator 3: *processing time*) and their behavior of deciding on the last inspected material without revisiting websites (indicator 4: *satisficing*) should be suitable to differentiate between different stopping rules.

Finally, we can represent our target attributes (SE, SL, WE, WL) as a synthesis of four process indicators (SERP-to-website time ratio, number of websites visited, processing time, and satisficing). This synthesis may be accomplished by setting thresholds for the process indicators that evaluate whether the displayed behavior indicates the presence or absence of a particular attribute (e.g., Hahnel et al., 2019). However, due to a lack of sufficient information about precise thresholds, we will aggregate them using cluster analysis. This approach also allows us to examine whether the theoretically expected groups can be recovered empirically and to evaluate our approach of deriving theory-informed process indicators from the large pool of possible process indicators.

The Current Study

With the aim to provide a methodological exemplar of the theory-based construction of process indicators from log data, we introduced a theoretical rationale about the heuristic use of search and stopping rules in web search and derived, based on our assumptions, four process indicators that should represent groups of different combinations of those rules. We used log data from the EVON assessment to create the process indicators and performed a cluster analysis to investigate whether the assumed groups can be empirically recovered. To interpret the empirical clusters, we examined the characteristics of the found clusters and compared them with theory-based expectations about the relational orders between the mean values of the process indicators (Table 1). Additionally, we explored the consistency of cluster membership across different EVON items and associated cluster membership with the skill of evaluating online information.

Methods

Design and Sample

We reanalyzed the result and process data of Hahnel and colleagues (2020) for four EVON items that presented students with five clickable links and websites (the other items showed only three links, reducing variation in the process indicators). As part of a half-an-hour online assessment, the items were administered in a fixed order (positions 2, 3, 5, 7). Notably, each item was of a different item type that varied how immediately the target link/website (i.e., the correct solution) was identifiable as the optimal choice in terms of relevance and credibility compared to competing non-targets. The first item (position 2) provided clear credibility cues on the target website that were not obvious from the SERP information (*Cold*, type 3; Figure 1). The non-target websites of the second item displayed flaws not indicated by the SERP (*Paper*, type 4). For the third item, the SERP and website information matched completely (*Diving*, type 2). These three items had in common that all their search results were semantically related to a task at hand. That was not the case for the last item (*Email*, type 1), where the non-targets were only related to the task on a surface level. Hahnel et al. (2020) showed that the item type predicted item difficulty, with type 1 being the easiest and type 3 the most difficult item type. Apart from this, the EVON items are relatively uniform (e.g., they are of the same structure and navigation complexity). Given that the EVON addresses university students, the readability of provided texts is of low to moderate complexity (Table S1).

After excluding two cases with incomplete log data, we analyzed the data of 150 students. The sample was on average 23.2 years old (*SD* = 3.42), comprised 66.4% female students, and was enrolled in different programs (bachelor = 54.05%, master = 14.19%, German Staatsexamen [a degree earned after long-cycle programs, such as law, medicine, teaching] = 31.76%).

Measures

The four process indicators were derived per item using the R package *LogFSM* (Kroehne, 2019). Table S2 presents their means, standard deviations, and intercorrelations. They were operationalized as follows:

- The *SERP–website time ratio* is the time spent on the SERP, over the total time on websites (plus a constant of 1 ms). The ratio was log-transformed. Accordingly, positive values indicate that students spent most of their time on the search results; negative values indicate higher time investment for websites.
- The *number of websites visited* is a count indicator of different websites accessed while students processed an item (range: 0–5 websites).
- *Processing time* is the time interval from the item start event to the event of students clicking the next-item button. The indicator was log-transformed for the cluster analysis.
- *Satisficing* was represented as a dichotomous indicator that marked students as showing satisficing behavior if their behavior met the following criteria: They selected the website they had visited last as their response (independent of the response's correctness); did not visit websites multiple times; and did not visit any websites after their response.

Table 1. Evidence identification rules that assign the attribute labels according to relational orders of the process indicators' mean values

Group	SERP–websites time ratio	Number of websites visited	Processing time	Satisficing
SE	$M_{SE} > M_{SL} > 0$	$M_{SE} < M_{SL/WE/WL}$	$M_{SE} < M_{SL/WE/WL}$	$M_{SE} < M_{WE}$
SL	$M_{SL} > 0$	$M_{SE} < M_{SL} < M_{WL}$	$M_{SE} < M_{SL} < M_{WL}$	$M_{SL} < M_{WE}$
WE	$M_{WE} < 0$	$M_{SE} < M_{WE} < M_{WL}$	$M_{SE} < M_{WE} < M_{WL}$	$M_{WE} > M_{SE/SL/WL}$
WL	$M_{WL} < M_{WE} < 0$	$M_{WL} > M_{SE/SL/WE}$	$M_{WL} > M_{SE/SL/WE}$	$M_{WL} < M_{WE}$

Note. SE = SERP-focused/early stop; SL = SERP-focused/late stop; WE = website-focused/early stop; WL = website-focused/late stop.

For further exploratory analyses of the relationship between the empirical groups of web search decision-making and students' skill to evaluate online information, we used the EVON score estimated by Hahnel and colleagues (2020; EAP reliability = .62, p. 10 in their paper).

Data Analysis

For synthesizing the process indicators, we performed several cluster analyses per item (R package *cluster*; Maechler et al., 2021). Note that we deviate from the typical application of cluster analysis since we do not primarily focus on detecting unknown patterns but rather demonstrate the results of a particular feature selection strategy. We first determined a dissimilarity matrix using Gower's distance that handles mixed data types. It computes a similarity value between individuals by variable type (continuous vs. categorical) and averages the partial dissimilarities across individuals (Everitt et al., 2001). Afterward, we used a partitioning around medoids (PAM) algorithm. PAM clustering determines a set of objects within a cluster (medoids) and calculates other cases' dissimilarities. The process is iterative, which means any case that reduces the average dissimilarity is used as a new cluster medoid until new medoids cannot be found anymore. The final medoids represent the most centrally located objects in the clusters.

PAM clustering requires predefining the number of clusters. Although we expected 4 groups, we ran the algorithm multiple times for each items to produce partitions into 2–10 groups to evaluate alternative clustering. We determined the average silhouette width as a coherence measure for each run. Silhouette widths close to 1 indicate a high within-group homogeneity and a low overlap with other groups. A justifiable cluster structure is characterized by an average silhouette width larger than 0.5 (Everitt et al., 2001).

Finally, we investigated students' cluster membership across the items based on the identified clusters. With four clusters within four items, we expected 256 unique combinations to recode according to how many behaviors students showed across the items (e.g., various combinations of SE and SL behaviors would be summarized as "SE/SL behaviors"). Using regression analysis, we used this new membership indicator to predict students' performance in evaluating online information.

Results

Clustering

We decided on the three-cluster solutions, as their average silhouette widths were the highest (average values between .70 and .79; Figure S1). Table 2 shows the descriptive statistics of the process indicators and the cluster size per cluster (also Table S3). The identified clusters were highly similar across the items. Therefore, we assigned the same label to similar clusters.

In line with the evidence identification rules for assigning the attribute labels (Table 1), two cluster sets matched the attributes of SE and WL. SE students spent most of their time on the SERP (high positive SERP–website time ratios), visited hardly any websites, and completed the items rather quickly (on average, about 22 s). In contrast, WL students spent most of their time on the websites (negative SERP–website time ratios), visited almost all of them (4–5 pages), and showed the most extended processing times.

The remaining cluster set revealed characteristics of both SL and WE. These students tended to spend more time on the SERP than on the websites (low but positive SERP–website time ratios), favoring the SL interpretation, but predominantly displayed satisficing behavior compared to the students of the other clusters, favoring the WE interpretation. In line with both interpretations, these students included website information in their decision-making process (on average, 2–4 websites visited), and invested a moderate amount of time in item processing compared to the SE and WL clusters. Accordingly, we assigned a new label to this mixed cluster (MX).

Exploratory Analyses

First, we inspected the item success rates and the average EVON score for the identified clusters (Table 2). The item success rates for the SE clusters did not considerably deviate from the chance level (at .20) for two items but were lower for the item *Paper* and higher for the item *Email*. The SE attribute was descriptively associated with a below-average EVON score. The MX clusters were fairly successful in item processing, while their behavior seemed less suitable for solving the item *Paper*. Their average EVON score, though, varied considerably between items (averages between 0.08

Table 2. Results from partitioning around medoids clustering (means and standard deviations of the process indicators) and average item and test performance per cluster and item

Item	Chosen label	*n*	Log time ratio SERP–websites	Number of websites visited	Processing time (seconds)	Satisficing*	Item score*	EVON score
2: Cold	SE	46	9.99 (0.72)	0.00 (0.00)	27.47 (17.60)	0.00	0.24	−0.76 (0.63)
	MX	35	0.69 (0.76)	2.14 (1.35)	61.15 (37.78)	1.00	0.70	0.17 (0.61)
	WL	69	−0.14 (0.65)	4.25 (1.13)	109.47 (45.88)	0.00	0.58	0.42 (0.72)
3: Paper	SE	52	9.82 (0.61)	0.00 (0.00)	22.27 (14.09)	0.00	0.13	−0.73 (0.64)
	MX	39	0.30 (0.62)	2.56 (1.25)	58.91 (21.16)	0.74	0.63	0.08 (0.71)
	WL	59	−0.51 (0.48)	4.71 (0.53)	104.14 (40.03)	0.02	0.76	0.58 (0.56)
5: Diving	SE	53	9.28 (2.03)	0.06 (0.23)	21.56 (11.99)	0.00	0.23	−0.77 (0.59)
	MX	34	0.19 (0.72)	2.82 (1.47)	56.10 (29.32)	1.00	0.73	0.32 (0.71)
	WL	63	−0.37 (0.58)	4.59 (0.75)	88.90 (39.25)	0.00	0.73	0.47 (0.59)
7: Email	SE	63	9.51 (1.26)	0.02 (0.13)	18.26 (10.48)	0.00	0.45	−0.63 (0.69)
	MX	39	0.42 (0.73)	2.03 (1.18)	52.75 (24.15)	1.00	0.82	0.30 (0.62)
	WL	48	−0.20 (0.69)	4.10 (1.12)	87.07 (40.04)	0.00	0.89	0.58 (0.60)

Note. *The relative frequency of category “1” (dichotomous variable). Performance measures were not included in the clustering. SE = SERP-focused/early stop; MX = mixed behaviors; WL = website-focused/late stop.

and 0.32) indicating different group compositions (i.e., students seemed to exhibit MX behavior under specific circumstances rather than consistently across items). The WL clusters showed high item success rates (except for the item *Cold*) and the highest average EVON scores.

Next, we inspected students’ cluster membership across items. Figure 2 shows that most students who exhibited SE behavior in the first item also showed this behavior in other items. Transitions to other clusters occurred but were less likely. For WL behavior, we also found that a considerable proportion of students stuck exclusively to WL behavior, but it was lower than for SE behavior. There were comparatively few students who adhered exclusively to MX behaviors. Instead, students exhibiting MX behavior at some point tended to demonstrate a mix of different behaviors across items. Notably, the least transitions occurred between the SE and WL behaviors, the two most distinct behaviors given their assumed search and stopping rules.

Finally, we recorded the behavior sequences into a categorical variable distinguishing between which and how many different behaviors were exhibited during item processing (reference category “SE behavior only”; Table S4). The variable was used to predict the EVON score. The results in Table 3 show that 49.7% of the variability in the EVON score was explained. Students who displayed MX or WL behaviors, or a combination of both, demonstrated higher skill in evaluating online information than students who did not exhibit these behaviors.

Discussion

Our study demonstrates how theoretical expectations about a mental process, such as decision-making in web search, contribute to drawing inferences from process data produced by assessments like the EVON, which assesses students’ skills in evaluating online information. The results illustrate the strengths and challenges of this approach. Following Goldhammer and colleagues (2021), we were able to derive new process indicators (SERP-to-website time ratio, satisficing) and improve the informational value of other more generic ones (number of pages visited, processing time). Developing evidence identification rules was particularly valuable, as it allowed us to investigate log data for uses that were not considered during the design of the EVON items, demonstrating the potential of reusing existing computer-based test instruments. Although we did not find all anticipated groups, we were able to assign meaningful interpretations to the emerged groups with ease, guided by our evidence identification rules and theoretical assumptions about search and stopping rules (see Gigerenzer & Gaissmaier, 2011). Having found the expected SE and WL attributes supports the plausibility of the validity argument for the derived process indicators, despite a lack of evidence for separable SL and WE attributes (i.e., finding of the MX group). Still, the appropriateness of our preferred interpretation needs to be carefully evaluated against alternative interpretations and limitations of this study.

Alternative interpretations of the SE and WL cluster could result from construct-irrelevant influences. For example, some SE students showed very low processing times, suggesting that their observed behavior might be due to rapid guessing or test disengagement. However, this interpretation is opposed by the comparatively high solution rate in the item *Email* (the target link stood out in relevance) and the solution rate below the chance level in the item *Paper* (the non-target websites displayed deficiencies but not their links). Another example is that many of the WL students visited all available websites. Although this

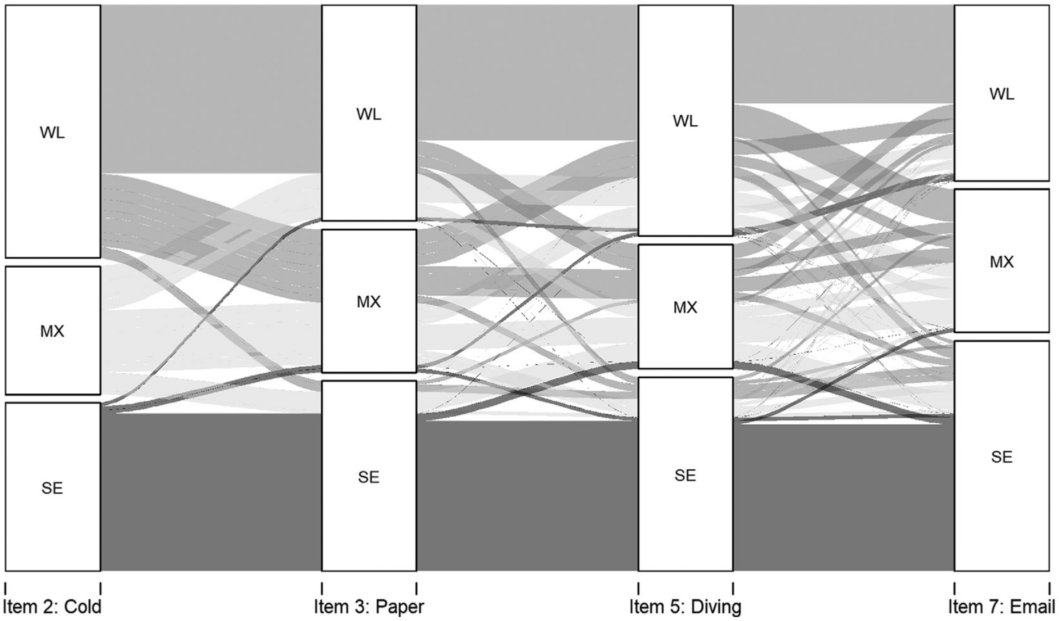


Figure 2. Flow chart of students' cluster membership across items. Box sizes are proportional to cluster sizes; colors represent students' membership in the first item; line thickness indicates the sample size of students in boxes connected by the line.

Table 3. Results of the exploratory regression analysis to predict the EVON score (reference category "SE behavior only")

Predictor	*b*	*p*	*b* 95% CI [LL, UL]	Fit
(Intercept)	−0.84***	< .001	[−1.03, −0.65]	
MX behavior only	1.30***	< .001	[0.81, 1.79]	
WL behavior only	1.54***	< .001	[1.24, 1.84]	
SE/MX behaviors	0.51**	.010	[0.13, 0.90]	
SE/WL behaviors	0.34	.243	[−0.23, 0.91]	
MX/WL behaviors	1.23***	< .001	[0.97, 1.49]	
Mix of all behaviors	0.84***	< .001	[0.44, 1.23]	
				R^2 = .497** 95% CI [.36, .57]

Note. *b* = unstandardized regression weights; LL = lower limits of a confidence interval; UL = upper limits of a confidence interval; SE = SERP-focused/early stop; MX = mixed behaviors; WL = website-focused/late stop. ***p* < .01; ****p* < .001.

behavior is consistent with the definition of sampling (Reader & Payne, 2007), it also marks an effective test-taking strategy in items asking for an optimal choice. The proportion of WL students in the item *Email* could be taken as evidence for this interpretation. However, this cluster was descriptively smaller than the WL clusters in the other items. Accordingly, serious attempts to challenge the SE/WL interpretations of the clusters could be to provide students with experimentally varied additional instructions that request them to focus on a specific aspect (e.g., decide as fast as possible vs. consider all available information).

The question remains of why we did not find evidence for the assumed SL and WE attributes. Instead, we repeatedly observed the MX cluster, which appears to combine characteristics of both attributes. A reasonable explanation could be that the activity space of the items was too limited to separate the assumed attributes empirically. The EVON items present only five link-website pairs without further website navigation options. Therefore, a suitable validation approach could be to attempt to identify the clusters when students work on web search tasks under more natural conditions (i.e., in a large enclosed or open information space). Alternatively, the MX group might also point out that our theoretical assumptions oversimplify possible search and stopping rules and combinations of those. Currently, our theoretical assumptions speak for mutually exclusive attributes, not covering phenomena such as switching heuristics or following multiple searches and stopping rules. It is reasonable to assume that advancing the differentiation of our theoretical assumptions about different rules could lead to more differentiated attributes and, consequently, process indicators. Finally, the absence of separate SL and WE clusters might also be due to the small sample of university students who are comparably homogenous regarding their cognitive abilities or the high correlations between the continuous process indicators (Table S2). Especially the latter marks a challenging issue as it can influence cluster formation. It might be overcome by providing a larger activity space, eventually affecting the variation of the process indicators.

The presence of the MX cluster also raises the question of whether uninterpretable behavioral patterns could exist. Such patterns could reflect construct-irrelevant behavior, such as patterns with long times of inactivity (e.g., responding after several minutes without any website visit) or short times of activity (e.g., visiting all websites in a short period). In fact, observing new clusters would be highly informative, as they can indicate flawed evidence identification rules or the need to adapt and extend the underlying theoretical assumptions. Additionally, the cluster-analytical approach could contribute to masking the occurrence of certain groups. Although we see strength in synthesizing multiple process indicators to indicate the presence or absence of a target attribute, unpredicted behavior might get grouped with somewhat similar behavior if it is inconspicuous enough in the data. If such cases are not revealed, the decision-making process of these students will be misclassified. With a larger sample, information on the individual fit to a class can be obtained by using a model-based approach such as latent class analysis (see Greiff et al., 2018). Provided that it is reasonable that the theoretical groups exist, it is also possible to use fuzzy clustering, a technique that specifies the strength of a case's membership in each assumed cluster without an underlying probabilistic model. For our study, we can at least note that the substantial average silhouette widths indicate well-separated and coherent clusters.

Finally, the exploratory analyses indicated individual and task-specific differences, providing interesting directions for future validation efforts. The results show that the clusters could be consistently identified for all four items, albeit in varying proportions across items. This variability hints at students' capability to switch their behavior between items to adapt to the current task requirements. Accordingly, some pattern combinations across items might be more likely than others, assuming that students with high web search abilities are able to choose context-appropriate search and stopping rules. This adaptivity is particularly evident in students who exhibit WL and MX behaviors across items. They differed little in their EVON scores from students who only engaged in WL behaviors, which are more demanding in their completion (i.e., higher time and effort investment due to exposure to more information). However, the direction and implications of these results must be considered with utmost care since our analyses only included one item per item type. Accordingly, any effect of cluster membership is potentially confounded with individual factors (e.g., skill, interest, motivation) or item and test characteristics (e.g., item position within the test, the position of the target link within an item) and requires attention in future research.

Despite its limitations, our study holds value in demonstrating how theoretical assumptions can guide the selection and construction of meaningful process indicators. It also illustrates a tradeoff between a construct's breadth and its theoretical elaboration on processes. The broader the construct of interest, the more difficult it is to break down its complexity in terms of observable behavior on a micro-level. In contrast, overly specified assumptions might not be generalizable to observations in other contexts anymore (see also the discussion of generic vs. task-specific process indicators in Goldhammer et al., 2021). The theoretical assumptions in our study were specific enough to guide the identification of evidence for the assumed cognitive processes to derive meaningful process indicators. However, this came at the cost of simplifying possible search and stopping rules into dichotomous heuristics. In summary, it is challenging to formulate specific theoretical assumptions about processes that cover a construct appropriately and motivate process indicators with a clear, preferably unambiguous interpretation. Nevertheless, our study suggests that even the attempt contributes to clarifying what evidence from log data is needed and how it can be identified as process indicators to represent the construct of interest.

References

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education, 53*(4), 1207–1217. https://doi.org/10.1016/j.compedu.2009.06.004

Browne, G. J., Pitts, M. G., & Wetherbe, J. C. (2007). Cognitive stopping rules for terminating information search in online tasks. *MIS Quarterly, 31*(1), 89–104. https://doi.org/10.2307/25148782

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education, 128*, 1–12. https://doi.org/10.1016/j.compedu.2018.08.004

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). Oxford University Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology, 62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

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. https://doi.org/10.1186/s40536-021-00113-5

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education, 126*, 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

Hahnel, C., Eichmann, B., & Goldhammer, F. (2020). Evaluation of online information in university students: Development and scaling of the screening instrument EVON. *Frontiers in Psychology, 11*, 1–16. https://doi.org/10.3389/fpsyg.2020.562128

Hahnel, C., Jung, A. J., & Goldhammer, F. (2023, April 22). *Data and supplementary materials for "Theory matters: An example of deriving process indicators from log data to assess decision-making processes in web search tasks"*. https://osf.io/5vpmn

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*, 524–537. https://doi.org/10.1111/bjep.12278

Kroehne, U. (2019). *LogFSM: Analyzing log data from educational assessments using finite state machines (LogFSM)*. http://www.logfsm.com

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). *cluster: Cluster analysis basics and extensions*. https://CRAN.R-project.org/package=cluster

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication, 60*(3), 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 1*, 1–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science, 29*(3), 343–373. https://doi.org/10.1207/s15516709cog0000_20

Reader, W. R., & Payne, S. J. (2007). Allocating time across multiple texts: Sampling and satisficing. *Human-Computer Interaction, 22*(3), 263–298. https://doi.org/10.1080/07370020701493376

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

Wirth, W., Böcking, T., Karnowski, V., & von Pape, T. (2007). Heuristic and systematic use of search engines. *Journal of Computer-Mediated Communication, 12*(3), 778–800. https://doi.org/10.1111/j.1083-6101.2007.00350.x

History
Received March 3, 2022
Revision received March 7, 2023
Accepted April 12, 2023
Published online August 9, 2023
EJPA Section / Category Methodological topics in assessment

Open Science

We report how we determined our sample size, all data exclusions, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact *p* values, effect sizes, and 95% confidence or credible intervals.

Open Data: The information needed to reproduce all of the reported results are not openly accessible. The data is available on request from the authors of Hahnel et al. (2020).

Open Materials: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology. The materials and the supplementary materials (Table S1–S4 and Figure S1) are available under https://osf.io/5vpmn/ (Hahnel et al., 2023).

Preregistration of Studies and Analysis Plans: This study was not preregistered.

Funding

This work was supported by the Centre of International Student Assessment (ZIB). Open access publication enabled by DIPF | Leibniz Institute for Research and Information in Education.

ORCID
Carolin Hahnel
https://orcid.org/0000-0003-2394-3944
Alexander J. Jung
https://orcid.org/0000-0003-3699-9066
Frank Goldhammer
https://orcid.org/0000-0003-0289-9534

Carolin Hahnel
DIPF | Leibniz Institute for Research and Information in Education
Rostocker Strasse 6
60323 Frankfurt
Germany
hahnel@dipf.de