Evaluating Knowledge Gain in Search Environments: An Exploratory Study of Learning Measurement

Marcelo Tibau[1], **Rafael Tavares da Silva**[1],
Sean Wolfgand Matsui Siqueira[1], **Bernardo Pereira Nunes**[2]

[1] Department of Applied Informatics – Federal University of the State of Rio de Janeiro Av. Pasteur, 458, Urca – 22.290-240 – Rio de Janeiro – RJ – Brazil

[2]School of Computing – Australian National University
108 North Rd, Acton ACT 2601 – Canberra – Australia

{marcelo.tibau,sean}@uniriotec.br,

rafaeltavares.silva@edu.unirio.br, bernardo.nunes@anu.edu.au

Abstract. ***Research Context:*** *Searching as Learning (SaL) frames web search as a process where users construct and refine knowledge. However, measuring knowledge gain in natural search environments remains a methodological challenge.* ***Scientific and/or Practical Problem:*** *Traditional behavioral proxies (e.g., dwell time, clicks) scale well but fail to capture conceptual change, while pre/post-tests provide richer insights but are intrusive. This gap limits the development of search systems that can evaluate and promote learning.* ***Proposed Solution and/or Analysis:*** *This study advances a computational measure based on entropy reduction and semantic similarity, and novelly operationalizes it through a browser plug-in that enables real-time measurement in natural search environments, extending prior formalizations and prototype-based validations of the DKG metric.* ***Related IS Theory:*** *The study draws on Shannon's Information Theory and Information Processing Theory in IS to conceptualize knowledge gain as uncertainty reduction supported by socio-technical processes.* ***Research Method:*** *An experiment combined three structured search tasks, pre/post-tests, and Concurrent Think-Aloud protocols. Quantitative measures (Transfer of Learning scores, also known as ToL, and values from the proposed metric) were triangulated with qualitative coding using OISS and ESKiP frameworks.* ***Summary of Results:*** *Statistical analysis showed a moderate positive correlation between ToL and the proposed metric ($r = 0.62$, $p < 0.01$). Bland–Altman analysis revealed systematic differences in scale, with ToL showing higher values, yet relative patterns were consistent. Transcripts emphasized how strategies such as query specialization, evaluation of sources, and persistence in reformulation aligned with higher values.* ***Contributions and Impact to IS area:*** *The study contributes a validated computational metric and artifacts for measuring knowledge gain in real search environments. It reinforces the sociotechnical view of IS by linking human strategies, processes, and technological advantages, and points to adaptive search systems that could measure and promote learning.*

1. Introduction

The Web has become one of the main environments for knowledge acquisition; beyond locating information, users frequently construct, refine, and integrate knowledge while interacting with search systems. This perspective, known as Searching as Learning (SaL), emphasizes the cognitive and metacognitive dimensions of search activities [Marchionini 2006] [Vakkari 2016] [Tibau et al. 2022] and has gained increasing attention in the fields of Information Retrieval and Human–Computer Interaction because it positions search systems as mediators of learning processes.

Despite this relevance, evaluating how much users actually learn during a search session remains a persistent challenge. Standard behavioral measures, such as dwell time, the number of clicks, or query reformulations [Xu et al. 2020] [Gritz et al. 2021] [Tibau et al. 2019], offer useful signals but do not adequately capture the conceptual changes that characterize knowledge acquisition. More structured methods, such as pre- and post-tests or open-ended assessments [Roy et al. 2020] [Câmara et al. 2021], are informative but intrusive, limiting their applicability in naturalistic settings [Roy et al. 2020]. This methodological gap has motivated the development of different procedures capable of automatically and reliably estimating the knowledge gain.

In this context, the Degree of Knowledge Gain (DKG) metric was proposed to quantify knowledge acquisition in search environments [Tibau 2024]. The DKG builds on Shannon's Entropy to represent reductions in uncertainty across search sessions, complemented by semantic similarity measures between queries and clicked documents. Previous studies have formalized the metric and provided empirical validation with diverse populations, including students and professionals, emphasizing its interpretability and potential as an automated indicator of learning [Tibau et al. 2022, Tibau et al. 2023, Tibau 2024].

Building on this foundation, the present study advances the operationalization of the DKG metric through the development of two technological artifacts. The first was a prototype search engine embedding the metric directly into its retrieval logic, serving as a proof-of-concept but limited by issues of result quality and scalability. To tackle these obstacles, a second artifact was developed: a browser plug-in capable of computing the DKG in real time while users employ their preferred search engines.

To assess the applicability of the plug-in, an experiment was conducted that combined three structured search tasks, pre- and post-tests, and the Concurrent Think-Aloud protocol [Kelley et al. 2015]. The tasks addressed topics from introductory science concepts to applied domains (e.g., prosthetics), allowing analysis of the DKG's behavior across varied levels of complexity. The novelty lies in the operationalization of DKG while preserving users' natural search behavior, combined with a mixed-methods validation that explicitly connects automatic measurements to learning strategies and test-based outcomes.

This research is situated at the intersection of people, processes and technologies, exploring how distinct persons take part in knowledge construction (people), how search and learning strategies unfold during tasks (processes), and how computational metrics and tools can support this activity (technologies). The application domain of this study is online learning and science education, where search systems act as mediators

of conceptual understanding and inquiry-based exploration. In doing so, it responds to the GranDSI-BR research agenda for Information Systems (IS) in Brazil, particularly the challenges of designing IS for the open world, where mobility, transparency, knowledge sharing, and diversity must be supported [Boscarioli et al. 2017], and the sociotechnical view of IS, which emphasizes the integration of human and technological dimensions in solving real-world problems [Boscarioli et al. 2017], contributing as a methodological advance for evaluating learning in digital contexts. The proposed DKG plug-in directly addresses the need for empirical mechanisms to evaluate information systems in real operational environments, as emphasized in the discussions on interoperability, evaluation of IS impacts, and sociotechnical integration (Chapters 1, 9, and 11), which aligns with the Brazilian IS community's long-term priorities.

The work also advances three distinct contributions: technical, methodological, and empirical. The introduction of the browser-based plug-in that computes the DKG metric in real time during users' natural search sessions is presented as the technical contribution; the proposition and empirical validation of a mixed-methods framework for learning measurement that triangulates an automatic entropy-based metric (DKG) with pre/post knowledge tests and qualitative coding of search strategies is presented as the methodological contribution; and empirical evidence linking search strategies, such as query specialization, evaluation, and persistence, to higher levels of measured knowledge gain is presented as the empirical contribution.

This paper begins by situating the metric within the Information Systems theories (Section 2) and SaL literature (Section 3), then details its theoretical foundations and formal definition (Section 4). Next, it presents the technological artifacts developed to operationalize the metric (Section 5) and the experimental methodology adopted (Section 6). The results are reported using both quantitative analyses and qualitative observations (Section 7), followed by an integrative examination of their meanings, limitations, and avenues for future research (Section 8).

2. Theoretical Background

The DKG metric is anchored in two complementary theoretical traditions, namely Information Theory, particularly Shannon's notion of entropy, and Information Processing Theory as applied in the field of Information Systems. These practical standpoints provide the quantitative foundation for measuring knowledge gain and the organizational lens through which the processing of information is understood as a mechanism of learning and adaptation.

Information Theory, introduced by Claude Shannon in 1948, established a mathematical structure for representing, transmitting, and measuring information in communication systems. Central to this theory is the concept of *entropy*, which quantifies the level of uncertainty or unpredictability within a probability distribution. In essence, entropy measures the information required to resolve uncertainty about a set of possible outcomes [Shannon 1948].

In the context of search and learning, entropy provides a means of modeling the uncertainty faced by users before encountering new information. As searchers reformulate queries and interact with documents, their uncertainty is progressively reduced [Jaynes 1985]. The DKG derives from this principle as it operationalizes knowledge gain

as a measurable decrease in entropy across a search session, supplemented by semantic similarity between queries and retrieved documents. This theoretical foundation ensures that the metric captures the amount of information processed and its conceptual alignment with the user's evolving knowledge state.

While Information Theory offers a formal quantitative basis, the theoretical anchoring of this research within Information Systems is provided by Information Processing Theory. Originally developed in cognitive psychology and later extended to organizational studies, e.g., [Galbraith 1973] and [Daft and Lengel 1986], this theory views individuals and organizations as information-processing entities whose effectiveness depends on their capacity to reduce uncertainty and handle complexity [Haußmann et al. 2012].

In IS research, Information Processing Theory underlines how people, processes, and technologies interact to collect, interpret, and act upon information. It has been widely applied to explain how organizations cope with information overload, adapt to dynamic environments, and design structures that match their information-processing requirements [Haußmann et al. 2012]. Within this approach, uncertainty decrease is a socio-technical process shaped by human strategies and technological affordances.

By bringing together Shannon's entropy and Information Processing Theory, the DKG metric establishes a dual grounding that connects a quantitative foundation with an Information Systems theoretical perspective. In this sense, the decrease in entropy provides an accurate measure of knowledge acquisition during search, while Information Processing Theory situates this measurement within the IS domain, focusing on the interaction between users (people), their search behaviors (processes), and the computational tools that support them (technologies).

This study adopts a pragmatist epistemological stance, common in Information Systems research, in which the value of knowledge lies in its practical consequences and explanatory usefulness [Boucher 2018]. From this perspective, learning is treated as a functional construct that can be approximated through multiple complementary indicators (e.g., behavioral traces, test-based outcomes, and qualitative evidence) [Kivinen and Ristela 2003].

3. Related Works

Measuring learning in search sessions is a central concern of SaL research [Urgo and Arguello 2022]. Assessment methods vary depending on whether the focus lies on the searcher or on the search engine. While pre/post-tests, open-ended methods, and self-reports target the user directly, search engines typically rely on assumed evidence of learning, inferred from behavioral traces such as clicks, dwell time, and query reformulations [Tibau 2024].

Studies based on assumed evidence of learning focus on how systems can infer user knowledge states implicitly. Early work by [Chi et al. 2016] and subsequent research [Xu et al. 2020] [Otto et al. 2021] [Yu et al. 2021] [Gritz et al. 2021] [Otto et al. 2022] [El Zein and da Costa Pereira 2022] analyzed behavioral proxies, such as query changes, reading sequences, and document interaction patterns, as signals of evolving knowledge. More recently, [El Zein and Da Costa Pereira 2023] and [Liu et al. 2023] advanced this approach with knowledge-graph-based models to approximate conceptual growth during

exploratory tasks. Eye-tracking studies, such as [Gritz et al. 2024], reinforced the connection between reading strategies and knowledge acquisition. Similarly, [Câmara 2024] proposed frameworks to integrate behavioral data into search engines, emphasizing query refinements and task attributes, while [Liu et al. 2025] identified 16 attributes shaping exploratory search, from domain expertise to motivation.

These efforts collectively show that search systems can move beyond retrieval to stimulate and measure learning. However, current behavioral proxies face two main limitations. First, they often treat learning as a byproduct of interaction, rather than modeling it as a structured cognitive process. Second, measures of assumed evidence, while scalable, struggle to capture qualitative changes in knowledge, especially in complex or open-ended tasks.

The Degree of Knowledge Gain metric builds on these insights by introducing a more cognitively grounded approach, and extends assumed-evidence approaches by offering a scalable behavioral indicator and a formalized model of knowledge evolution, making it especially suitable for search environments designed with educational intent, where both the quantity and the quality of learning must be assessed. Unlike these prior approaches, which infer learning indirectly from isolated behavioral signals or post-hoc analysis, the DKG enables real-time estimation of knowledge gain during unconstrained, natural search sessions, as seen in Table 1.

Table 1. Comparison between DKG and existing automatic learning measurement approaches

Approach	Learning signal	Intrusiveness	Ecological validity	Strategy sensitivity
Gritz et al.	Reading sequences, gaze	Low	Medium	Limited
Otto et al.	Multimedia consumption	Low	Medium	Limited
El Zein & Pereira	Behavioral proxies	Low	Medium	Moderate
Câmara	Query reformulation patterns	Low	Medium	Moderate
DKG metric	Entropy reduction + semantic similarity	None	High	High

4. The DKG Metric

The DKG metric is based on Shannon's Entropy (H), which formalizes uncertainty as a probability distribution of possible outcomes [Prieto-Guerrero and Espinosa-Paredes 2019]. In a search context, the entropy reflects the user's uncertainty before encountering new information. As learning occurs, entropy decreases, indicating refinement of knowledge. Expanding on this idea, DKG incorporates semantic similarity between queries and clicked documents using the Jaccard similarity coefficient, estimating the degree of term overlap between reformulations and ensuring that knowledge gain reflects both novelty and conceptual alignment.

The formulation of additive information gain [Milne 2012] underpins the qualitative dimension of DKG. Each new information element β contributes incrementally to what is already known from α, formalized as:

$$IG(\alpha \wedge \beta, \Gamma) = IG(\alpha, \Gamma) + IG(\beta, \Gamma \cup \alpha) \tag{1}$$

where $IG(.)$ denotes an additive information gain function that quantifies the contribution of newly acquired information relative to an existing knowledge base and Γ denotes the

user's knowledge base. This represents how new knowledge is iteratively built across search sessions.

The quantitative model reformulates the entropy to reflect query-level contributions:

$$H(Q) = -\sum_{k=1}^{n} Q(\alpha_k) \log P(\alpha_k, \Gamma) \tag{2}$$

Here, $Q(\alpha_k)$ is the probability distribution over queries, and $P(\alpha_k, \Gamma)$ the probability that query α_k contributes relevant information to the knowledge base Γ.

The DKG is then defined as [Tibau 2024]:

$$\deg_{KG} = \left(1 - \sum_{i=1}^{n} p_i \log \frac{m_i}{p_i}\right) \times 0.01 \tag{3}$$

In this formulation:

- p_i is the detection probability of a clicked document at rank i, modeled from empirical user-click data [Dupret and Piwowarski 2008];
- m_i is the observation coefficient, combining a click indicator (c_i)[1] with the Jaccard similarity between consecutive queries.

Thus, m_i captures the joint effect of user actions (query reformulation, clicking) and system responses (ranking), while p_i reflects how likely the observed interaction would occur, and the scaling factor adjusts interpretability.

5. Development of the Artifacts

The operationalization of the DKG metric required the development of two technological artifacts[2] intended to test how the metric could be embedded in real search environments, but differing significantly in scope, design, and overall effectiveness.

The first artifact was a prototype search engine that embedded the DKG metric directly into its retrieval logic; this proof-of-concept demonstrated that the metric could be integrated into ranking algorithms, enabling the system to adjust document ordering according to reductions in entropy. However, the tool presented several limitations, as the search results often failed to reflect the intent of the user, affecting the quality of the sessions. Additionally, the system relied on an external API that required a high token consumption per session, making it costly and impractical for scaling experiments. These limitations weakened the user experience and the viability of the approach, restricting its role to an exploratory validation of the metric.

To tackle these issues, a second artifact, a browser plug-in for Google Chrome, was developed. Unlike the stand-alone prototype, the plug-in operated seamlessly within the user's preferred search engine, thereby preserving the ecological validity

[1]The click indicator is defined as $c_i = 1$ if a result is clicked, and $c_i = 0$ otherwise.

[2]The artifacts are openly available at `https://doi.org/10.5281/zenodo.17203699`.

[Schmuckler 2001][3] of the experiments. Its main contribution was the ability to compute the DKG metric in real time by intercepting queries and document clicks during the search process. Each interaction was transformed into probability distributions and similarity measures that fed directly into the entropy-based calculation, providing immediate estimates of knowledge gain without interfering with the search routine.

The plug-in also introduced a structured data pipeline with all interaction logs, including queries, clicked results, timestamps, and computed DKG values, transmitted to a Supabase database. This ensured consistent storage, eliminated the costs of external APIs, and provided a sustainable infrastructure for subsequent analysis. As a result, the tool improved the ecological validity and usability of the experiments and enabled a more scalable methodology to evaluate knowledge gain in naturalistic settings.

Beyond technical improvements, the second artifact enables research questions that were not feasible with the earlier prototype search engine. Because the metric operates unobtrusively within users' habitual search environments, it becomes possible to study longitudinal learning trajectories, strategy adaptation over time, and the interaction between search behaviors and external tools (e.g., AI-assisted summaries) without constraining the retrieval system itself.

To underscore the evolution from conceptual validation to experimental applicability, Table 2 summarizes the main differences between the prototype search engine and the browser plug-in.

Table 2. Comparison between the prototype search engine and the browser plug-in

Aspect	Prototype Search Engine	Browser Plug-in
Integration with search	Stand-alone system simulating a search engine; DKG embedded in ranking logic	Embedded in Chrome, intercepting queries and clicks in the user's preferred search engine
Result quality	Limited relevance; API often returned results that did not match user intent	Results provided by established search engines chosen by participants, ensuring higher quality
Computation of DKG	Applied within the retrieval algorithm, affecting ranking order	Computed in real time during natural searches, based on entropy reduction and semantic similarity
Data storage	No structured storage; session data tied to API responses	Structured pipeline using Supabase for queries, clicks, timestamps, and DKG values
Costs and scalability	High maintenance costs due to external API tokens; limited scalability	No additional cost; scalable for multiple participants and longer sessions
Ecological validity	Artificial environment restricted to the prototype's interface	Natural environment; participants used their usual search engines without disruption
Research role	Proof-of-concept for embedding DKG in search retrieval	Mature tool enabling controlled experiments in real-world search contexts

6. Methodology

Five participants, recruited voluntarily from a Brazilian university setting, took part in the experiment designed to assess the DKG metric during online searches. They had heterogeneous academic backgrounds, including computing, education, and related fields. All procedures were approved by the institutional ethics committee (CAAE 82881624.0.0000.5285), and participants provided informed consent before the sessions. Interactions were anonymized and transcribed for analysis.

[3]Ecological validity refers to the extent to which experimental outcomes can be applied to real-world contexts, reflecting situations and environments encountered in daily life.

The present study adopted a five-user design after established evidence in usability research that this number is sufficient to achieve data saturation in interaction testing [Lu 2025] [Lazar et al. 2017]. [Virzi 1992] demonstrated that the majority of usability problems are identified within the first five users. [Nielsen and Landauer 1993] further formalized this through a Poisson model, showing that approximately 80–85% of usability issues can be uncovered with five participants, with diminishing returns from additional users. Subsequent studies, e.g., [Borsci et al. 2013], reinforced this view, particularly when methods such as the Think-Aloud protocol are applied. Importantly, this choice reflects the exploratory and theory-building nature of the study, the goal was to examine whether the DKG metric behaves coherently across users, tasks, and strategies, and whether its values align meaningfully with qualitative evidence of learning processes.

Each participant was asked to perform three search tasks of varying complexity, ranging from a simple factual task to an open-ended exploratory task:

- Task 1: Intro to weight, mass, volume, density. Explain these concepts and compare differences between weight and mass. 20 min.
- Task 2: Skeletal & muscular system. Identify main bones and muscles. Locate examples of movable joints in the human body. 25 min.
- Task 3: Limb prosthesis. Explain main characteristics of movable joints in a prosthetic limb. 25 min.

None of the participants reported formal training in all task domains (physics, anatomy, prosthetics), ensuring that tasks required active information seeking rather than recall. Differences in prior familiarity were expected and treated as part of the phenomenon under study rather than as noise. This design ensured exposure to different levels of cognitive demand and encouraged the use of diverse search strategies. Tasks were performed individually in a controlled online environment, with participants sharing their screens and verbalizing their thoughts using the Concurrent Think-Aloud (CTA) protocol [Kelley et al. 2015]. Each session lasted approximately 60–90 minutes and was fully recorded, including search queries, result clicks, and spoken reflections.

To measure learning, participants completed pre- and post-tests that evaluated their knowledge of the task topics; the resulting scores served as an external benchmark against which the behavior-based DKG metric was compared, and the test scores themselves were computed using the following equation [Hart et al. 2019]:

$$\text{ToL Score} = \frac{(\text{PostTestScore} - \text{PreTestScore})}{\text{TotalTestScore}} \tag{4}$$

The DKG was automatically calculated from query and click logs using Equation 3, modeling the progressive reduction of uncertainty throughout the search process.

Qualitative analysis complemented this measurement. Transcripts and logs were coded using two frameworks: the Online Information Searching Strategies (OISS) framework [Reisoğlu et al. 2019], which captures behavioral, procedural, and metacognitive strategies (Table 3), and the ESKiP Taxonomy of Query State [Tibau et al. 2019] (Table 4), which classifies transitions in query formulation (e.g., broadening, narrowing, or shifting focus). These schemes support a structured interpretation of both what users do during search and how their strategies evolve to refine or redirect their information needs over

time, and reliability was ensured by cross-checking annotations. Formal inter-rater reliability statistics (e.g., Cohen's kappa) are planned for future larger-scale studies, where multiple independent coders and extended datasets can support more robust reliability estimation.

All sessions were fully transcribed and anonymized[4], with identifiers replaced by user numbers. The transcripts were first segmented by task (weight/mass/volume/density, anatomy, prosthetics) to preserve comparability [Miles et al. 2013, p.89-90]. Analysis proceeded in two complementary stages. In the first, descriptive accounts were produced to capture how participants articulated concepts, navigated search engines, and perceived task complexity [Miles et al. 2013, p.243-247]. In the second stage, transcripts were systematically coded using the cited OISS framework and the ESKiP taxonomy of query states, enabling the identification of behavioral, procedural, and metacognitive patterns across participants [Saldaña 2021, 149-152]. The detailed mappings of search terms, query states, and OISS codes are available in a dataset[5].

Table 3. Online Information Searching Strategies (OISS) indicators

Category	Indicators
Behavioral (Control)	C1: Using familiar search engine; C2: Typing engine name in browser; C3: Entering website name in engine; C4: Entering URL in address bar; C5: Using "home" button; C6: Using back/forward; C7: Boolean operators; C8: Using images/videos/maps; C9: Advanced search in images/videos/maps; C10: Advanced search options.
Behavioral (Disorientation)	D1: Giving up after failure; D2: Using unrelated terms; D3: No idea how to search; D4: Feeling bad after failure.
Procedural (Trial & Error)	TE1: Modifying keywords; TE2: Trying other engines; TE3: Opening multiple sites.
Procedural (Problem-Solving)	PS1: Attempting to resolve problems; PS2: Seeking reasons for problems.
Metacognitive (Purposeful Thinking)	PT1: Narrowing field; PT2: Accessing additional sites; PT3: Searching multiple sources; PT4: In-site search.
Metacognitive (Select Main Ideas)	SMI1: Opening known site; SMI2: Typing specific terms; SMI3: Following suggestions; SMI4: Following in-site results; SMI5: Using hyperlinks; SMI6: Reading titles; SMI7: Tracking relevant info; SMI8: Using Ctrl+F.
Metacognitive (Evaluation)	E1: Evaluating relationships; E2: Comparing sources; E3: Deciding worth of info; E4: Combining and presenting data.

Table 4. ESKiP Taxonomy of Query State

State	Definition	Example
Initial (IS)	Query contains 1 term, start of a search.	Lake Como
Return (RS)	Query repeats earlier terms.	Lake Como → Lake Como Vacation → Lake Como
Generalization (GE)	Qi+1 has fewer terms than Qi.	Brazilian Flag Colors → Brazilian Flag
Specialization (SC)	Qi+1 has more terms than Qi.	Brazilian Flag → Brazilian Flag Colors
Repeat (RP)	Same terms, different order/format.	Brazilian Flag → Flag Brazilian
Word Substitution (WS)	Same length, different terms.	Brazilian Flag → Colombian Flag
New (NW)	No shared terms.	Brazilian Flag → Dog Breeds
Related (RE)	No shared terms, but related meaning.	Girl with the Dragon Tattoo Actress → Rooney Mara images

Finally, a mixed-methods integration was carried out [Plano Clark 2017]. DKG scores were triangulated with pre/post-test results and with behavioral codes from OISS and ESKiP, allowing the analysis to connect quantitative gains with the strategies and challenges that shaped them.

[4]The anonymized transcripts supporting this study are openly available at the following repository: `https://doi.org/10.5281/zenodo.17195699`.

[5]The dataset is openly available at `https://doi.org/10.5281/zenodo.17202855`.

7. Results and Findings

The results of this study combine quantitative and qualitative evidence to assess how learning gains emerged during the experimental tasks. In line with qualitative and mixed-methods traditions, analytic generalization rather than population inference was sought, focusing on pattern coherence across data sources over effect size estimation. Quantitative analyses are based on two complementary measures: the Transfer of Learning score and the Degree of Knowledge Gain metric. In parallel, transcript-based qualitative observations further contextualize these outcomes by accentuating participants' strategies, perceived task complexity, and sources of difficulty or disorientation.

7.1. Results

The quantitative outcomes of the experiment are presented in Table 5 and visualized in Figure 1. For each participant and task, two measures were computed, specifically the ToL score, derived from pre- and post-test performance, and the DKG metric, automatically calculated from search interactions. While ToL captures the change in outcome knowledge measured before and after the tasks, DKG reflects process-level evidence derived from interaction logs. In this sense, the two measures are intentionally complementary rather than interchangeable.

Table 5. ToL scores and DKG metric per user and task

User	Task	ToL Score	DKG
User_1	1	0.1	0.028
User_1	2	0.0	0.009
User_1	3	0.1	0.019
User_2	1	0.1	0.014
User_2	2	0.1	0.037
User_2	3	0.4	0.017
User_3	1	0.3	0.033
User_3	2	0.3	0.028
User_3	3	0.6	0.031
User_4	1	0.1	0.020
User_4	2	0.3	0.037
User_4	3	0.2	0.025
User_5	1	0.2	0.025
User_5	2	0.5	0.067
User_5	3	0.4	0.049

Overall, the ToL scores revealed heterogeneous learning outcomes across participants. Users 1 and 2 exhibited modest or low gains (ranging from 0.0 to 0.4), while Users 3 and 5 showed the largest improvements, reaching 0.6 and 0.5, respectively. These differences suggest variability in prior knowledge and strategy use, consistent with the transcript-based observations reported in Subsection 7.2. In contrast, the DKG metric produced lower but more fine-grained values (0.009–0.067), capturing refined distinctions across tasks and users. For example, User 2 achieved the highest ToL score in Task 3 (0.4) but showed a relatively low DKG (0.017), whereas User 5 produced the highest

DKG values (0.067 in Task 2 and 0.049 in Task 3), reflecting steady involvement with query reformulation and evaluation.

Comparison of the two measures indicates partial convergence: higher ToL scores often coincided with higher DKG values (e.g., User 3 in Task 3, User 5 in Task 2), though exceptions emphasize their complementary perspectives. While ToL explicitly captures knowledge change through tests, DKG accounts for the evolving process underlying search interactions. The scatterplot (Figure 1) shows a positive tendency, it is possible to observe that participants with higher ToL gains generally displayed higher DKG values, although the relationship was not strictly linear.

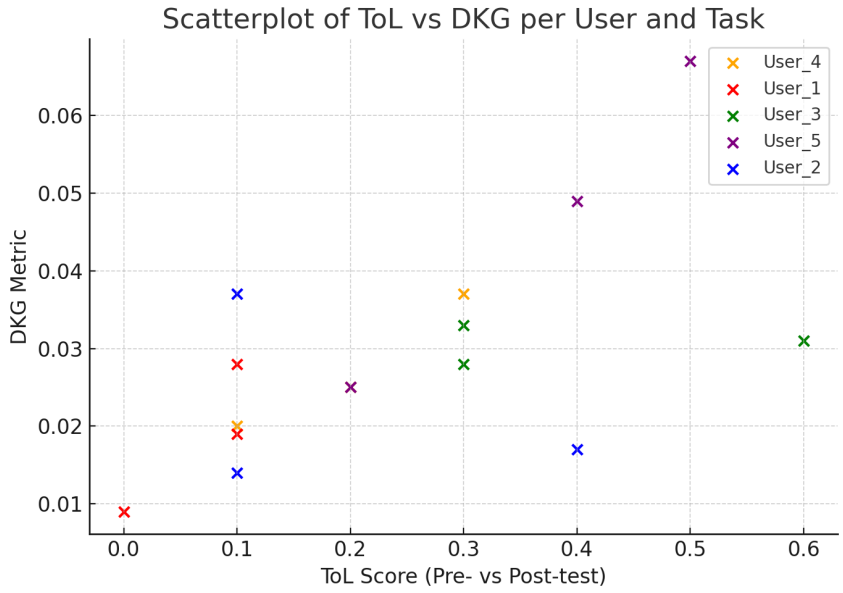

Figure 1. Scatterplot of ToL scores and DKG values per user and task.

To further examine these relationships, several statistical analyses were conducted[6] (Table 6). Pearson's correlation revealed a moderate positive association ($r = 0.62$, $p < 0.01$), indicating that higher ToL scores tend to align with higher DKG values. A simple linear regression confirmed this tendency, with ToL explaining 38% of the variance in DKG ($R^2 = 0.38$, $\beta = 0.052$, $p = 0.01$). However, given the small sample size of search sessions ($N = 15$), these results should be interpreted with caution.

From a construct validity perspective, a moderate correlation is both expected and theoretically appropriate [Strauss and Smith 2009] as the ToL score and the DKG metric operationalize different constructs. ToL captures outcome-oriented knowledge change measured at discrete time points, whereas DKG reflects process-oriented evidence derived from moment-to-moment search interactions. As such, strong convergence would risk construct redundancy, while weak or absent correlation would undermine convergent validity [Strauss and Smith 2009, Smith 2005]. The observed moderate association supports the interpretation that DKG captures a related but non-identical aspect of learning.

ANOVA tests were applied to assess differences in DKG across tasks and users.

[6]The Python scripts used to perform the statistical analyses are available in the same repository as the anonymized transcripts.

No significant differences emerged across tasks ($F(2, 12) = 1.12$, $p = 0.36$), suggesting that task type did not strongly affect metric outcomes. In contrast, significant differences were found across users ($F(4, 10) = 4.27$, $p = 0.02$), indicating that individual factors such as prior knowledge, fatigue, and strategy use played a stronger role in shaping DKG values.

Finally, agreement between ToL and DKG was examined with a Bland–Altman analysis. The mean difference was positive (bias = 0.17), with limits of agreement between -0.05 and 0.39. This reflects a systematic bias, since ToL consistently produced higher values. The observed bias reflects an expected scale difference between the two measures. ToL scores are linearly computed from test performance, while DKG is based on a logarithmic entropy formulation that compresses value ranges (Equation 3). Bland–Altman analysis is therefore used here not to assess interchangeability, but to examine consistency of relative patterns across scales. Nevertheless, most points fell within the limits of agreement, indicating that although the two measures differ in scale, their relative patterns across participants were consistent (Figure 2).

Table 6. Summary of statistical analyses comparing ToL scores and DKG values

Analysis	**Result / Interpretation**
Pearson Correlation	$r = 0.62$, $p < 0.01$. Indicates a moderate positive correlation between ToL and DKG, suggesting that higher ToL scores tend to align with higher DKG values.
Linear Regression	Regression of DKG on ToL: $R^2 = 0.38$, slope $\beta = 0.052$, $p = 0.01$. ToL explains 38% of the variance in DKG, though the limited sample size (5 participants, 15 observations) warrants cautious interpretation.
ANOVA (by task)	No significant differences in mean DKG across tasks ($F(2, 12) = 1.12$, $p = 0.36$). Suggests that task type did not strongly affect DKG outcomes.
ANOVA (by user)	Significant differences in mean DKG across users ($F(4, 10) = 4.27$, $p = 0.02$). Indicates that user-specific factors played a stronger role in explaining DKG variability than task type.
Bland–Altman Analysis	Mean difference (bias) = 0.17 (ToL higher). Limits of agreement: $[-0.05, 0.39]$. ToL regularly results in higher values due to its linear computation, while DKG's logarithmic basis compresses score ranges. Agreement is moderate but points methodological differences.

7.2. Empirical Observations

Qualitative analysis of the transcripts revealed how participants approached the three tasks and how their strategies shaped their search process. Users 1 and 2 frequently relied on familiar educational websites, such as *Brasil Escola*, and verbalized their reasoning clearly, although User 2 overlooked one of the required concepts in the first task and later compensated by reformulating the queries. In the anatomy task, User 2 systematically listed bones, muscles, and joints, while in the prosthesis task, the same participant turned

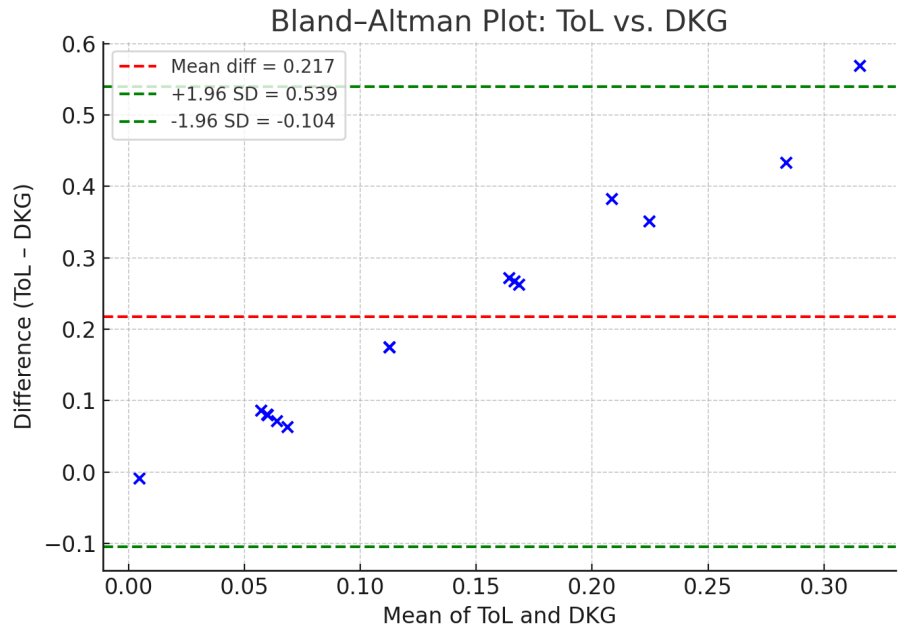

Figure 2. Bland–Altman plot comparing ToL scores and DKG values. The red dashed line represents the mean difference (bias = 0.217), while the green dashed lines indicate the limits of agreement (-0.104 to 0.539). The plot shows that ToL systematically yields higher values than DKG due to its linear computation, whereas DKG compresses score ranges through its logarithmic formulation.

to AI-generated summaries when relevant details were difficult to locate. Users 1 and 2 alike perceived the first task as low in complexity, the second as high, and the third as medium.

User 3, who declared being less accustomed to organic searching [Tomczyk et al. 2024][7], expressed preference for AI tools but completed all tasks using a search engine as instructed. In the first task, the participant articulated the distinction between mass and weight with examples and later identified different types of joints. In the prosthesis task, the participant emphasized adjustability, stability, and usability in daily activities. User 3 classified the first task as highly complex, given a limited background in physics, but found the anatomy and prosthesis tasks easier and closer to their expertise. The transcript indicates awareness of how search practices have evolved, showing a shift from exploratory browsing to AI-mediated synthesis.

User 4 used Bing with Copilot and focused on gaps identified in the pre-tests, often consulting videos to visualize abstract concepts such as density. The participant explained weight, mass, and volume in detail and used structural breakdowns to describe bones, muscles, and prosthetic components. Queries were frequently reformulated when results were incomplete, complemented by written notes to organize reasoning. User 4 classified Tasks 1 and 3 as highly complex, and Task 2 as medium, noting that fatigue contributed to the perceived difficulty.

[7]The term organic searching refers to the process of manually formulating queries, browsing search engine results, and selecting information sources, as opposed to relying on AI-generated summaries or automated answer tools.

User 5 had minor technical issues during the plug-in installation, but continued to complete every task. In the first, the participant gave consistent explanations of mass, weight, volume, and density, using concrete examples to distinguish between different contexts. In the anatomy task, User 5 distinguished between synovial and fibrous joints, described stages of bone repair, and listed key skeletal and muscular structures. For prosthetic joints, the participant emphasized amplitude, stability, durability, and ease of maintenance, complementing textual information with videos when technical content was dense. The first task was rated as medium complexity and the others as high.

Together, the transcripts illustrate how participants adapted their strategies according to their familiarity with the topics and the difficulties encountered. Query reformulation was extensive, most often through Specialization or Word Substitution, while Generalization was rare. The prosthesis task consistently emerged as the most difficult, leading some participants to consult AI responses or academic resources. The perception of task complexity varied by user profile, and emotional responses, such as frustration or fatigue, influenced engagement. Despite these difficulties, persistent reformulation and source evaluation often enabled participants to complete the tasks successfully.

7.3. Findings

The initial analysis of the transcripts showed the use of various search strategies, often reflecting the participants' familiarity with search engines and their ability to adjust the process. Procedural behaviors such as Trial-and-Error (e.g., modifying keywords, trying different sites) appeared frequently, especially when participants found it difficult to refine their queries. For instance, User 1 repeatedly reformulated queries when the results were inadequate, demonstrating iterative problem-solving and strategic keyword adjustment.

Metacognitive strategies such as Purposeful Thinking (e.g., narrowing scope or in-site searching) were more evident in participants with higher ToL scores. User 3, for example, critically examined sources by comparing and contrasting information through multiple sites. Disorientation indicators were also evident. User 2 showed signs of frustration when initial queries failed, manifesting behaviors such as giving up after failure and a negative emotional response. Such behaviors suggest a clear association between cognitive control and knowledge gain results.

Using the ESKiP taxonomy, it was observed that most participants consistently applied Specialization (SC) and Word Substitution (WS) to refine their queries as noted in Subsection 7.2. For example, User 5 began with general queries, then moved to more detailed formulations, adding specific terms or modifying them entirely to improve precision. Specialization was particularly common when tasks required synthesis of current information. Word Substitution captured attempts to match query vocabulary with relevant documents, showing adaptability. Interestingly, few Generalization (GE) moves were observed, suggesting that participants tended to narrow their searches instead of expanding them when difficulties occurred.

The pre- and post-test comparison confirmed measurable increases in knowledge for all participants, though to different degrees. Those with more evaluative and integrative strategies (e.g., cross-comparing sites, identifying main ideas) had the highest DKG values, while those who displayed more signs of disorientation achieved lower gains.

The qualitative analysis reinforces that DKG is sensitive to strategy use. Auto-

matic measures alone can detect changes in test performance, but the transcript analysis reveals why those changes occur. For instance, although User 2 attempted multiple Trial-and-Error reformulations in Task 3, these were fragmented (e.g., repeatedly narrowing a specific query into near-duplicate queries without yielding new information) and eventually replaced by reliance on AI summaries, leading to limited gains. By contrast, User 4's fewer but more targeted refinements, combined with evaluative strategies, contributed to eventual success in the same task.

8. Conclusion

This study examined the Degree of Knowledge Gain as an automatic indicator of learning in search, advancing it from a conceptual proposal to a deployable, real-time metric embedded in a browser plug-in. Based on Shannon's entropy and located within Information Processing Theory, DKG operationalizes knowledge gain as uncertainty decrease reflecting evolving user information needs. Across three tasks, the quantitative analyses showed a *moderate positive association* with test-based gains (ToL) (e.g., $r = 0.62$, $R^2 = 0.38$), *significant differences by user but not by task* (ANOVA), and *systematic scale differences* in a Bland–Altman analysis, with ToL score producing higher values because of its linear computation, and DKG compressing ranges due to its logarithmic form. These findings indicate that DKG and ToL capture related yet distinct elements of learning, since ToL reflects outcome change, while DKG underlines processual evidence during search. From a pragmatic measurement standpoint, the lack of scale equivalence between ToL and DKG is not a limitation but a consequence of their distinct theoretical foundations and intended uses, one as an outcome benchmark, the other as a process-sensitive indicator.

Qualitative results clarify why scores diverge, that is, higher DKG values coincided with purposeful strategies (specialization, evaluation, and triangulation of sources), whereas lower values were often accompanied by disorientation or fragmented reformulations (e.g., alternating between overly broad and overly specific terms before defaulting to AI-generated summaries). Taken together, these results support DKG as a promising proxy for knowledge gain provided it is interpreted alongside behavioral evidence (OISS/ESKiP) and, when possible, an external benchmark such as pre/post testing.

From an Information Systems perspective, the contribution is twofold: technologically, a plug-in and data pipeline that compute DKG unobtrusively in natural search environments were provided, improving ecological realism over a bespoke prototype engine; methodologically, a mixed-methods evaluation was demonstrated that links interaction logs, cognitive proxies, and learning outcomes, advancing the GranDSI-BR agenda by integrating people (searchers and their strategies), processes (search and learning workflows), and technologies (metrics and tools) in a sociotechnical assessment relevant to the online learning / science education domain.

The availability of a plug-in that computes DKG in real time also opens new research avenues, including the study of learning across extended periods, comparisons between organic searching and AI-mediated exploration, and adaptive interventions that respond dynamically to detected learning plateaus. Beyond educational settings, the study has implications for organizational information systems in contexts where learning, sensemaking, and knowledge acquisition are embedded in daily work practices. In knowledge-intensive organizations, such as public administration, healthcare, journalism, and soft-

ware development, search activities are tightly coupled with decision-making, problem solving, and organizational learning. The ability to unobtrusively capture learning signals during real search sessions enables organizations to assess how information systems support knowledge work without disrupting established workflows.

Limitations include the small and single-site sample, short task windows, and topic scope, which warrant caution regarding over-generalization and constrain statistical power. DKG's magnitude is not directly comparable with test scores; rather, its value lies in *sensitivity to strategy* and *moment-to-moment progression*. Future work should scale to larger and more diverse samples and domains; run longitudinal studies to assess retention and transfer; calibrate DKG to external outcomes via mixed-effects models or per-task normalization; enrich the metric with additional semantic/contextual cues (e.g., query intent, source credibility) and incorporate fairness/ethics checks; and explore *adaptive search* that closes the loop, using DKG in real time to recommend next actions (e.g., broadening/narrowing queries, comparative reading) that promote learning, transparency, and inclusion.

References

Borsci, S., Macredie, R. D., Barnett, J., Martin, J., Kuljis, J., and Young, T. (2013). Reviewing and extending the five-user assumption: A grounded procedure for interaction evaluation. *ACM Trans. Comput.-Hum. Interact.*, 20(5).

Boscarioli, C., Araujo, R. M., and Maciel, R. S. P. (2017). *I GranDSI-BR: Grand Research Challenges in Information Systems in Brazil 2016–2026*. Brazilian Computer Society (SBC), Special Committee on Information Systems (CE-SI).

Boucher, S. (2018). Stances and epistemology: Values, pragmatics, and rationality. *Metaphilosophy*, 49(4):521–547.

Câmara, A. (2024). *Designing Search-as-Learning Systems*. Ph.d. thesis, Delft University of Technology. ⟨DOI: 10.4233/uuid:0fe3a6bb-1bc1-40e2-86b0-ec3d3aef9c77⟩, ⟨ISBN: 978-94-6384-569-4⟩.

Câmara, A., Roy, N., Maxwell, D., and Hauff, C. (2021). Searching to learn with instructional scaffolding. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, page 209–218, New York, NY, USA. Association for Computing Machinery.

Chi, Y., Han, S., He, D., and Meng, R. (2016). Exploring knowledge learning in collaborative information seeking process. In *SAL@SIGIR*.

Daft, R. L. and Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, 32(5):554–571.

Dupret, G. E. and Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 331–338, New York, NY, USA. Association for Computing Machinery.

El Zein, D. and da Costa Pereira, C. (2022). User's knowledge and information needs in information retrieval evaluation. In *Proceedings of the 30th ACM Conference on User*

Modeling, Adaptation and Personalization, UMAP '22, page 170–178, New York, NY, USA. Association for Computing Machinery.

El Zein, D. and Da Costa Pereira, C. (2023). The evolution of user knowledge during search-as-learning sessions: A benchmark and baseline. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, page 454–458, New York, NY, USA. Association for Computing Machinery.

Galbraith, J. R. (1973). *Designing Complex Organizations*. Addison-Wesley, Reading, MA.

Gritz, W., Hoppe, A., and Ewerth, R. (2021). On the impact of features and classifiers for measuring knowledge gain during web search-a case study.

Gritz, W., Hoppe, A., and Ewerth, R. (2024). On the influence of reading sequences on knowledge gain during web search. In Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., and Ounis, I., editors, *Advances in Information Retrieval*, pages 364–373, Cham. Springer Nature Switzerland.

Hart, S. L., Steinheider, B., and Hoffmeister, V. E. (2019). Team-based learning and training transfer: a case study of training for the implementation of enterprise resources planning software. *International Journal of Training and Development*, 23(2):135–152.

Haußmann, C., Dwivedi, Y. K., Venkitachalam, K., and Williams, M. D. (2012). A summary and review of galbraith's organizational information processing theory. *Information Systems Theory*, pages 71–93.

Jaynes, E. T. (1985). *Entropy and Search Theory*, pages 443–454. Springer Netherlands, Dordrecht.

Kelley, T., Capobianco, B., and Kaluf, K. (2015). Concurrent think-aloud protocols to assess elementary design students. *International Journal of Technology & Design Education*, 25(4):521 – 540.

Kivinen, O. and Ristela, P. (2003). From constructivism to a pragmatist conception of learning. *Oxford review of education*, 29(3):363–375.

Lazar, J., Feng, J. H., and Hochheiser, H. (2017). Chapter 10 - usability testing. In Lazar, J., Feng, J. H., and Hochheiser, H., editors, *Research Methods in Human Computer Interaction (Second Edition)*, pages 263–298. Morgan Kaufmann, Boston, second edition edition.

Liu, C., Song, X., and Hansen, P. (2023). Characterising users' task completion process in learning-related tasks:a search pace model. *Journal of Information Science*, 49(6):1462–1480.

Liu, Y., Qin, C., Ma, X., Chen, J., He, H., and Mao, J. (2025). Characterising exploratory search tasks: Evidence from different fields. *Journal of Information Science*, 0(0):01655515251330611.

Lu, J. 2025). A systematic literature review of usability definitions and psychometric properties of instruments in the field of learning design and technology. *Journal of Research on Technology in Education*, 0(0):1–22.

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Commun. ACM*, 49(4):41–46.

Miles, M. B., Huberman, A. M., and Saldaña, J. (2013). *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications, Thousand Oaks, CA, 3 edition.

Milne, P. (2012). Probability as a Measure of Information Added. *Journal of Logic, Language and Information*, 21(2):163–188.

Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, page 206–213, New York, NY, USA. Association for Computing Machinery.

Otto, C., Rokicki, M., Pardi, G., Gritz, W., Hienert, D., Yu, R., von Hoyer, J., Hoppe, A., Dietze, S., Holtz, P., Kammerer, Y., and Ewerth, R. (2022). SaL-Lightning Dataset: Search and Eye Gaze Behavior, Resource Interactions and Knowledge Gain during Web Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 347–352, Regensburg Germany. ACM.

Otto, C., Yu, R., Pardi, G., von Hoyer, J., Rokicki, M., Hoppe, A., Holtz, P., Kammerer, Y., Dietze, S., and Ewerth, R. (2021). Predicting knowledge gain during web search based on multimedia resource consumption. In Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., and Dimitrova, V., editors, *Artificial Intelligence in Education*, pages 318–330, Cham. Springer International Publishing.

Plano Clark, V. L. (2017). Mixed methods research. *The Journal of Positive Psychology*, 12(3):305–306.

Prieto-Guerrero, A. and Espinosa-Paredes, G. (2019). 7 - nonlinear signal processing methods: Dr estimation and nonlinear stability indicators. In Prieto-Guerrero, A. and Espinosa-Paredes, G., editors, *Linear and Non-Linear Stability Analysis in Boiling Water Reactors*, Woodhead Publishing Series in Energy, pages 315 – 398. Woodhead Publishing.

Reisoğlu, I., Çebi, A., and Bahçekapılı, T. (2019). Online information searching behaviours: examining the impact of task complexity, information searching experience, and cognitive style. *Interactive Learning Environments*, 0(0):1–18.

Roy, N., Moraes, F., and Hauff, C. (2020). Exploring users' learning gains within search sessions. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, page 432–436, New York, NY, USA. Association for Computing Machinery.

Saldaña, J. (2021). *The Coding Manual for Qualitative Researchers*. SAGE Publications, London, 4 edition.

Schmuckler, M. A. (2001). What is ecological validity? a dimensional analysis. *Infancy*, 2(4):419–436.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Smith, G. T. (2005). On construct validity: issues of method and measurement. *Psychological assessment*, 17(4):396.

Strauss, M. E. and Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, 5:1–25.

Tibau, M. (2024). *Quantifying Knowledge Gain in Online Searches: The DKG Metric*. Tese (doutorado em informática), Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Rio de Janeiro. Programa de Pós-Graduação em Informática.

Tibau, M., Siqueira, S. W. M., and Nunes, B. P. (2022). The impact of non-verbalization in think-aloud: Understanding knowledge gain indicators considering think-aloud web searches. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 107–120, New York, NY, USA. Association for Computing Machinery.

Tibau, M., Siqueira, S. W. M., and Nunes, B. P. (2023). Accounting for the knowledge gained during a web search: An empirical study on learning transfer indicators. *Library & Information Science Research*, 45(1):101222.

Tibau, M., Siqueira, S. W. M., Nunes, B. P., Nurmikko-Fuller, T., and Manrique, R. F. (2019). Using query reformulation to compare learning behaviors in web search engines. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161-377X, pages 219–223.

Tomczyk, P., Brüggemann, P., Mergner, N., and Petrescu, M. (2024). Are ai tools better than traditional tools in literature searching? evidence from e-commerce research. *Journal of Librarianship and Information Science*, 0(0):09610006241295802.

Urgo, K. and Arguello, J. (2022). Learning assessments in search-as-learning: A survey of prior work and opportunities for future research. *Information Processing & Management*, 59(2):102821.

Vakkari, P. (2016). Searching as learning: A systematization based on literature. In *Journal of Information Science, 42(1)*, pages 7–18.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Hum. Factors*, 34(4):457–468.

Xu, L., Zhou, X., and Gadiraju, U. (2020). How does team composition affect knowledge gain of users in collaborative web search? In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 91–100, New York, NY, USA. Association for Computing Machinery.

Yu, R., Tang, R., Rokicki, M., Gadiraju, U., and Dietze, S. (2021). Topic-independent modeling of user knowledge in informational search sessions. *Information Retrieval Journal*, 24(3):240–268.