frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 29 April 2024
DOI 10.3389/fcomp.2024.1366934

Differences in source selection and their effect on the quality of written statements during a critical online reasoning performance assessment

Dominik Braunheim*, Lisa Martin De Los Santos Kleinz and Lukas Trierweiler

Department of Business and Economics Education, Johannes Gutenberg University Mainz, Mainz, Germany

Check for updates

OPEN ACCESS

EDITED BY
Katarzyna Anna Bobrowicz,
University of Luxembourg, Luxembourg

REVIEWED BY
Andy Lücking,
Goethe University Frankfurt, Germany
Katri Kleemola,
University of Helsinki, Finland

*CORRESPONDENCE
Dominik Braunheim
✉ dobraunh@uni-mainz.de

RECEIVED 10 January 2024
ACCEPTED 12 April 2024
PUBLISHED 29 April 2024

CITATION
Braunheim D, Martin De Los Santos Kleinz L and Trierweiler L (2024) Differences in source selection and their effect on the quality of written statements during a critical online reasoning performance assessment. *Front. Comput. Sci.* 6:1366934.
doi: 10.3389/fcomp.2024.1366934

The importance of researching information on the Internet is ever increasing. While ease of use and accessibility are advantages, there is a risk of not being able to adequately assess the relevance and reliability of online sources. With data from the BRIDGE-project (*N* = 143) we assess the online movements of respondents and track how they researched digitally to answer a task on a generic, accessible topic. We then draw conclusions about their search behavior during this open online performance assessment. The controlled and monitored environment allowed to identify differences in their source selection and how those had an impact on the quality of their written statements respective to the given task.

KEYWORDS

higher education, critical online reasoning, media use, source selection, performance assessment

1 Introduction

Increasing digitization has revolutionized the higher education landscape and significantly changed learning materials available to students. Across all academic disciplines, the Internet has emerged as one of the most important learning media. Digital learning materials and online media have replaced offline sources (i.e., textbooks and print media) to a large degree for students in higher learning (Gasser et al., 2012; Steffens et al., 2017; Maurer et al., 2020). While the ease of use and accessibility are valid advantages, a central drawback of online sources is the danger of lacking relevance, factuality, and reliability (assured through skilled editors, which textbooks usually provide). Critically selecting, reflecting on, and arguing with online sources is vital in an environment that heavily relies on this medium (Brooks, 2016; Wineburg and McGrew, 2017). Assessing online media becomes a responsibility for users who often lack the skills required to evaluate sources and avoid the risks associated with inaccurate information or targeted misinformation (Fogg et al., 2003; McGrew, 2021; Kiili et al., 2022). The ability to discern topic relevance is closely related to the process of selecting and evaluating source material. The most prominent factor in source assessment for online settings, particularly for weaker learners, is the evaluation of relevance (Goldman et al., 2012). Its significance is anticipated to further grow with the development of more intricate search engines using algorithms reinforcing existing beliefs, stereotypes, and prejudices and the increasing demand for up-to-the-minute technology, like prompt engineering (Ćurković and Košec, 2018; Zhang, 2022; Meskó, 2023). To effectively derive meaningful knowledge from information and resources available on the Internet, students must possess the skills to critically search, assess, choose, and evaluate online information and sources based on quality criteria relevant to the topic, which we defined as *Critical Online Reasoning* (COR) (for details, see Molerov et al., 2020). The purposeful use of digital media in conjunction with critical understanding, evaluating and selecting of online information are considered central skills for successful study, and can be classified as fundamental generic skills that graduates should develop for successful professional activities as they usually have no professional experience or practical knowledge and very often use the Internet as their primary research tool (Harrison and Luckett, 2019; Kuhn et al., 2020; Molerov et al., 2020; Zlatkin-Troitschanskaia et al., 2021a; Osborne et al., 2022).

The range of factors that might influence source selection and how readers' interpret information is quite broad, featuring previously held beliefs (Zlatkin-Troitschanskaia et al., 2020), biases in the media outlets (Groeling, 2013), differences in media socialization (Schweiger, 2007), but also previous knowledge and expertise in the form of online media literacy (Ashley et al., 2013). There is a multitude of possibilities and different approaches when trying to research digitally. How do students start their search on the Internet? How do they select sources, ascribe relevance and reliability? When are they satisfied with the information they researched? Digital research can be seen as a process with many steps and many influencing factors on multiple ends, be it the digital information landscape or personal differences on the users' end. Open web simulation, here in the sense of the COR assessment, is a useful tool to gain insight into a respondent's research behavior, while reducing outside restrictions through the preselection of sources or other artificial stimuli on part of the task design (Shavelson et al., 2019). With this method it is possible to differentiate interpersonal differences between the respondents, processual differences in search behavior, website selection and synthesis, contextual differences based on different tasks used and lastly performance differences within the written statements. This allows for a narrower framework while analyzing the complexities of digital research. On basis of the findings from this type of assessment our guiding questions are: How do students and young professionals research and select sources given a specific problem? How does their source selection influence the quality of their answers on the given problem?

COR outlines skills such as searching, selecting, accessing, processing, and making critical inferences from online information to solve specific general or domain-specific problems or tasks. Subsequent to this, a novel COR assessment (CORA) was created and underwent its first validation (Molerov et al., 2020; Zlatkin-Troitschanskaia et al., 2021b). According to the COR construct definition, the CORA consists of various real-life situational tasks in an online environment, to objectively and reliably assess the COR skills of young professionals through a realistic performance assessment. Initially, respondents were asked to independently search for sources on the subject, assess their reliability (for each of the sources), and write a cohesive argument regarding the various tasks. In this open-web performance assessment, we capture real-time online information usage in a direct and valid way and collect process data (e.g., log files with timestamps and an activity log), performance data (e.g., performance scores), and personal data (e.g., questionnaires). Based on the data at hand, we are able to summarize the selected online sources, the reliability respondents attributed to each of them, how they incorporated the source materials into their written arguments and how they performed on the critical reasoning assessment (more on the scoring scheme in chapter 3). Aside from the performance and process data, there is extensive additional data on the respondents themselves. The dataset allows to control for differences in sociodemographic factors, cognitive ability, psychological characteristics, and educational background. Of major interest within this design was the quality of the websites that respondents researched in their limited time and how they evaluated the quality of these websites. To be able to differentiate the diversity of websites used, we relied on a website typology, mainly distinguishing the entities who authored the respective information (more in chapter 2.2). Gaining a better understanding of the differences in media use and their effects on *Critical Online Reasoning* (COR) skills (see chapter 2.1) should lead to diagnostic advantages to foster learning opportunities in the respective disciplines. Additionally, a more systematic perspective on the patterns of specific use of digital media outlets and their relationship to fostering COR ability is a promising way to adopt and emphasize targeted media formats in the future.

In this paper, sources that respondents utilized to answer a generic COR task as well as exploring trends in source selection are analyzed. For this purpose, the quality of selected sources was evaluated along a rating scheme built on different theories and perspectives of media studies and communication science (see chapter 2.2). Furthermore, we examine how differences in the quality of source selection affect the overall quality of the written statements in the assessment. Following this research interest, this paper relies on data from the *BRIDGE* project (Sample $N = 143$), which primarily investigates the ability of young professionals of three different domains (medicine, law, and teaching) to critically select and assess online sources and extract their content into an informed and reflected argument. Participating respondents underwent up to two generic COR tasks. The 1st measurement, which is also focus of the analysis at hand, features an open access internet search on the potential health benefits of e-bikes. This task will be focus of the following analysis, which does not cover the domain-specific variants of the assessment (which were different for each of the three domains).

In chapter 2 we introduce the *BRIDGE* project, the COR framework and the concept of the website categorization, followed by detailed research questions (RQ). Chapter 3 introduces the assessment method, the task specifics, and the scoring scheme. Chapter 4 features the results of our analysis, which are then discussed in relation to the RQ's in chapter 5. At last, we mention limitations of the design/data and an outlook for future work in chapter 6.

2 The BRIDGE project

BRIDGE is focused on the analysis and promotion of competencies in critically dealing with online media sources and content (for more details, see Zlatkin-Troitschanskaia et al., 2021b) he target demographic are young professionals from three domains (teaching, medicine, and law) at the intersection of the last year of their studies and their first year in their respective professional careers, a probationary training period common in a lot of German professions. This allows for the analysis of not only generic but also domain-specific competencies with advanced students. Competencies are framed holistically as dispositions, knowledge, and ability to prepare and answer a realistic task. The focus on task related competencies allows us to monitor whether students would be able to solve problems they might encounter in their everyday lives (generic) or within their professional day to day (domain-specific). This can be seen as a performance-oriented approach (Shavelson et al., 2019), which is focused on problem-solving within a specific assignment, producing an open-ended solution. There is still a lot of untapped potential for university curricula to adopt and incorporate opportunities for students to foster critical media and source evaluation abilities. These diagnostics of COR skills are a step to make them more targeted and shape them not only towards the students' needs, but also towards the specific challenges open web-search poses to them. In addition, the *BRIDGE* project offers specific training modules, which a subsample of respondents joined voluntarily, to support and enhance students' COR abilities. At the end of each survey, participants received an expense allowance and personalized feedback on their task completion and training performance.

2.1 Critical online reasoning

Based on previous research, we analyzed existing concepts and models theoretically in terms of connections and intersections between them, with a focus on how students evaluate and analytically justify and weigh online information (see Molerov et al., 2020). To have a valid measurement of students' critical task-solving ability we use the COR construct, which combines a holistic conceptual framework of researching online media, assessing, and evaluating them, synthesizing the information, and writing a problem-specific statement, with a computer-based measurement approach. It challenges students to differentiate the relevance and reliability of sources, while also promoting freely written answers based on the information, they deem suitable. These key facets of critical and analytical thinking are increasingly important for the ability to select suitable sources in a growing digital environment (Wineburg and McGrew, 2017).

Molerov et al. (2020) define the COR-concept as process, content, domain, and development oriented. It covers the abilities of students in searching, selecting, evaluating, and processing online information given a specific problem to solve. To describe these processes in their entirety it relies on three overlapping facets: "(i) *Online Information Acquisition* (OIA) abilities (for inquiry-based learning and information problem solving) (ii) *Critical Information Evaluation* (CIE) abilities to analyze online information particularly in terms of its credibility and trustworthiness, and (iii) abilities to use the information for *Reasoning based on Evidence, Argumentation, and Synthesis* (REAS), weighting (contradictory) arguments and (covert) perspectives, while accounting for possible misinformation and biases" (Molerov et al., 2020, p. 7). Additionally, it is assumed that the process of COR needs a stimulant or impulse to start. Critical reflection and reasoning are not necessarily the default mode of cognitive processes, which can also be automatic, i.e., following habits and heuristics (see Verplanken and Orbell, 2022). Metacognitive Activation (MCA) is necessary to prompt critical reflection on the given information (Molerov et al., 2020). In *BRIDGE* the MCA is given by the task design, which asks respondents explicitly to research sources, evaluate their suitability and confronts them with substantiating their reasoning.

In detail the COR construct is intended to represent the process a person goes through when researching online information: OIA is focused on the initial search phase, covering the use of search engines and databases, specifying queries and the assessment of relevance; CIE in the next step addresses the selection within the found websites, the evaluation of the website features, as well as the general reliability and quality of the information found; REAS covers the synthesis of the information, within sources deemed reliable, into a weighted, evidence-based argument (Molerov et al., 2020). Arguments are not necessarily deterministic. On a complex topic, a well-founded argument leaves room for both uncertainty and deliberation between opposed positions (Walton, 2006). In this sense, there is no single correct answer to the given task, but an open spectrum of valid pathways of argumentation, based on information of differing quality and reliability.

2.2 Website categorization and evaluation

To differentiate between the reliability of websites we categorized them within an ordinal typology (Table 1) (see Hesse, 2018; Nagel et al., 2020).

While the use of search engines like *Google* was tracked and counted, since they provide a meaningful first step, even for experienced researchers (Speicher et al., 2015; McGrew et al., 2017), they were not given a value in the typology. The reason is, that they can be seen as an in-between-step of getting to the websites which finally hold the information the students were using. This is also the method used for other websites, like translation sites, which helped students with their interpretations, but were not the primary source they drew from.

The other categories go from 1 to 5, with 1 being the lowest value, covering online shops, social media and miscellaneous sites of the World Wide Web. The value 2 was given to encyclopedias, 3 to news pages and associations, 4 to curated professional journals, and Google

TABLE 1 Website categorization.

	Category	Reliability score
5	Lecture notes	5
	University database	5
	Scientific database	5
	Scientific journal	5
	Scientific research institute	5
	Government bodies	5
4	Specialist magazine	4
	Google scholar	4
3	News page	3
	Associations	3
2	Encyclopedia	2
1	Online-shop	1
	Social media	1
	World wide web	1
0	Search engine	0

Scholar. The highest value 5 was given to scientific databases, lecture notes of universities, governmental bodies, university library catalogues and databases, scientific journals, and research institutes (see Table 1). The categories reduce the complexity of the web to depict search behavior and are in line with various theoretical strains of media studies.

Social media, while occasionally moderated, do not tend to have set standards on the quality of content or the reliability of the information presented, which makes them regarded as generally untrustworthy (Ciampaglia, 2018; Maurer et al., 2018). This is not too surprising, since their main feature is often seen as connecting people and providing open communication channels for users, who are free to voice their opinions (Bendel, 2018). Social media articles and posts often miss any citation, which can lead to negative information acquisition (Wolfsfeld et al., 2016). Posts are often presented in a shortened, exaggerated manner, which only highlight specific parts of the contexts they cover (Guess et al., 2019). This type of representation induces partial, subjective, and often one-sided interpretations of the topic given. Reliable information should be neutral in tone and provide all necessary information in its entirety (Kelly et al., 2018). Online-Shops fall in the same category because they act within a vested interest, framing their content to maximize economic gain. The miscellaneous category of "World Wide Web" responds to sites where the authorship is unclear and no instances of content moderation are obvious to visitors of the site, which is also detrimental to the reliability. The ease for people to create a new website and feature content has also shown a rise in blogs and non-mainstream websites advocating conspiracy theories and specific agendas, which poses its own risk (Schultz et al., 2017). For those reasons, these types of media outlets are categorized as the lowest quality (Category 1).

In contrast to social media sites, an encyclopedia's (Category 2) main purpose is to be informative, but the open-access nature of free editing can often pose an, at least temporal, risk to the quality of information (Lucassen et al., 2013). At the same time, the editing feature allows for the correction of false or misleading information, which can be seen as advantageous when contrasted with the permanence of incorrect social media posts, that stay on the platforms for years (Voss, 2005). Encyclopedias are a frequently used means for students to acquire information online and often allow a starting point in research as an alternative to search engines (Head and Eisenberg, 2009; Brox, 2012).

News pages and associations (Category 3) tend to provide full-length articles including their sources, making it possible for readers to check where the presented information stems from (Andersen et al., 2016). Mass media are an important means for many people to form opinions and have a large impact on societal perspectives (Hasebrink, 2016). While potentially being skewed in the topics they report on and with rising critiques on political biases, mass media still tend to operate under professional journalistic standards, which also apply to their digital outlets (Müller-Brehm et al., 2020). The limitations here are, that the necessity of a greater reach can lead to media outlets succumbing to sensationalism (Leif, 2001) in both topics and content framing, which is particularly a problem for the wording of headlines.

Like news outlets, specialist magazines usually have standards for publishing they must adhere to. Those can be scientific, in terms of systematic, understandable and reproducible knowledge gain (Schröder, 1994) or journalistic in the same sense as newspapers. There is a higher need for neutrality and objectivity in the presentations, leading to this type of media being categorized as the value 4.

Category 5 mainly includes established public institutions. In this sense governmental bodies and news outlets (like databases and journals) of universities and academia are held to higher standards of completeness, neutrality and factuality, with a lower level of freedom of expression. They are of societal importance by providing empirical and falsifiable information in an open manner, which generally leads to a higher level of trust from the public (Wagschal et al., 2020).

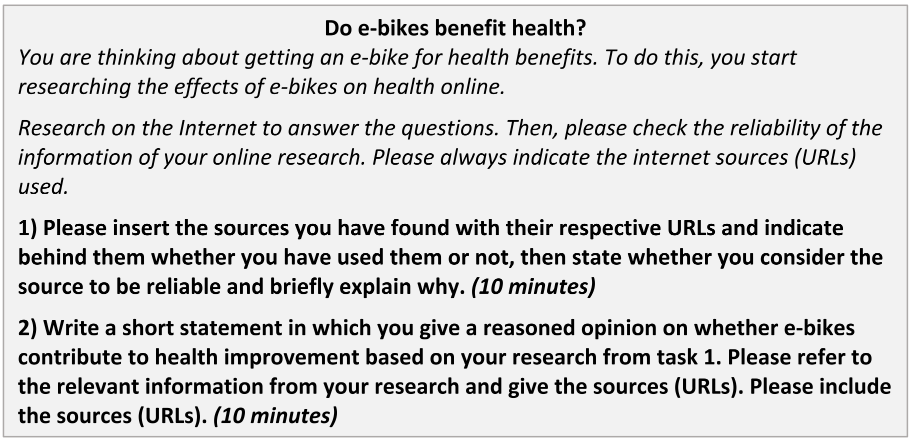
Based on the conceptualizations of the COR-construct and the categorization of different types of websites we formulate the following research questions to structure our later analysis:

1) Are there significant differences in the frequencies of use for different website categories during the assessment?
2) What is the relation between generally researched sources and sources cited for the written arguments?
3) Does the quantity of selected sources affect the quality of the research?
4) Does the quality of the research affect the quality of the final written arguments?

3 Method

The COR skills of young professionals from three domains – medicine, law, and teaching – were analyzed using performance and process data. Participants were asked to perform an open-ended web search, evaluate online information or sources, and write an open-ended response to the short CORA task (max. 20 min). The tasks were generic in nature and not reliant on domain-specific previous knowledge. The research was carried out through an online assessment platform, where participants accessed the platform individually using provided login details. Prior to the survey, participants were notified that their online activities would be recorded, and their involvement was entirely voluntary; all participants provided explicit consent through a declaration to utilize their data for research purposes. Following this, participants completed a standardized questionnaire lasting approximately 10 min, which gathered sociodemographic information such as gender, age, and general media consumption habits using the validated scale developed by Maurer et al. (2020). Participants received detailed instructions on the test procedure and the lab environment in which the assessment took place. Within a virtual machine hosted on Microsoft Azure Labs, we recorded participants' browsing histories as they completed the assigned web searches. The histories were stored in log files that included timestamps and an activity log. Participants' open-ended written responses were collected as well. The participants had access to the software required for the completion of the tasks (Internet browser, MS Office) in the virtual Lab. They were granted the flexibility to conduct the task at their preferred time and place, utilizing their personal computer, within a designated time frame.

To complete a CORA task, participants were asked to perform an open web search for reliable sources to answer a given question, e.g., related to 'health promotion through e-bikes' in the first measurement (Figure 1). The task comprised a brief contextual description with two



Do e-bikes benefit health?

You are thinking about getting an e-bike for health benefits. To do this, you start researching the effects of e-bikes on health online.

Research on the Internet to answer the questions. Then, please check the reliability of the information of your online research. Please always indicate the internet sources (URLs) used.

1) Please insert the sources you have found with their respective URLs and indicate behind them whether you have used them or not, then state whether you consider the source to be reliable and briefly explain why. *(10 minutes)*

2) Write a short statement in which you give a reasoned opinion on whether e-bikes contribute to health improvement based on your research from task 1. Please refer to the relevant information from your research and give the sources (URLs). Please include the sources (URLs). *(10 minutes)*

FIGURE 1
CORA task.

TABLE 2 Scoring scheme.

Scoring dimensions (and related COR facets)	Weights
1. Concreteness (REAS): degree of clarity and lack of ambiguity in the judgment.	10%
2. Comprehensibility (OIA, CIE, REAS): to what extent are the sources and the written statement sufficiently linked to the health benefits of e-bikes?	10%
3. Quality of Sources (OIA): content-related suitability and actuality of the sources used.	20%
4. Accuracy of source evaluation (CIE): whether and how appropriately the sources used were evaluated.	25%
5. Deliberation (REAS): are advantages and disadvantages considered in the argumentation?	15%
6. Quality of argumentation (REAS) – is the reasoning structured, conclusive and convincing?	20%

open-ended question formats: Conducting an internet search without constraints, this first part of the task focused on OIA and CIE; and then critically evaluating the online information found and writing a response for each subtask. This subsequent section, on the other hand, requires abilities in REAS. A pre-determined maximum completion time for each task and post-hoc monitoring of browsing history ensured that participants conscientiously solved the task in a controlled testing environment. Nagel et al. (2022) provide a detailed account of the evidence-based design (Mislevy et al., 2017) and validation process for the new COR assessments, which comprehensively captures research, evaluation, and online information use in a realistic Internet context. Aside from the assessment and feedback on the respective performances the project also offered optional training modules (for more detailed descriptions of training modules see Braunheim et al., 2023) on research and information synthesis skills.

The performance data was rated by three trained, independent raters along a fixed, multi-faceted rating scheme covering the three main facets of COR (OIA, CIE and REAS) across six rating dimensions (Table 2). OIA was mainly covered by the quality of sources (dimension 3); CIE was assessed within dimension 4. Rating dimensions 1, 5 and 6 are aligned to REAS, while rating dimension 2 is a general quality marker of the written statements, which intersects with all COR facets. The scores were also weighted, with more importance given to the quality (OIA) and assessment (CIE) of sources and the general quality of the written statement (REAS). Achievable points in each of the dimensions were scaled from 0 to 4 with distinctions of half a point. Final performance scores were computed based on the mean ratings by the first two raters. In cases of higher discrepancy (<0.5 points within a facet), the rating of rater 3 was included. Interrater reliability was measured by intraclass correlation coefficient (ICC) for the three independent ratings, which was satisfactory for the full scoring at $r = 0.77$ and $p = 0.00$.

Regarding research queries, we extracted and categorized all online sources that respondents entered in the two response fields of the COR assessment. All cited sources from the answers, plus the sources viewed during their research (from process data) were then evaluated by three raters according to their credibility using previously established criteria on a scale from 0 (search engines) over 1 (social media, online shops) to 5 (government bodies, scientific journals) (see Table 1). This allows us to compare which sources they have visited during their research but not included in their argumentation and draw conclusions about the selection behavior of the participants.

4 Results

The analyzed sample consists of 143 young professionals from the three domains – 68 from medicine, 30 from law, and 45 from teaching. The participants from various locations in Germany had already completed their studies and were in the practical part of their training after the 1st state examination at the time of the measurement. Gender distribution within the sample reveals that the majority (65.03%) of participants are female. The mean age of the cohort was 27.15 years, with a standard deviation of 3.51 with a range spanned from 23 to 46 years. The university entrance qualification was examined, revealing a mean grade of 1.76 with a standard deviation of 0.64. The range of grades spans from 1 to 3.6, indicating some variation in participants' prior academic achievements (Table 3). According to their own statements, 129 of the participants communicate best in German, 11 communicate as well in German as in another language, and only 4 communicate best in other languages.

4.1 Characteristics of website use

By tracking the log files during task completion, it is possible to see which websites were visited during the assessment. This makes it easier to track the number of searches and sources used by participants to find information. The first thing we looked at was how many sites from each category were visited during the research, and in a second step we analyzed which of these were used in respondents' arguments. Logfiles from 137 participants are available, as few participants had technical problems with the virtual machine. In total, the participants accessed 1.332 different websites, with each site URL only counted once per participant. Out of these, 497 were different queries on search engines, which for most respondents (135 out of 137) represented the start of their research on the topic during the 20-min assessment. In the initial searches of the 137 respondents, variations of the words "e-bike" and "health" in German were used in 111 cases. Eleven participants started their search with variations of "e-bike" and "health" in English and 14 participants searched for variations of "e-bike" without reference to health aspects. For this first task-related search, *Google* was used in 125 cases, while *Google Scholar* was used in four cases. Three participants began their research directly on the medical database *PubMed*, while others used *DuckDuckGo*, *Ecosia*, *Wikipedia*, and the website of the German Ministry of Health. Search engine views were characterized by a mean score of 3.71 different search engine requests per participant for the entire duration of the assessment, and a standard deviation of 2.69 among the 137 participants. The data indicates a wide range of search engine requests, with unique search requests ranging from zero search engine visits (attributed to technical difficulties with the logfiles of one respondent, otherwise min. 1) to 18 different search requests. The distribution is positively skewed (skewness = 1.79), indicating that a significant fraction of participants demonstrated relatively reduced interest towards search engines during the assessment. In contrast, a more extended right tail denotes a group of participants who heavily relied on search engine resources. There are also notable disparities in the conduct of the participants concerning the quantity of the 835 distinct source-views (search engine queries excluded). On average they visited 6.14 different source URLs, with a standard deviation of 3.21. The data displays a range from a minimum of one to a maximum of 18, thereby highlighting the variance in participants' information-seeking behaviors. The distribution, exhibiting a skewness of 1.04, implies a minor positive skew, inferring that most participants held a moderate number of source views, while a sub-group delved into a more comprehensive exploration of information sources.

TABLE 3 Sample description.

Sample	*N*	Male	Age	Final school grade*
Mean (variance)				
Full	143	34.97%	27.15 (12.32)	1.76 (0.41)
Medicine	68	29%	26.03 (2.61)	1.37 (0.39)
Law	30	23%	26 (1.53)	2.10 (0.63)
Teaching	45	50%	29.54 (4.32)	2.11 (0.62)

*Passing school grades in Germany are scaled from 1 "highest" to 4 "lowest".

During the assessment, the source code of each accessed website was saved, allowing us to see which results were displayed to the participants on *Google*. The sources most frequently visited by the test participants can be found in the results on the first page of *Google*, for example, when searching for "e bike gesundheit [health]" (the most frequently used search term). These sources, in descending order, include *quarks.de* (a news magazine published by a German public broadcaster) (*N* = 76), the specialist magazine *zeitschrift-sportmedizin.de* (*N* = 44), and *emotion-technologies.de* (an association of bicycle dealerships) (*N* = 41). Upon comparing the first page of Google search results with the sources accessed, it becomes apparent that some are among the most visited sources, while others were barely accessed at all. Notably, the *emotion-technologies.de* page was accessed by a large number of participants, whereas the *emotion-ebikes.de* page was hardly accessed (*N* = 2) although both sites belong to the same association of bicycle dealerships. The website *ebike-gesundheit.de* appeared among the top results on *Google*, but received little attention (*N* = 13) despite being a reliable source affiliated with the *Hannover Medical School* and highly relevant to the topic at hand.

The logfiles of the visited websites (excluding search engine views) show a mean source credibility score of 3.29 (sd = 0.83) per participant and the entire range of sources from category 1 to 5 was consulted (Table 4). The distribution, with a skewness of 0.03 and negative kurtosis (−0.48), suggests a near-normal distribution with a slight tendency towards less trustful sources. Sources in reliability category 3, above all news pages, were accessed most frequently accounting for 31.02% of the total number of sources accessed and by the majority (84.67% of respondents visited at least one source in this category). In descending order, 26.95% of the sources received a score of 5 (visited by 48.91% of the participants), while 24.07% were assigned a score of 4, like specialist magazines. A total of 16.53% of the visited sources were assigned a score of 1 (visited by 56.20% of the participants) and the least frequent utilization was attributed to sources assigned a reliability score of 2, accounting for 1.44% of the overall distribution.

4.2 Citation of visited websites in the written answers

We focused not only on analyzing the sources used to gather information, but also on the number and credibility of the sources used as references in their arguments. This two-stage analysis is aimed to detect patterns in participants' responses with regards to source selection, evaluation, and integration. By comparing the visited sources with the referenced sources, our aim was to shed light on how participants make decisions to determine the relevance, credibility, and applicability of information in their responses to the task. A total of 491 of the 834 sources originally visited have been included in the

TABLE 4 Logfile data of the initial search process.

N*	Mean	sd*	Median	Trimmed	Mad*	Min*	Max*	Range	Skew*	Kurtosis	se*
Number of unique search engine requests											
137	3.71	2.69	3	3.3	1.48	0	18	17	1.79	5.01	0.23
Number of unique source views											
137	6.14	3.21	5	5.77	2.97	1	18	17	1.04	1.06	0.27
Source credibility											
137	3.29	0.83	3.25	3.29	0.86	1	5	4	0.03	−0.48	0.07

*N = sample size; sd = standard deviation; mad = average absolute deviation; min = minimum; max = maximum; skew = skewness; se = standard error.

TABLE 5 Characteristics of used sources in the written responses.

N*	Mean	sd*	Median	Trimmed	Mad*	Min*	Max*	Range	Skew*	Kurtosis	se*
Number of cited sources											
143	3.45	1.6	3	3.19	1.48	1	11	10	2.09	6.32	0.13
Source credibility											
143	3.25	0.89	3.09	3.24	0.85	1	5	4	0.09	−0.39	0.07
Difference in source credibility answer texts vs. logfiles from initial search											
136	−0.04	0.45	0	−0.07	0.29	−1	1.8	2.8	0.84	2.23	0.04

*N = sample size; sd = standard deviation; mad = average absolute deviation; min = minimum; max = maximum; skew = skewness; se = standard error.

answer texts. The sample size of 143 responses shows that respondents utilized 3.45 sources on average to support their arguments in response to the assessment question (Table 5). The standard deviation of the mean was 1.6, indicating moderate variability in the utilization of sources. The range of source references across responses is highlighted by the minimum and maximum values of 1 and 11. A rightward skew is emphasized by the skewness of 2.09, suggesting that a subset of responses exhibits a relatively higher number of cited sources. Furthermore, the kurtosis of 6.32 indicates a distribution with heavier tails and more extreme values than a normal distribution, implying a higher variability in sourcing practices.

To be able to compare the mean values of the sources visited and the sources cited, we subtracted the individual mean values of the sources from the response texts from the mean values from the research phase. The findings reveal a disparity in the mean quality of sources accessed during the research phase, versus those later cited in the responses. The negative average deviation of −0.04 (sd = 0.45), obtained from a sample of 136 cases, suggests a slight decline in the quality of selected citation sources when compared to the original research sources. The research and citation phases vary in source quality, as indicated by minimum and maximum differences of −1 and 1.8, respectively. From the difference values three groups of participants can be identified. The participants with positive scores ($N = 61$) excluded the less reliable sources from their research when selecting sources to support their argumentative claims. Among the subjects with no difference in mean score ($N = 33$), slightly less than a third did not exclude any sources from their research. The remainder kept the ratio of reliability scores consistent despite their selection. Subjects with negative values, on the other hand, excluded a larger number of sources with higher reliability scores.

Table 6 presents the sources visited by participants during the research and subsequently cited according to scores and categories. Each website was only counted once per respondent. The 'Sources from the Research Cited in the Answer Texts' column indicates the proportion of sources visited that were referenced in the response texts. It is evident that specialist magazines with a reliability score of 4 were cited most frequently in relation to the number of visits. In absolute figures, news pages (score = 3) were accessed and cited most frequently. For sources with a reliability score of 5, only a limited number were utilized for argumentation (see Scientific Research Institute, Government Bodies), as well as Google Books and Scholar and encyclopedias. In total, 58.87% of the visited sources were used as supporting evidence in the answer texts.

4.3 Task performance and domain differences

The COR scoring was done along 6 distinct scoring dimensions (see chapter 3). Table 7 shortly describes the results of the ratings in each of the dimensions. Dimensions 3 and 4 are targeted directly at the quality of the respondents searched sources and how accurately they evaluated them.

The partial scores within the assessment all tended to be on the higher end, indicating a potential ceiling effect which reduces the variance in the sample (Table 7). This was slightly less pronounced for the two dimensions on the quality of research and its evaluation accuracy (dimensions 3 and 4) and deliberation during the argument (dimension 5). Comparing between respondents of the three different domains, there are significant mean differences between each of them (*t*-test: $p < 0.05$ for medicine/teaching and law/teaching; $p < 0.10$ for medicine/law). Respondents of the law domain performed best on the task, with a total mean score of 3.51 and a low standard deviation of 0.36, which is likely enhanced by the subsample being the smallest. The law respondents performed strongly on the dimensions regarding quality of argumentation (dimensions 1, 2, 5 and 6), while also being only marginally weaker on the quality of research (dimensions 3 and 4) compared to respondents of medicine (Table 7). Respondents of

TABLE 6 Frequencies across formats and categories.

Score	Source category	Research sources	Cited sources	Sources from the research cited in the answer texts
5	Government bodies	11	2	18.18%
	Lecture notes	9	4	44.44%
	University database	15	3	20.00%
	Scientific journal	75	46	61.33%
	Scientific database	76	27	35.53%
	Scientific research institute	39	4	10.26%
4	Specialist magazine	148	128	86.49%
	Google books	2	0	0.00%
	Google scholar	50	1	2.00%
3	News page	209	142	67.79%
	Associations	50	34	68.00%
2	Encyclopedia	12	1	8.33%
1	Online shop	111	82	73.87%
	World wide web	27	15	55.55%
0	Search engines	496	2	0.40%
	Total	834 (1330)*	491	58.87%

*Total: 834 (1,330 including search engines).

TABLE 7 COR Scores along the 6 rating dimensions.

*N** = 143	Rating dimension	Mean (min: 0, max: 4)*	sd*
1	Concreteness	3.52	0.76
2	Comprehensibility	3.74	0.52
3	Quality of researched sources	3.32	0.70
4	Accuracy of source evaluation	3.25	0.84
5	Deliberation	2.97	1.00
6	Quality of answer	3.41	0.83
	Total score	3.33	0.59

Dimension	Medicine (*N* = 68)	Law (*N* = 30)	Teaching (*N* = 45)
1	3.47 (0.72)	3.72 (0.47)	3.47 (0.94)
2	3.75 (0.49)	3.89 (0.34)	3.64 (0.64)
3	3.46 (0.54)	3.41 (0.61)	3.05 (0.88)
4	3.39 (0.73)	3.36 (0.65)	2.99 (1.03)
5	2.87 (0.99)	3.24 (0.70)	2.96 (1.16)
6	3.44 (0.81)	3.69 (0.50)	3.18 (0.98)
Total	3.38 (0.49)	3.51 (0.36)	3.15 (0.80)

**N* = sample size; min = minimum; max = maximum; sd = standard deviation. Mean values (Min. = 0, Max. = 4) and standard deviations in ().

the subdomain teaching performed worst on all scoring dimensions, except for deliberation (dimension 5), while also featuring the highest level of standard deviations for each of the respective dimensions (Table 7). While the performance on the task was generally high, the distribution of performance is most scattered for teaching respondents and most cohesive within the law domain. Respondents of medicine performed particularly strongly on the research part of the task. This seems intuitive, since the task context is loosely health related, and few respondents of medicine were the only ones starting their initial search on a specialized database (i.e., Pubmed). The subsample was still outperformed by respondents of the law domain regarding the quality of argumentation and the written statements.

4.4 Saturation on the quantity of different sources

When looking at the returns of additional observed literature in terms of answering pre-defined assessment tasks, the question arises whether, and at what specific point, there are negative implications of citing too many articles. As the problem solving of the task contains an underlying time constraint, the amount of time allocated for filtering scientific sources is vastly limited in its own capacity. This potentially means that after a certain point, the quality of individual source selection and subsequent problem-solving will diminish with each additional piece of literature and might even yield negative returns. Thus, this regression model aims to find the optimal number of sources when solving the assessment task and after what exact number additional literature becomes counter effective.

The corresponding Ordinary-Least-Squares (OLS) regression model (Table 8) therefore uses a quadratic term of its main regressor, which captures the number of sources used. This allows the (OLS)-regression to display potential non-linearities in the effect of additional literature. In this specific case, as the first, linear and non-quadratic term takes a positive sign, while the quadratic term takes on negative values, the diminishing returns were clearly captured within the data frame. The dependent variable here is dimension 3 within the COR scoring scheme, depicting the quality of the researched sources. The regression model is also controlling for domain-affiliation with medicine being the reference group.

After taking the first derivative of the regression equation and calculating the optimal number of sources, the estimation suggests that on average after around 7 sources, every additional piece of literature lowers the expected score that participants achieve. Furthermore, the negative sign of the quadratic regressor also shows that the returns of additional literature themselves are diminishing from the start, even though there is a clear maximum effect around the number of 7 sources. To further enhance the robustness of this underlying model, a set of control variables as well as heteroscedasticity robust standard errors were deployed.

TABLE 8 Quadratic regression model.

Variables	Quality of researched sources
Number of sources	0.235** (0.102)
Number of sources squared	−0.018** (0.008)
Gender	0.064 (0.125)
Age	−0.017 (0.019)
Parental origin	−0.113 (0.099)
Native language	−0.00752 (0.110)
Final school grade	−0.189 (0.127)
Domain (law)	0.043 (0.141)
Domain (teaching)	−0.161 (0.183)
Constant	3.658*** (0.657)
Observations	143
R^2	0.158
Adjusted R^2	0.101
p-value	0.054

Standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Domain reference group: medicine.

4.5 Effects of the quality of research on COR performance

Table 9 covers the general effect of the mean quality of the cited sources, as measured by the website categorization, onto 4 different COR rating dimensions (Models 1 to 4) and their sum score (Model 5). The two dimensions of rated quality of researched sources (rating dimensions 3 and 4) were omitted. Hence, it only consists of dimensions 1,2,5 and 6 of the rating scheme, which were weighted appropriately. The reason for the omission of the rating dimensions 3 and 4 was that those dimensions are another measurement of the quality of the researched sources (as estimated by the raters), therefore being partially redundant to and dependent on the quality of research as estimated by the categorization.

In terms of the model, the estimation method of choice was a linear regression model with heteroscedasticity robust standard errors and a corresponding set of sociodemographic control variables. While the models themselves have mostly low explanatory power (R2), Model 4 shows a significant effect of the quality of sources onto the answer scores. An increase of 1 additional answer grade leads to an average increase of 0.173 points for the dimension 6 answer score, which is significant at the 95% significance level. In terms of the other dimensions, the quality of source selection seemingly has a positive effect for dimension 2 and the sum score, although the statistical power is insufficient to reject the corresponding null hypothesis. With respect to dimension 1 and 5, the estimations are close to 0 and not significant, which leads to the conclusion that there likely will be no observable effect.

5 Discussion

Regarding RQ 1, whether there are significant differences in the frequencies of different Website categories, Table 6 shows stark variety across formats and categories. Excluding search engines, the most used formats were news pages (category 3), specialist magazines (category 4) and online shops (category 1). Out of the total of used sources category 3 (31.02%), category 5 (26.95%) and category 4 (24.07%) were the most frequent, while category 1 (16.53%) and category 2 (1.44%) were rare. This emphasizes a slight focus on more adequate source materials during the assessment, while still showcasing a shortcoming in the terms of scientific materials incorporated (mostly found in categories 4 and 5). Differences in preferences between the categories can be seen, but there was no overwhelming majority for any single category. Most respondents (84.67%) visited a news page in the preparation of their written assignment. At the same time, while only roughly half as many different online shops were visited by the full sample, still 56.20% of respondents visited such a site during their research. This likely indicates that online shops were seen as relevant to the task by many, but the variety of different news outlets on the internet is larger than the variety of online shops for e-bikes, leading to more visits on the same sites (59 out of 111 visits were to the same online shop for e-bikes, which is also highly ranked on a Google search). Within category 5 the frequencies of websites also varied across the formats. Scientific journals and databases were considered more often than governmental bodies, university databases or research institutes. Overall differences in the frequencies of use can be seen, but they also vary within specific formats in the same categories.

The differences in frequencies of use translate into the frequencies to sources cited in the written arguments to a large degree. Table 6 shows that the most visited formats also have some of the highest rates of citation (86.49% for specialist magazines). In turn, the rate of discarding found sources was often low, even for the less adequate categories (73.87% citation rate for online shops). Some outliers are scientific databases, which were close to the top for visited formats within category 5, but only had a citation rate of 35.53%. This was even more pronounced for scientific research institutes, which were visited less frequently, and only had a citation rate of 10.26%. The rate of incorporating sources from the highest category 5 was generally low compared to categories 1, 3 and 4. This suggests a general preference for more easily accessible information during the assessment, which makes sense regarding the time constraints, but might be detrimental to the quality of the written statements. Regarding RQ 2, this indicates a higher rate of discarding quality sources, even after those were initially accessed, with a propensity for including lower quality sources at higher rates. Since 125 of the respondents started their initial search using the Google search engine, with 111 of them using some variation of the words "e-bike" and "health" (as directly taken from the task prompt) the differences in source selection are unlikely to be explained by this first step in the approach. The common preferences for the top results for this search query are in line with the higher rates of discarding quality sources in Table 6. *Ebike-gesundheit.de*, a reliable website by the *Hannover Medical School*, and a top result for the common search query was rarely visited ($N = 16$), while websites of *ebike* dealerships and news magazines were frequented more often.

The quantity of researched sources on the individual level seemed to have an impact on the rated quality of the research in general (RQ 3). Some of this effect can be attributed to the rating scheme as it was a necessity to have multiple sources for a deliberate and informed argument. Which was only an issue when respondents built their argumentation on a single source. At the same time, the non-linearity

TABLE 9 Linear regression models on the different dimensions of COR score.

Variables	Model 1 concreteness	Model 2 comprehensibility	Model 3 deliberation	Model 4 quality of argument	Model 5 full COR score
Quality of used sources	−0.012 (0.084)	0.097** (0.049)	0.030 (0.113)	0.188** (0.076)	0.303 (0.263)
Gender	−0.086 (0.149)	−0.081 (0.102)	0.101 (0.173)	−0.162 (0.157)	−0.228 (0.476)
Age	−0.028 (0.021)	−0.013 (0.015)	−0.033 (0.025)	−0.029 (0.023)	−0.103 (0.070)
Native language	−0.156 (0.138)	−0.042 (0.067)	−0.271** (0.129)	−0.035 (0.111)	−0.503 (0.383)
Final school grade	−0.041 (0.155)	−0.107 (0.092)	−0.234 (0.181)	−0.093 (0.149)	−0.475 (0.510)
Domain (law)	0.247 (0.171)	0.261** (0.110)	0.527** (0.229)	0.398** (0.181)	1.432** (0.566)
Domain (teaching)	0.133 (0.226)	0.106 (0.160)	0.363 (0.261)	0.072 (0.242)	0.674 (0.796)
Constant	4.599*** (0.703)	4.054*** (0.488)	4.149*** (0.828)	3.913*** (0.695)	16.71*** (2.302)
Observations	143	143	143	143	143
R^2	0.058	0.084	0.088	0.112	0.094
Adjusted R^2	0.009	0.037	0.041	0.066	0.047
p-value	0.362	0.097	0.114	0.021	0.086

Standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$. Domain reference group: medicine.

of more sources not leading to more information and a better argumentation was not by design. The quadratic regression model (Table 8) showcases a saturation point during the task assessment around 7 different incorporated sources. Past that point, additional research seemed increasingly detrimental to the performance on the task. Within the analysis it is unclear whether the reason for this is time management on the task or issues with cognitive load while adding up information from an increasing number of sources, which then must be synthesized into the argument. The regression model shows a steep increase in the quality of research for the initial extra sources with diminishing returns for each one added. Expecting a level of redundancy between different sources, this finding is within expectations.

The impact of the quality of the researched sources on the quality of the written arguments (RQ 4) was covered in the 5 linear regression models (Table 9). Model 1 (covering concreteness of the answer) and model 2 (comprehensibility of the argument) were mostly added for controlling parts of the analysis. Both facets can rather be seen as individual writing ability and would only be impacted by the quality of the researched sources if the respondents copied text passages from the sources directly. In a lesser sense, this is also true for model 3 (covering the deliberation aspect within the argument) which is dependent on different perspectives within the researched sources, but at the same time an argumentative competency, which the respondents would have to apply during the written assignment. Model 4 covered the quality of the argument, which is also tied to general reasoning competencies. Raters were asked to give little emphasis to linguistic markers within the written responses and focus on the number and quality of single arguments within the entire argumentation and the overall argumentative synthesis. This should have lessened the impact of differences in individual writing capabilities and enforce an emphasis on the content which respondents drew from their research. Model 4 shows a significant effect of the quality of the researched sources along the categorization of websites on the quality of the written argument. The model parameters are at a *p*-value of 0.02 with an adjusted R^2 of 0.07 and significant, even considering the rather low sample size. In both model 4 and 5 (the model for the aggregated COR score) increased age of respondents corresponded to a lower quality of the written arguments. This effect was not found within models 1 through 3, which covered more general writing and argumentation capabilities, indicating a relation to the online research aspects of the assessment. With the low effect strength and the rather small sample this result is mostly tentative, especially since the assessment covered a specific age group of young professionals at the start of their careers with only a few outliers with increased age.

There is evidence of meaningful variance in the source selection along categories within the sample, but the data set does not seem robust enough to make clear claims regarding the influence of the quality of researched and cited websites on the quality of the written arguments.

6 Conclusion, limitations and outlook

In conclusion, respondents in the assessment shared a common initial response on how they started their search. The overall performance on the task was high, with law respondents performing best on average and teaching respondents performing worst (Table 7). Both the selection of initial websites for the top results within the most common search query and the higher rates of discarding the most trustworthy sources (Table 6) indicate a tendency to rely on less trustworthy sources on average. This is also featured in the small, but overall negative, change in source credibility from researched sources to sources used and cited within the written statements (Table 5). Additional sources helped the overall source quality until a saturation point of 7 sources, at which adding more led to an overall decline (Table 8). The quality of incorporated sources was most significant for the quality of the arguments within the written statements and to a slightly lesser degree to the overall performance (Table 9).

A limiting factor in this analysis is the framing of the website categorization with primary emphasis on the publishing entities of the websites. Qualitative content analysis of the websites would add deeper insights into the information the respondents encountered during their visits but is not feasible for the 834 different websites. Methods of quantitative content analysis would be a way to add information for each of the sources, since theoretically an online shop or encyclopedia could provide relevant and reliable information on the given topic. Parameters such as factuality, neutrality of tone, visual aspects etc. could provide a more complete scoring rubric for the websites themselves. Exemplary cases could also be drawn from the sample and analyzed in a qualitative manner. This is an undergoing angle which is currently being followed up on in which reconstructive hermeneutics and narrative analysis are used to gain a detailed understanding on how respondents interpreted single passages within sources and how they incorporated those into their later written arguments.

Another issue, specifically for the quantitative analysis provided here, was the sample size ($N = 143$) within the *BRIDGE* project. The assessments are time intensive and effortful on part of the respondents, which naturally leads to challenges in respondent acquisition. The lack of sample size was also exacerbated through the high performance on all COR dimensions during the first survey wave, which led to a ceiling effect and rather low variance on the scoring. This was possibly a result of a task not challenging enough, since it did not repeat itself in a later survey with a different task context. In a larger scale follow up project we are preparing new tasks and are aiming for a much larger longitudinal sample, which will provide a more robust data set. The tentative findings here provide an early step for future data collection and analysis in the years to come. They highlight differences in search behavior, source selection and the synthetization of information into an argument given a specific task. They already indicate that differences in source selection are to be expected during a task with open research possibilities and that they are meaningful for the information that respondents can draw from them and the quality of arguments that they can build on them, while more sources are not always necessarily better.

Data availability statement

The data supporting the conclusions of this article will be made available by the authors, upon reasonable request.

Ethics statement

The studies involving humans were approved by Gemeinsame Ethikkommission Wirtschaftswissenschaften der Goethe-Universität Frankfurt und der Johannes Gutenberg-Universität Mainz. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

DB: Writing – original draft, Writing – review & editing. LM: Writing – original draft, Writing – review & editing. LT: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study is part of the Bridge project (Funding no. 01JD1906), which was funded by the German Ministry of Education and Research (BMBF).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andersen, K., Bjarnøe, C., Albæk, E., and De Vreese, C. H. (2016). How news type matters. *J. Media Psychol.* 28, 111–122. doi: 10.1027/1864-1105/a000201

Ashley, S., Maksl, A., and Craft, S. (2013). Developing a news media literacy scale. *Journal. Mass Commun. Educ.* 68, 7–21. doi: 10.1177/1077695812469802

Bendel, O. (2018). *Soziale Medien. Gabler Wirschaftslexikon*. Wiesbaden: Springer.

Braunheim, D., Zlatkin-Troitschanskaia, O., and Nagel, M. (2023). Erfassung und Förderung von Kompetenzen zum kritischen Umgang mit Online-Informationen bei Rechtsreferendarinnen und -referendaren. *Zeitschrift für Didaktik der Rechtswissenschaft* 10, 145–167. doi: 10.5771/2196-7261-2023-2-145

Brooks, C. (2016). *ECAR study of students and information technology*. Louisville, KY: ECAR.

Brox, H. (2012). The elephant in the room: a place for Wikipedia in higher education? Nordlit. *Tidsskrift I Litteratur Og Kultur* 16:143. doi: 10.7557/13.2377

Ciampaglia, G. L. (2018). "The digital misinformation pipeline" in *Positive Learning in the Age of Information*, O. Zlatkin-Troitschanskaia, G. Wittum, and A. Dengel Eds. (Wiesbaden: Springer), 413–421.

Ćurković, M., and Košec, A. (2018). Bubble effect: including internet search engines in systematic reviews introduces selection bias and impedes scientific reproducibility. *BMC Med. Res. Methodol.* 18:130. doi: 10.1186/s12874-018-0599-2

Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. (2003). How do users evaluate the credibility of web sites? In Proceedings of the 2003 conference on designing for user experiences, ed J. Arnowitz (New York, NY: ACM), 1–15.

Gasser, U., Cortesi, S., Malik, M., and Lee, A. (2012). *Youth and digital media: From credibility to information quality*. Cambridge: The Berkman Center for Internet and Society.

Goldman, S. R., Braasch, J. L. G., Wiley, J., Graesser, A. C., and Brodowinska, K. (2012). Comprehending and learning from internet sources: processing patterns of better and poorer learners. *Read. Res. Q.* 47, 356–381. doi: 10.1002/rrq.027

Groeling, T. (2013). Media Bias by the numbers: challenges and opportunities in the empirical study of partisan news. *Annu. Rev. Polit. Sci.* 16, 129–151. doi: 10.1146/annurev-polisci-040811-115123

Guess, A. M., Nagler, J., and Tucker, J. A. (2019). Less than you think: prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* 5:eaau4586. doi: 10.1126/sciadv.aau4586

Harrison, N., and Luckett, K. (2019). Experts, knowledge and criticality in the age of 'alternative facts': re-examining the contribution of higher education. *Teach. High. Educ.* 24, 259–271. doi: 10.1080/13562517.2019.1578577

Hasebrink, U. (2016). Meinungsbildung und Kontrolle der Medien. Available at: https://www.bpb.de/gesellschaft/medien-und-sport/medienpolitik/172240/meinungsbildung-und-kontrolle-der-medien?p=all (Accessed December 01, 2023).

Head, A. J., and Eisenberg, M. B. (2009). Lessons learned: how college students seek information in the digital age. *Soc. Sci. Res. Netw.* doi: 10.2139/ssrn.2281478

Hesse, W. (2018). Die Glaubwürdigkeit der Wikipedia. *Inform. Wissenschaft und Praxis* 69, 171–181. doi: 10.1515/iwp-2018-0015

Kelly, Y., Zilanawala, A., Booker, C., and Sacker, A. (2018). Social media use and adolescent mental health: findings from the UK millennium cohort study. *EClinicalMedicine* 6, 59–68. doi: 10.1016/j.eclinm.2018.12.005

Kiili, C., Bråten, I., Strømsø, H. I., Hagerman, M. S., Räikkönen, E., and Jyrkiäinen, A. (2022). Adolescents' credibility justifications when evaluating online texts. *Educ. Inf. Technol.* 27, 7421–7450. doi: 10.1007/s10639-022-10907-x

Kuhn, S., Müller, N., Kirchgässner, E., Ulzheimer, L., and Deutsch, K. (2020). Digital skills for medical students—qualitative evaluation of the curriculum 4.0 "medicine in the digital age". *GMS J. Med. Educ.* 37:Doc60. doi: 10.3205/zma001353

Leif, T. (2001). Macht ohne Verantwortung. Der wuchernde Einfluss der Medien und Desinteresse der Gesellschaft. *Aus Politik und Zeitgeschichte* 41, 6–9.

Lucassen, T., Muilwijk, R., Noordzij, M. L., and Schraagen, J. M. (2013). Topic familiarity and information skills in online credibility evaluation. *J. Assoc. Inf. Sci. Technol.* 64, 254–264. doi: 10.1002/asi.22743

Maurer, M., Quiring, O., and Schemer, C. (2018). "Media effects on positive and negative learning" in *Positive learning in the age of information*. eds. O. Zlatkin-Troitschanskaia, G. Wittum and A. Dengel (Wiesbaden: Springer VS)

Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., and Jitomirski, J. (2020). "Positive and negative media effects on university students' learning: preliminary findings and a research program" in *Frontiers and advances in positive learning in the age of InformaTiOn (PLATO)*. ed. O. Zlatkin-Troitschanskaia (Cham: Springer)

McGrew, S. (2021). Skipping the source and checking the contents: an in-depth look at students' approaches to web evaluation. *Comput. Sch.* 38, 75–97. doi: 10.1080/07380569.2021.1912541

McGrew, S., Ortega, T., Breakstone, J., and Wineburg, S. (2017). The challenge That's bigger than fake news: civic reasoning in a social media environment, *Am. Educ.* 41, 4–9.

Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: tutorial. *J. Med. Internet Res.* 25:e50638. doi: 10.2196/50638

Mislevy, R. J., Haertel, G. D., Riconscente, M., Rutstein, D. W., and Ziker, C. (2017). "Evidence-Centered assessment design" in *Assessing model-based reasoning using evidence-centered design (Cham: Springer)*, 19–24.

Molerov, D., Zlatkin-Troitschanskaia, O., Nagel, M.-T., Brückner, S., Schmidt, S., and Shavelson, R. J. (2020). Assessing university students' critical online reasoning ability: a conceptual and assessment framework with preliminary evidence. *Front. Educ.* 5:577843. doi: 10.3389/feduc.2020.577843

Müller-Brehm, J., Otto, P., and Puntschuh, M. (2020). Kommunikation, Medien und die öffentliche Debatte. *Informationen zur politischen Bildung* 344, 8–15.

Nagel, M.-T., Zlatkin-Troitschanskaia, O., and Fischer, J. (2022). Validation of newly developed tasks for the assessment of generic critical online reasoning (COR) of university students and graduates. *Front. Educ.* 7:914857. doi: 10.3389/feduc.2022.914857

Nagel, M., Zlatkin-Troitschanskaia, O., Schmidt, S., and Beck, K. (2020). "Performance assessment of generic and domain-specific skills in higher education economics" in *Student learning in German higher education: Innovative measurement approaches and research results*. eds. O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper and C. Lautenbach (Wiesbaden: Springer), 281–299.

Osborne, J., Pimentel, D., Alberts, B., Allchin, D., Barzilai, S., Bergstrom, C., et al. (2022). *Science education in an age of misinformation*. Stanford University, Stanford, CA.

Schröder, W. (1994). "Erkenntnisgewinnung. Hypothesenbildung und Statistik" in *Neuere statistische Verfahren und Modellbildung in der Geoökologie*. eds. W. Schröder, L. Vetter and O. Fränzle (Wiesbaden: Vieweg Teubner Verlag), 1–15.

Schultz, T., Jackob, N., Ziegele, M., Quiring, O., and Schemer, C. (2017). Erosion des Vertrauens zwischen Medien und Publikum? *Media Perspektiven* 5, 246–259.

Schweiger, W. (2007). *Theorien der Mediennutzung*. Wiesbaden: Springer.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Mariño, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309

Speicher, M., Both, A., and Gaedke, M. (2015). SOS: Does your search engine results page (SERP) need help? In proceedings of the 33rd annual ACM conference on human factors in computing systems. New York: Association for Computer Machinery.

Steffens, Y., Schmitt, I. L., and Aßmann, S. (2017). *Mediennutzung Studierender: Über den Umgang mit Medien in hochschulischen Kontexten. Systematisches Review nationaler und internationaler Studien zur Mediennutzung Studierender*. Köln: Universität zu Köln, Humanwissenschaftliche Fakultät, Department Erziehungs- und Sozialwissenschaften 2017, 56 S.—URN: urn:nbn:de:0111-pedocs-154685

Verplanken, B., and Orbell, S. (2022). Attitudes, habits, and behavior change. *Annu. Rev. Psychol.* 73, 327–352. doi: 10.1146/annurev-psych-020821-011744

Voss, J. (2005). *Measuring Wikipedia*. E-prints in library & information science. Available at: http://eprints.rclis.org/6207/ (Accessed December 01, 2023).

Wagschal, U., Jäckle, S., Hildebrandt, A., and Trüdinger, E.-M. (2020). *Politikpanel Deutschland – Ausgewählte Ergebnisse der zweiten Welle einer Bevölkerungsumfrage zu den Auswirkungen des Corona-Virus*. Freiburg: Seminar für Wissenschaftliche Politik.

Walton, D. (2006). *Fundamentals of critical argumentation. Critical reasoning and argumentation*. Cambridge: Cambridge UP.

Wineburg, S., and McGrew, S. (2017). Lateral Reading: Reading less and learning more when evaluating digital information. In Stanford history education group working paper no. 2017-A1. Available at: https://ssrn.com/abstract=3048994 (Accessed November 27, 2023).

Wolfsfeld, G., Yarchi, M., and Samuel-Azran, T. (2016). Political information repertoires and political participation. *New Media Soc.* 18, 2096–2115. doi: 10.1177/1461444815580413

Zhang, X. (2022). Analysis on how algorithms reshape People's existence, cognition, and relationships. *Adv. Soc. Sci. Educ. Hum. Res.* 664, 597–601. doi: 10.2991/assehr.k.220504.109

Zlatkin-Troitschanskaia, O., Beck, K., Fischer, J., Braunheim, D., Schmidt, S., and Shavelson, R. J. (2020). The role of students' beliefs when critically reasoning from multiple contradictory sources of information in performance assessments. *Front. Psychol.* 11:2192. doi: 10.3389/fpsyg.2020.02192

Zlatkin-Troitschanskaia, O., Brückner, S., Nagel, M.-T., Bültmann, A.-K., Fischer, J., Schmidt, S., et al. (2021b). Performance assessment and digital training framework for young professionals' generic and domain-specific online reasoning in law, medicine, and teacher practice. *J. Supranat. Pol. Educ.* 13, 9–36. doi: 10.15366/jospoe2021.13.001

Zlatkin-Troitschanskaia, O., Hartig, J., Goldhammer, F., and Krstev, J. (2021a). Students' online information use and learning progress in higher education – a critical literature review. *Stud. High. Educ.* 46, 1996–2021. doi: 10.1080/03075079.2021.1953336