Integrating measures of replicability into scholarly search: Challenges and opportunities

Chuhao Wu
The Pennsylvania State University
State College, Pennsylvania, USA
cjw6297@psu.edu

Tatiana Chakravorti
The Pennsylvania State University
State College, Pennsylvania, USA
tfc5416@psu.edu

John M. Carroll
The Pennsylvania State University
State College, Pennsylvania, USA
jmc56@psu.edu

Sarah M. Rajtmajer
The Pennsylvania State University
State College, Pennsylvania, USA
smr48@psu.edu

ABSTRACT

Challenges to reproducibility and replicability have gained widespread attention, driven by large replication projects with lukewarm success rates. A nascent work has emerged developing algorithms to estimate the replicability of published findings. The current study explores ways in which AI-enabled signals of confidence in research might be integrated into the literature search. We interview 17 PhD researchers about their current processes for literature search and ask them to provide feedback on a replicability estimation tool. Our findings suggest that participants tend to confuse replicability with generalizability and related concepts. Information about replicability can support researchers throughout the research design processes. However, the use of AI estimation is debatable due to the lack of explainability and transparency. The ethical implications of AI-enabled confidence assessment must be further studied before such tools could be widely accepted. We discuss implications for the design of technological tools to support scholarly activities and advance replicability.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

literature search, replicability, reproducibility, explainable artificial intelligence

ACM Reference Format:
Chuhao Wu, Tatiana Chakravorti, John M. Carroll, and Sarah M. Rajtmajer. 2024. Integrating measures of replicability into scholarly search: Challenges and opportunities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3643043

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3643043

1 INTRODUCTION

Isaac Newton famously said, "If I can see further, it is because I am standing on the shoulders of giants." The principle that scientific inquiry is built upon insights derived from an existing body of work is central to the scientific method [32], and literature review has been proposed as the first and foundational step in any research project [11, 56]. Despite its important role, for many years, the lack of systematization challenges validity and undermines the reproducibility of findings based on meta-analyses [60, 71]. Several recommendations and guidelines have been proposed to help scholar report their literature search and review in a clear and reproducible manner. For example, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [92] and its literature search extension [102] have provided comprehensive checklists for reporting systematic reviews. However, besides methodological issues, there is also a technological component that needs to be addressed [14], especially tools that can assess the quality of the results from search engines [64]. Systems for information retrieval, such as search engines, mainly rely on keyword matching, yet scientific article search ought to consider other factors such as citation patterns, publication venues, and disciplines.

To tackle this problem, a number of information retrieval models and paper recommenders have been developed to incorporate citation context [79] and scholars' preferences [119] in their algorithm, and leverage text-mining to automate manual tasks and save time for researchers [51]. However, a comparative analysis of existing tools in the literature and quality metrics is needed to improve the usability, ease of use, reliability, performance, and support for future designs [27]. Al-Zubidy and Carver [1] show that researchers could have diverse needs for literature tools and there can be disciplinary differences. Yet, there is a lack of understanding of user experiences and challenges with these tools. For example, many researchers rely on digital databases and search engines such as Google Scholar for their literature search, but these digital libraries have their different features and may limit user experience and fail to meet user needs [107]. As most of them are proprietary, we have little information about how users interact with and assess the search results. Understanding current practices around literature search can significantly contribute to the human-centered design of tools to support scholarly activities.

A primary challenge to the search and assembly of existing research is the massive and increasing number of publications across

virtually all fields of study. The rise of predatory journals is flooding the literature with low-quality work [12, 75, 106], heightening the stakes for effective search and, in parallel, motivating the inclusion of indicators of credibility into search outputs. Generative AI will likely exacerbate this problem in the coming years because of the potential misuse of AI in writing [63]. Even putting aside predatory journals and AI-generated papers, broader concerns about confidence in published work have come to the foreground over the past decade. Across the natural and social sciences, researchers have recognized that many studies are difficult or impossible to replicate, e.g., [50, 85]. A number of contributing factors have been proposed including publication bias favoring novel, affirmative results, manipulative statistical analyses, e.g., p-hacking, and lack of transparency when describing research methods [10, 34, 88].

In response, the open science and science of science communities have made significant efforts to improve confidence in research, calling for changes to both policy and practice [52, 84]. Likewise, technologies are emerging to automate aspects of confidence assessment for published findings. There have been a number of recent efforts, for example, using supervised learning over features extracted from papers' text and metadata to predict outcomes study replications [3, 94, 100, 116, 118].

Yet, ways in which open science indicators, e.g., preregistration, open materials, as well as scores and explanations returned by automated approaches for confidence assessment, could be integrated into scholarly search and literature review, remain unexplored. To address this gap, we conduct semi-structured interviews with PhD students and postdoctoral researchers from social & behavioral science departments in the U.S. We ask participants about their literature search and review practices, and about their concerns with respect to the credibility of work in their area. We also demonstrate an AI-driven replicability estimation tool to understand whether and how such a tool could assist them in their workflows. Specifically, our work is motivated by three primary research questions:

- RQ1: What are researchers' current approaches to literature search and review?
- RQ2: What challenges do researchers encounter during literature search and review?
- RQ3: How should signals of credibility be integrated into researchers' literature search and review?

It should be noted that there can be significant differences in the literature search process and perceptions of reproducibility across disciplines [53, 54, 86]. Despite some common practices used regardless of discipline, the current study does not seek to make statements that can be generalized without further validation. As a context, this study only focuses on social and behavioral science research, with all participants and study materials coming from this field. With this qualification, the study makes several important contributions. First, we connect with and contribute to studies on design that support scholarly search and management by empirically documenting researchers' strategies for literature search, review, and evaluation, highlighting the need for more flexible and intelligent approaches. Second, we link researchers' literature review practices with their perceptions of credibility, demonstrating opportunities for the inclusion of reproducibility and replicability metrics. Finally, we explore scholars' perceptions of an AI-driven replicability estimation tool and how to integrate the replication prediction into the literature review. We believe that the findings and discussions can motivate the research community to further ponder the design implications and ethical considerations for systems enabling automated assessment of confidence in published findings.

2 RELATED WORK

Our work builds upon and brings together literature in the areas of scholarly search and scientific integrity and its assessment. In this section, we presented some prior progress made for augmenting literature search and reviewing, promoting research reproducibility and replicability, and assessing these dimensions of the quality of publications.

2.1 Literature search and review

Effective literature search and review is foundational for research in all evidence-based fields, yet finding, assembling, and contextualizing relevant work is a painstaking and primarily *ad hoc* manual process [22]. Researchers in various disciplines have attempted to organize protocols to guide effective and efficient literature search. Through the analysis of nine guidance documents on a systematic review of topics in the social sciences, Cooper et al. [38] summarize eight key stages of literature search, starting from "who should literature search" and ending with "managing references and reporting the search process". Other researchers have demonstrated that apparently slight differences in any stage of a literature search can lead to distinct results. For example, comparing search strategies used in 152 applied psychology systematic reviews, Harari et al. [60] highlight the important impact of database selection. They find that publisher databases tended to perform poorly with low yields, while Google Scholar returned the largest proportion of articles yet with poor precision. When using these databases, search queries also play an important role and sophisticated techniques may be required for users to construct an effective search query [8, 20].

Recognizing the challenges inherent to literature search and opportunities for technologies to provide assistance, Human-Computer Interaction (HCI) research has extensively explored ideas for supporting scholars in this work. Choe et al. [30] design an interactive system (Papers101) to accelerate the discovery of literature by recommending relevant keywords and ranking papers based on keyword similarity, publication year, citation count, and other metrics. Action Science Explorer (ASE) [46] and LitSense [109] support reference management and help researchers develop a holistic understanding of the literature through network visualizations, e.g., topic graphs and citation networks. While these papers have demonstrated the preliminary usability of their tools, evaluations were conducted with a limited, fixed set of publications. Therefore, they focused more on the effectiveness of visualizations and metrics rather than the whole search experience. Another line of work focuses on scaffolding scholarly reading and evaluation processes. Prior studies have developed novel interfaces to facilitate the understanding of technical terms and symbols used throughout a paper by providing tooltips and auto-generated glossaries [61], prioritizing and contextualizing inline citations based on researchers'

reading history [28], and even augmenting documents with auto-generated paragraph headings and filtering redundant information [93]. With the advancement of natural language processing (NLP) techniques, we certainly expect continued innovation in support of the literature review.

With the widespread popularity of scholarly search engines [111], understanding search practices is foundational to the design and development of these tools. Maloney and Conrad [81] summarize four types of scholarly search strategies: *Exploring*, *Finding*, *Refinding*, and *Serendipity* based on whether users know what they are seeking and where to find it. They emphasize that information providers should particularly support serendipitous discovery. Through 368 survey responses, Soufan et al. [108] find that the literature search task is often identified as an exploratory search task characterized by unfamiliarity with the domain and dynamic information needs. As digital communication methods become diverse, literature search may also go beyond digital libraries and involve social resources and non-library technologies [67]. It is clear that with the rapidly increasing volume of information, users will need more customized support from search tools since the optimal search strategy often depends on the task [6]. It is critical for search tools to organize results for the optimal usefulness of the information to the users [112]. Therefore, understanding how users assess search outcomes plays an important role in improving the design, as does understanding the ways in which current systems satisfy or fail to meet researchers' needs.

2.2 The replication crisis

Another important object of the study is to explore how signals of replicability could be integrated into researchers' literature search and review process. Concerns have been raised over the past decade about the reproducibility, replicability, and robustness of published findings in the social sciences and beyond have gained significant attention, as large-scale efforts to replicate high-profile empirical studies have yielded low success rate [23, 34]. This revelation has been nicknamed the *replication crisis*, or equivalently the *reproducibility crisis*. There has been some ambiguity around terms; throughout this work, we adopt definitions from [90, 91, 95]. Namely, *reproducibility* refers to computational repeatability – obtaining consistent computational results using the same data, methods, code, and conditions of analysis; *replicability* means obtaining consistent results on a new dataset using similar methods. These two terms have been frequently mentioned together or even used interchangeably due to their close relatedness [87, 96]. *Robustness* is obtaining consistent results on the same data using a different analytical approach. *Generalizability* refers to the extent that results hold in other contexts or populations different from the original. Each of these terms captures some subset of qualities we might consider important for contextualizing confidence in a given claim or finding.

Baker [10] showed there was a crisis of reproducibility due to selective reporting, low statistical power, and other inappropriate practices. Other studies have confirmed that questionable research practices (QRPs) such as p-hacking and HARKing (constructing new hypotheses after the results are known) can yield false-positive results [62], and journals are less likely to publish replication studies, especially those contradicting prior findings [80]. Recognizing this replication crisis, the scientific community has proposed multiple initiatives to address these issues. For example, making study materials such as data and analysis codes openly available allows others to validate published results and identify errors more easily [69]. Platforms such as Open Science Framework (OSF) [52] also facilitated data sharing and promoted reproducible practices across the community.

While researchers' conscientiousness in maintaining good practices is critical, incentive and policy changes are also necessary. Despite exciting progress in both promoting and assessing reproducibility and replicability, to the best of our knowledge, there has been limited study of how these considerations should be integrated into scholarly search and review. Specifically, it is unclear whether researchers look for signals of reproducibility during their search processes, and to what extent this criterion may impact their selection of literature. Considering the significant role of literature review in the research workflow, the current study explores this issue by focusing on whether and how researchers would include automated assessments of the replicability of the findings in their workflows.

2.3 Assessment of reproducibility and replicability

A number of large-scale replication projects have set out to assess the state of reproducibility and replicability in specific domains. These have included efforts targeting psychology [35], economics [13, 24], experimental philosophy [39], social and behavioral sciences more broadly [25], cancer biology [50], neuroscience [18], computer science [36], and machine learning [99]. Generally speaking, these projects require massive investment in time and resources and, therefore necessarily limited in scale. The most common approach to assessment of reproducibility and replicability, short of actually repeating a study, is to apply a checklist of evaluation criteria to a paper in question. For example, to estimate the state of reproducibility of AI research, Gundersen and Kjensmo [58] use a list of binary variables (true or false) indicating how well the method, data, and experiment are documented in a paper. Similarly, for web measurement studies, Demir et al. [43] specify 18 criteria with three levels (omit, undocumented, and satisfy), describing the dataset, experiment design, and evaluation aspects of a paper. However, these checklists still require manual examination of the full paper. In recent years, researchers have begun to explore opportunities for automated estimation of replicability, e.g., [3, 94, 118] using supervised learning over features extracted from text and metadata of ground truth replication project outcomes. While these applications are promising, there are several important issues known in machine learning and AI-empowered tools that might impact user experience. As Chen and Eickhoff [29] suggested, there is an increasing need for search engines, recommenders, and similar systems to provide transparent and explainable results to users. Existing research on explainable AI (XAI) emphasized that systems should enable users to see how inputs and outputs are mathematically interlinked, using an intrinsic method that generates a human-readable explanation for the model's decision [2]. Some researchers pointed

out that previous approaches to XAI have been algorithm-centered, and understanding end-users' explainability needs and the socio-organizational context should be incorporated in the development of XAI [48, 70]. In addition, for decision-making AI systems, robust representation of uncertainty also plays an essential role during human-AI collaboration [26, 98]. Therefore, examining user perception of explainability, transparency, and uncertainty is critical for designing automated estimation of replicability.

3 AI PREDICTION MARKET FOR RESEARCH CONFIDENCE ASSESSMENT

The current study takes advantage of an AI-driven tool for confidence scoring the replicability of published findings [100]. The inner workings of the AI are powered by a synthetic prediction market, i.e., a market where trader bots are asked to buy and sell outcomes of a notional replication study for a given finding. The feature extraction framework for replicability prediction is designed to extract a comprehensive range of features from published papers and their metadata, spanning five distinct categories: bibliometric, venue-related, author-related, statistical, and semantic information. Currently, it extracts a total of 41 features, encompassing various elements such as p-values, sample size, the number of authors, and the acknowledgment of funding. This broad array of features provides a detailed perspective on the papers, aiding in the assessment of their replicability. In the market, agents are initialized with a set of extracted features that represent the paper. The base model of the artificial binary prediction market is described in [89]. In the artificial prediction markets, AI agents are trained using a genetic algorithm, where agents that generate profits are retained while those failing to do so are removed from the agent pool. During this process, the five most profitable agents are identified, from which the top three are selected for mutation and crossover. These agents are endowed with a predetermined amount of initial cash, enabling them to purchase assets. The logic governing these purchases is defined by a sigmoid transformation function, with the stipulation that agents may acquire only one share of an asset at a time. The genetic algorithm's objective function aims to minimize the root mean square error (RMSE) of the estimated score. The aggregate final market price serves as the score value for each paper, essentially representing the market's final price. Various hyperparameters, such as the liquidity constant, market duration, initial cash, number of generations, number of agents per market, agent inter-arrival rate, and run time, play a pivotal role in controlling market performance. The careful selection of these hyperparameters is crucial for optimal market functionality [101, 116]. Since agents are fully algorithmic, similar to neurons in a neural network, we do not need to have ground truth for test data points (research findings). Agents are, however, trained on 446 ground truth replication outcomes.

4 ETHICAL CONSIDERATIONS

The prediction tool described above is fully developed by prior research [100]. It was obtained with full consent from the corresponding author. The current study does not make any modifications or design decisions for the system. During our interviews, we only demonstrate this tool to participants and collect feedback. Our overarching aim is to examine whether and how technological assistance (whether via the tool or some other means) can facilitate researchers' estimation of confidence in published work for the purposes of literature review and synthesis. Nonetheless, we need to point out some important ethical considerations for such a tool. First of all, regarding the use of author-related information, the tool only utilizes information presented in a publication (e.g., affiliations). It does not collect or use any extra demographic information about the authors (e.g., age and gender). Second, we acknowledge that researchers may not be comfortable with AI-empowered replicability estimation, as AI predictions are often based on correlations instead of causality [55]. Therefore, to what extent these tools accurately estimate replicability is still uncertain and requires thorough examination. In addition, the sole focus on replicability does not necessarily improve the quality of research [44]. Therefore, the use of AI-empowered replicability estimation should be treated with extra caution. For this exploratory study, we aim to take a neutral position without enforcing certain attitudes on our participants. Instead, we investigate participants' perceptions of the issue and openly invite them to comment on the tool's negative and positive aspects.

5 METHODS

5.1 Recruitment and Data Collection

The study took place on the main campus of an R1 university in the mid-Atlantic region of the United States. Participants were selected using purposive sampling. Recruitment emails were sent to PhD students currently enrolled in, or who had recently graduated from, social and behavioral science (SBS) disciplines at the university—such as psychology, sociology, and anthropology programs. This specific focus was chosen because discussions around the replicability crisis have been particularly prominent in SBS fields, and the AI prototype was trained on publications from these disciplines. However, it should be noted that the resulting findings will, therefore, be confined by the characteristics of the chosen population and should not generalized to the other disciplines and cultures without further validation.

Participants' email addresses were sourced from the university's directory, and the study received approval from the institution's Institutional Review Board (IRB). Data collection occurred from November 2022 to May 2023. The recruitment email included a hyperlink, enabling participants to voluntarily schedule remote interviews. Prior to the interview, participants were asked to complete a demographic questionnaire and provide the title of a recently-read paper. Interviews were conducted and recorded via Zoom video conferencing. For the first part of the interview, questions are centered around participants' experiences with literature search and assessment. In the second part, we demonstrated the prototype interface [100] using the paper chosen by the participant. The web-based interface comprises three main functions: PDF Feature Extraction, Run Synthetic Market, and Explore Results (See Figure 1). In PDF Feature Extraction, users can upload PDFs of published papers for assessment, and the system automatically extracts relevant features. The Run Synthetic Market function then generates replicability estimates. Figure 2 displays an example of these estimates, which range from 0 to 100; higher scores indicate a greater likelihood of successful replication. This score corresponds to the final price

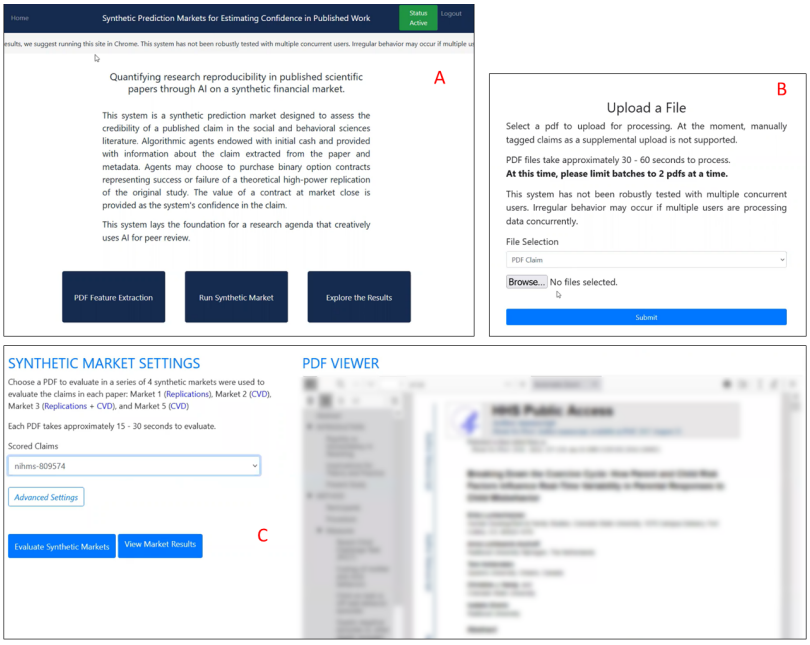

Figure 1: Web interface for replicability estimation prototype tool. A: Home page; B: Interface for uploading PDFs for scoring; C: Interface for paper evaluation, with the ability for users to adjust model hyperparameters using Advanced Settings.

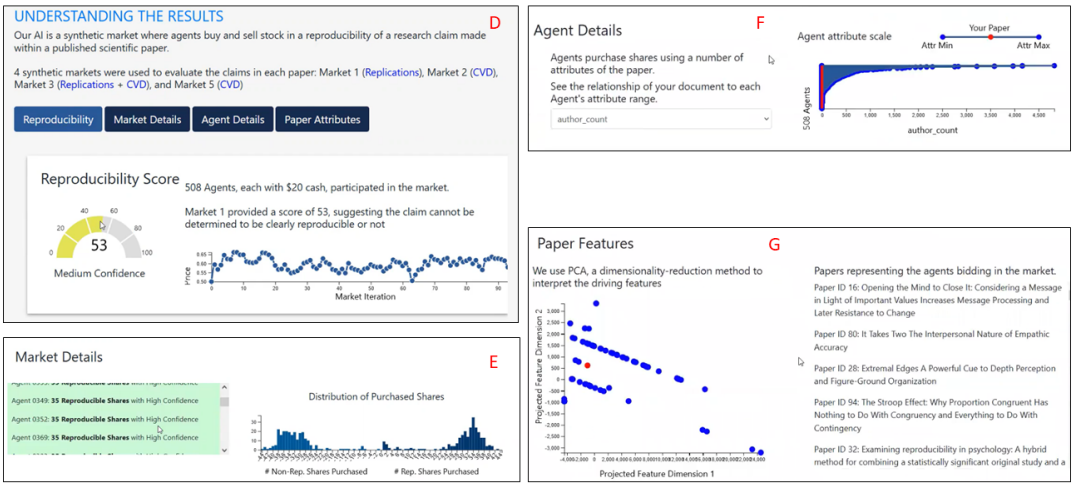

Figure 2: Example display of paper evaluation results. D: Replicability score. E: Explanations, presented as details about agents' behaviors/decisions; F: Publication's features in context; G: Visualization of principal components from extracted features.

in the synthetic market. Users can delve deeper into the results by examining the decisions of individual agents (Market Details), comparing the paper's features against agent value ranges (Agent Details), and scrutinizing the principal components extracted from the features (Paper Features).

This demonstration was followed by a series of questions exploring whether and how the replicability score might be useful in participants' research workflows. Additional questions were incorporated based on participants' responses during the interview. Figure 3 outlines the study protocol. All data were collected with informed consent from the participants and subsequently anonymized. Interview recordings were transcribed for further analysis.

Understanding of reproducibility. During the interview, we asked participants about their understanding of reproducibility and provided relevant explanations if participants were not able to answer the question. However, we did not assess whether participants' understanding aligns with the strict definition. This study seeks to report participants' understanding of reproducibility and how they assess it in their daily practices. Therefore, correcting

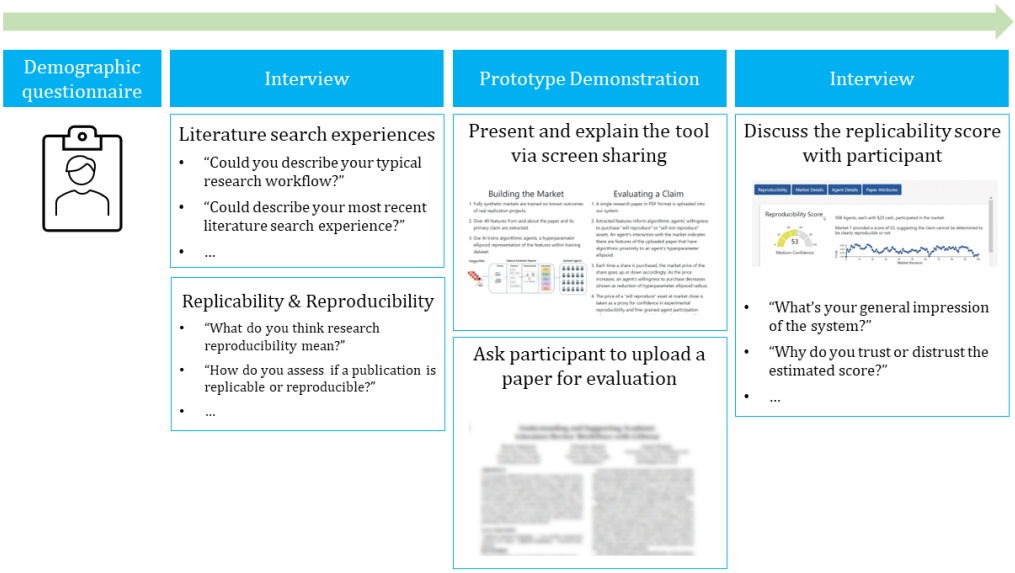

Figure 3: Illustration of the study protocol.

their understanding may lead to biased reflections on their previous experiences. In addition, the replicability score is estimated using various publications features instead of the definition of the concept. Following participants' current understanding of reproducibility enables us to obtain reactions to the system that align more closely with participants' needs and habits.

5.2 Data analysis

Interview transcripts were analyzed using a thematic analysis approach [21]. We adopted a collaborative and iterative coding process widely used by qualitative studies in HCI [4, 45, 65] to ensure the validity of the analysis process. Specifically, the first author thoroughly read transcripts multiple times to become familiar with the data. In the subsequent phase, open coding was performed to identify initial codes that represent meaningful segments that captured participants' opinions and behavioral patterns. These initial codes were then organized into potential themes pertinent to the primary research questions. After this initial coding phase, the first author met with all co-authors to discuss the meaning, similarities, and differences of the identified themes. Decisions were made collectively by the team regarding the retention, removal, or reorganization of these themes. For example, we grouped codes related to the author's reputation, journal reputation, and citation number under the theme "Metrics for filtering search results", as these qualities are frequently used by participants to select papers in the literature search. Subsequently, the first author continually reviewed and refined the themes by revisiting the transcripts and confirming the connections between themes and the high-level research questions. Once the themes were finalized, the team met again to interpret and develop them into narratives shown in Section 6. Analytic memos were employed throughout the process to supplement coding and facilitate reflections [40]. Analysis was done through NVivo 12, a qualitative data analysis software. The data analysis process was done iteratively by moving back and forth among the transcripts. All authors are active researchers in social science and closely related fields, including HCI. Our extensive experiences in literature review and management enable the team to make sense of the interview data. While we aim to avoid biases and stay neutral during data collection and interpretation, we acknowledge that findings could be subjective and depend on the authors' understanding.

6 FINDINGS

Seventeen participants were interviewed, with an average interview length of 42 minutes. Participants' self-reported demographic information is shown in Table 1, including gender, ethnicity, PhD progress (the number of years), and their main research field.

Table 1: Participants' demographic information.

ID	Gender	Ethnicity	PhD progress	Research field
P1	Female	Asian	1	Human-Computer Interaction
P2	Male	White	1	Management
P3	Female	Asian	1.5	Communications
P4	Female	White	5	Communications
P5	Male	Asian	2	Management
P6	Female	Pacific Islander	3	Marketing
P7	Female	White	1	Psychology
P8	Female	White	Postdoc	Psychology
P9	Male	White	3.5	Psychology
P10	Female	Other	3	Psychology
P11	Female	White	Postdoc	Psychology
P12	Male	White	3.5	Criminology
P13	Female	White	3	Sociology
P14	Male	Other	4	Political Science
P15	Female	Asian	3	Anthropology
P16	Male	Black	1	Anthropology
P17	Male	Asian	3	Human-Computer Interaction

6.1 Literature search: Strategies and challenges

6.1.1 Metrics for filtering search results. Literature search constitutes a critical step in a researcher's workflow, particularly during the early stages of a research project. Participants indicated that reviewing literature helped them to understand "what is still unknown", check "whether my questions are already answered", and establish "the empirical basis for the idea and the theoretical rationale for why I want to test the question".

Google Scholar emerges as the most commonly used search tool. Participants also utilized other platforms such as Web of Science and discipline-specific databases like PsycINFO. Although platforms like Web of Science offer advanced search queries and export functionalities, participants generally favored Google Scholar for its user-friendly interface. However, they noted that Google Scholar often yields overly broad and discipline-agnostic results. P6 mentioned that this issue diminished as her field expertise grew:

> *I had this question when I first entered the PhD program: basically, where do you find the most important literature? But I guess after years of experiments, I now purely rely on Google Scholar, because I can already know, what are the important journals? And what are the reliable sources? I can distinguish the most important and relevant information from Google Scholar [results].* -P6

As P6 noted, the choice of publication venue serves as a critical metric for evaluating search results. Other criteria include the number of citations, the author's reputation, and the recency of the publication. For instance, P5 values the reputation of authors or their institutions as indicators of research quality:

> *It depends on the quality of the journal and impact factor, and depends on if I know the authors because I'm mostly familiar with all the prominent authors in my area. If I recognize the authors or the institute that the paper is from. Then I do deeper [reading]. If it is from Harvard, Stanford, Wharton, Michigan, I know that it's of top quality.* -P5

Researchers new to a field, who are less familiar with established names, may rely more on citation counts as a measure of credibility:

> *I've studied psychology for about 10 years but as for Anthropology, I'm a newbie. So even if I see an author's name, there's a high chance I will not recognize it, even though he or she is a figure in the field. So, I have to rely on peripheral information like citation numbers. I would like to have at least two digits, even if it was published recently.* -P15

While P15 held a high standard of requiring at least two digits in the citation count, P8 was less persuaded by the impact of the citation number:

> *It's good if it's published in peer-reviewed journals and I don't tend to care about the impact factor or like how many citations.* -P8

P8 contended that the peer review process alone validates research quality. The majority of participants, however, paid heed to additional metrics. Author reputation not only influenced the evaluation of search results but also aided in the discovery of further relevant papers. Participants would follow authors' profiles on Google Scholar and institutional websites to find publications of potential interest:

> *Once I start to identify particular researchers who seem to be hitting on the topics that I am more interested in, I will go to Google the profiles of those researchers or look up their professional profiles from their institutions and try to read other stuff that they've done* -P7

Participants also employed citation networks to identify relevant literature. For instance, after identifying a foundational paper, they may explore its cited references or papers that have cited it to find related works. Instead of research quality, a citation network is mainly helpful in ensuring that the search is related to a specific topic:

> *One thing I'll do is I'll look up a foundational kind of article; pop it into Google Scholar and see who else cited it, and then see if there are any papers that are similar to the topic I want to go down.* -P4

Publication recency is another commonly used metric, particularly for empirical or experimental studies. The age of a paper is generally less critical for theoretical contributions:

> *Usually, I prefer a journal publication that was less than 10 years, but if I need to find some theory papers, I don't mind what year this paper was published.* -P1

P7 also added that it would be fine for review and meta-analysis papers to be a little bit old, because *"it'll still give me a summary of a lot of things and conceptualize thought pieces."* P4 stated that it's important to have a mixture of recent publications and foundational papers that may be very old, which can *"make sure that you are touching on those foundational pieces"*. Recency could also interact with other metrics during the assessment. For example, P11 would make exceptions for "outdated" papers if they were published by prominent researchers he could recognize. P10 pointed out the citation number may not be a valid indicator of quality as new papers tend to get fewer citations but can still be of good quality.

Despite some nuances in their importance, these easily accessible paper metrics were frequently used, and they helped participants quickly filter through the long list of search results. Still, these metrics provide limited information about a paper, and participants need to take further steps to assess papers and seek information useful to their own research.

6.1.2 Paper evaluation and information seeking. Evaluating papers in detail often commences with the abstract, which serves as *"the synopsis of the article"* for most participants. Once the abstract confirms the paper's relevance, the participants' approaches to reading the remaining content diverge. P7, for instance, prioritized the results section because the data allowed her to see if hypotheses were supported and then interpret the statistics in a relatively independent manner:

> *I want to see the general summary statistics that they can provide. Descriptive statistics and correlations and things. Just to give me the first bird's eye view of what might be going on with their data. And then I will read*

> *the hypotheses that they proposed beforehand and experiment... I prefer to look at the tables and the charts first, so that I can kind of draw my own conclusions and then see how they're presenting their information.* -P7

P7 believed that the numbers were important for understanding how a paper's conclusions were formulated. It seemed that participants who prioritized the result section were more sensitive to numbers and statistics, as P10 commented:

> *I have a lot of statistical background. So I kind of examine whether the statistics they chose match the question they were asking. And if I can follow why they chose the model, and if the results of that model seemed to speak to a meaningful finding, as opposed to maybe just like p-hacking. Those are the things I focus on.* -P10

By contrast, P3 believed that it wouldn't be helpful to focus on the results due to the publication bias favoring positive or statistically significant results:

> *It's almost never the results. Result is usually not the way that I filter. And part of that is the skepticism that for most papers, if it's hypothesized, then it will get published if the hypotheses are verified.* -P3

Similarly, P16 pointed out it was the method section that determined the integrity of the paper, instead of the results:

> *From the methods, you can know how detailed the sampling was. What was ethical compliance? And how they adjusted their methods for the research questions? So, for me, it's not really about the results, but it's about the method.* -P16

P13 also emphasized other sections over results, but due to the potential complexity in understanding the details. Instead, he focused on discussion and conclusion to understand a paper:

> *After the abstract, I'd say the intro, and then the discussion and conclusion. And then I usually go back and just quickly skim the literature review. Depending on what I'm trying to do, sometimes I'll read the data and methods as well. I feel like the results sometimes are just, there's too much going on here. So I'll just stick to the conclusion.* -P13

Despite the differences, participant' priorities in paper evaluation often depend on their current research phase or specific informational needs. For example, despite her emphasis on results, P7 would scrutinize the methods section when venturing into an unfamiliar discipline:

> *It really depends on what I was looking for. For example, if I don't know how to design the experiments for this specific type of study, I go to method part and also result part... If it's an experiment that's outside of my discipline, I look at the methods really heavily, because there are different standards and expectations in different disciplines.* -P7

When dealing with unfamiliar topics, P10 would instead read the introduction carefully. Similarly, P4 would pay more attention to the early sections of a paper when learning a new research topic:

> *It depends on where I'm at in my literature review. If it is a new topic, I'll be trying to figure out how people are defining certain concepts, then I probably do pay more attention to the literature review, just so I can figure out who are the big names in this field? What are the definitions?* -P4

The introduction and literature sections were helpful for participants to figure out the main topics and theories discussed and their definitions, whereas the method section can inspire new ideas for experiment designs. As P6 summarized, the iterative nature of research often prompts revisiting papers with a changing focus, depending on the project's stage:

> *I think that will be dependent on my goals. For example, when you are doing a literature review on the earliest stage, you may focus on what topics they cover, but as you proceed into the specific research question, you would probably look but more into the details of the specific methods they use. And at the end, when you try to write a paper then probably you will follow others' story lines and how they how they develop the whole paper. So I guess I will revisit a paper again from time to time during the whole project.* -P6

Overall, participants employ utility-oriented strategies for evaluating papers, focusing on aspects they believe would either effectively assess a paper's quality or contribute to their own research. As P9 commented, he needed to do a lot of research, and time efficiency really mattered in the literature search.

6.1.3 Challenges to literature search. Participants primarily used keyword-based searches to locate literature, but this approach often presented challenges. They found that search engines like Google Scholar did not always yield results consistent with their keyword expectations. This inconsistency was especially prevalent in interdisciplinary or innovative research, where the same concept may have multiple names:

> *My area of research is so interdisciplinary. Researchers in this field that I'm not familiar with, they might be studying this thing, and they are just calling it something that I wouldn't expect them to call it. So I feel like there is a level of randomness added to how I can discover papers on topics that I want to study. I just need to know what words to use. That's probably been my biggest challenge.* -P7

To address this, P4 suggested that search engines should offer keyword recommendations akin to a thesaurus. P6, however, was skeptical of recommendation systems, citing her experience with irrelevant alerts from Google Scholar. She proposed that a visual representation of the citation network would be more beneficial. In fact, several participants echoed the need for enhanced visualization features, such as a graphical representation of citations or co-authorship networks. While Google Scholar fulfills part of the expectations by providing the list of "Cited by" papers and the list of "Co-authors", participants desired more advanced functionalities:

> *Like a platform that you can give them a topic, and then it gives you the main authors that have published stuff related to that topic... I guess it could use metrics like*

> *the h-index or the number of citations, and I guess you can also use more subjective things like where do they work, or what journals they have published, or their academic ranking?* -P14

Despite options like date ranges and author filters, current search methods often fall short of capturing the nuanced information that researchers seek, e.g., theoretical framework, research methods, and main findings. P13 likened this to online shopping experiences, expressing a desire for more customized search criteria:

> *Almost like when you're shopping online: I want this size in this color. So this would be like, I want [papers that used] these types of dataset, [from] this time period, [from] these countries or regions even... The general method that they use, like if they're using like, quantitative methods versus qualitative methods.* -P13

This desire for more specific details also reflected participants' need for more intelligent search tools that can summarize scientific reports. While abstracts are designed for quick navigation of a paper, participants have found them "not standardized" or "too big picture". Instead, an intelligent tool that could offer more tailored summaries could address this gap:

> *If there was a way to click on a paper and see what survey items are used to measure each construct, to pull that out of a paper more automatically would be helpful. Along with it, results and directionality of results. If it's positively correlated, or negatively correlated with other things.* -P3

In summary, while existing search methodologies offer a baseline utility, there is a growing need for more sophisticated, flexible, and intelligent tools. These should accommodate not only standardized keywords but also offer advanced features like network visualization and content-specific search options.

6.2 Reproducibility and replicability: Importance and estimation

6.2.1 Perception of the importance. Except for P2 and P17, all participants reported that they've heard of the concept of research replicability or reproducibility. During the interview, many participants have used the terms interchangeably. Accordingly, we do not strictly distinguish between the two in the following sections.

Participants generally agreed that replication and reproduction of studies strengthen the credibility of scientific findings:

> *A lot of how we evaluate statistical effects in psychology is p-value, which is the assessment of how likely these results are due to chance. So any single finding can always be due to chance. And so if you're reproducing studies and if you find the same results, it gives you a lot more confidence that it's probably not due to chance.* -P7

Despite the emphasis on replication, P7's interpretation of p-values was incomplete and, therefore, biased his understanding of replicability. Specifically, a p-value actually measures the probability of observing the data, or something more extreme, if the null hypothesis is true [37]. The emphasis on 'chance' also oversimplifies the size of the effect, the precision of the estimates, and the quality of the studies. Similarly, other participants highlighted some relevant aspects that are not at the core of the replicability crisis, demonstrating a lack of a strict definition of replicability but rather a general understanding of qualities for good research. For instance, while P7 and P14 noted that emphasizing replicability minimizes errors due to chance and guards against research misconduct, P8 was more concerned with the representativeness of study populations, emphasizing the need for diversity to enhance generalizability:

> *I think a lot of our research has been based on what we call the WEIRD sample, which is Western, European, Industrialized [Rich, and Democratic population], you know. And I'm not surprised as our samples become more diverse in various ways, we can't find the same results. I think it's a distinct sample where a lot of psychology is based off, and a lot of our theories are based off... That could be one issue for the reproducibility crisis.* -P8

Generalizability is often highly related to replicability in that non-replicable studies are typically considered non-generalizable. Yet, a study may be replicable but not generalizable if the findings only apply under specific conditions [90]. It seemed that participants conflated concerns about replicability with generalizability to some extent. For example, for certain types of research, participants regarded replicability as less important or not applicable, especially when the research is qualitative in nature or involves the context of human interaction:

> *If you're doing interviews, it's impossible [to reproduce]. If someone were to interview me in a similar way that we're talking right now, but two weeks later, those interviews will be different. Even if someone was to just listen in our recorded interview. They still would miss the interaction that we're having.* -P3

> *I'm studying Organizational Psychology. And you cannot create the natural environments to understand people's behavior. In my field, context is the determinant. That's why I would understand that people cannot replicate studies* -P9

In addition, participants noted that even though replicability is important in many fields, the effort of validating studies is much less appreciated by scholars than innovative findings:

> *I've gotten reviews that say, it's important that this research (replication studies) is being done but the journals don't want to or not always interested in and publishing such work, because it doesn't advance the field forward, it only makes sure that the base that we have is solid. So while they appreciate the effort, they don't always want to publish it within the top tier journals.* -P12

Their understanding of replicability, while not accurate, may reflect some common perceptions among researchers, highlighting the need for curriculum interventions. However, this mixture of replicability and other properties of research quality does not equal misunderstanding and has, nevertheless, motivated participants to pay attention to good research practices. For example, during the literature search, while platforms lack explicit metrics

for reproducibility and replicability, participants have made assessments of these qualities to various extents when reading papers. As mentioned in Section 6.1.2, many participants paid significant attention to the method section of a paper, and by examining the rationale of the research methods, they were implicitly evaluating the replicability:

> *I am a bit skeptical while reading the design and methods because I saw they hypothesize something and they are measuring in a way that has many flaws. I would say it's not sure if they can replicate this study under the same conditions.* -P9

Again, this evaluation often included other aspects of the paper quality, such as generalizability:

> *I would be paying attention to things like how generalizable is the sample that they recruit? How many people did they recruit? A lot of studies, especially older studies, had like teeny tiny little samples, and that may not be valuable. And then assuming it's a method I'm really familiar with, like eye tracking, or imaging, I can use that section to evaluate what they did was accurate as opposed to poorly followed methodology.* -P10

More explicitly, participants said that clear descriptions of research methods, good rationale for research choices, and availability of research materials would convince them that a study is reproducible:

> *I am all for a very clear procedure, stimuli and measure section. For me, that is important. I feel like a lot of times, authors say if you want to see the stimuli, email us, if you want the full list of measures, email us, I think that's fine. But especially as a more junior researcher, it is helpful to be able to see the step-by-step [methods].* -P4

> *I think it is (reproducible), because it's a widely available dataset, and has very clear instructions. There's not a lot of wiggle room in that.* -P8

Some participants also strongly advocated for the practice of pre-registration. Pre-registration can ensure that a publication contains sufficient detail to permit reproduction and can reduce the bias in reporting significant results:

> *Ultimately, I really like the idea of the registered reports, because the journal is saying they're going to accept the article whether the results are null or not. So I think that's something I would want to focus on moving forward.* -P10

Generally, participants acknowledged that reproducibility and replicability are important concepts for the research community. Despite some confusion between replicability and other aspects of research quality, participants have paid attention to reproducible practices during their literature search. Rigorous research methods and availability of research materials were the main proxies that enabled participants to estimate paper quality and thereby impact their literature review process.

6.2.2 Trust in AI-estimated replicability. The AI-estimated score offers a quantitative metric for assessing the replicability of research without the need for actual replication. Some participants felt the score aligned well with their impression of the specific paper in question. However, others were hesitant to trust the AI system without understanding its underlying mechanisms:

> *For now, I don't think I trust it because I don't really understand how this system works. But once I understand how it work, and I think it's reasonable, I think I can trust more. A little bit of explanation of how these algorithms have been trained, and how many research papers are included in the training sets?* -P1

Participants less familiar with AI and machine learning techniques found it particularly challenging to interpret the scores. This is because machine learning features may not have an explicit causal relationship with the predictions:

> *I don't know anything about machine learning. I'm sure these variables connect in a way that I don't understand. But like, I don't understand why the number of authors makes a difference, or the citations, or where the authors are from. I think p-values are important, but I also think it depends on like, what were your stimuli? who were the participants? Was it a generalizable sample? You know, where did you get it from? Was it an experiment or survey? I feel like the methods play more of a role, whereas I feel like this is not looking at methods.* -P4

P4 expressed concerns that features like the number of authors or citations may not be reliable indicators of replicability. Similarly, P14 questioned whether recently published papers would be unfairly deemed less reproducible due to a lower citation count or if the system would be ineffective for qualitative studies that don't rely on p-values. According to the participants, the trustworthiness of AI-generated results depends on whether specific features serve as reasonable predictors for replicability. As P8 suggested, some descriptive summary accompanying the estimation score could aid in interpreting the results:

> *The conclusion is there, but it's very limited. If it could just say it's based on certain features that stood out, like, no citations, or the sample size, or whatever it's using. Not so much quantitative like in these numbers, but more of a narrative, that would be helpful.* -P8

The lack of sufficient explainability led participants to hesitate to rely on AI for estimating replicability. As P17 noted, while the system could be useful for making predictions, using it for assessments would require stronger evidence, given that accusations of non-replicability carry significant weight:

> *We need to have people redo the whole experiment (to decide reproducibility). The critics that others' work is not reproducible is a very strong accusation. So, you need to have very strong evidence to show that something is wrong.* -P17

Moreover, an indecisive estimation score further eroded participants' confidence in the system. For instance, a score of around 50 made it hard for participants to form a strong opinion. Similarly, a bimodal distribution of results from all agents (Figure 2) raised additional questions:

> *It's a really bimodal distribution with a lot of bots thinking that it's not reproducible, and a lot think that it is reproducible. So I would wonder what the factors that contributed to that? Without knowing why it's bimodal, I would probably trust the results less.* -P11

A bimodal distribution could signal high uncertainty, causing distrust. AI-estimated replicability was novel to most participants, and their trust might improve through continued interaction with the tool. P12 suggested that gaining confidence in the tool would require practice with multiple papers and analysis of result variability. Providing context for participants to compare and interpret scores could also enhance their trust:

> *I would like to have some reference scores. For example, in this research field, 80% of the empirical articles are receiving this score, so that's why 59 is a good score for this field. Comparisons would make me feel more confident about the result here.* -P9

In summary, although there was no outright rejection of AI-estimated replicability, participants were hesitant to trust the results fully. Providing detailed and transparent explanations of the AI system's workings and how each feature contributes to the estimation is essential for users to make informed decisions. Additionally, exposure to multiple papers or having reference points to interpret scores could further help users assess the tool's utility.

6.2.3 Potential use of AI-estimated replicability scores. While the prototype aims to offer researchers a level of confidence in published studies, participants expressed diverse opinions on its potential benefits. Firstly, some participants questioned the value of providing an estimated replicability score. As referenced in 6.2.1, P15 argued that the concept is less relevant to qualitative and context-sensitive studies, doubting the system's usefulness. Similarly, P2 felt that the score's relevance was minimal as his priority in the literature search was the paper's relevance to his research. P6, on the other hand, considered the citation count as a sufficient indicator of paper quality:

> *For the paper I just mentioned, I believe it has over 5000 citations. So that's definitely it's a good sign. This is publicly available and can be easily obtained. And if I already have this type of information, why bother using a more complicated system?* -P6

Others were more optimistic about integrating the system into their research workflow. P10, for instance, felt that the estimated score could enrich the literature review process by flagging potential methodological flaws:

> *I would imagine, as I was writing my literature review, I would be checking papers. And I would be saying, hey, in my literature review, I'm citing this paper, because it's foundational to the way we think about things. But there are concerns about whether it would be reproducible.* -P10

In this scenario, the estimation could help identify weaknesses in foundational literature, thereby improving the rigor of literature reviews. Similarly, P9 envisioned the tool as a means to quantify the strength of research arguments based on cited references:

> *If I have enough time, I would prepare a table with the empirical articles that I used and the scores that they have received to increase the credibility of my arguments. These empirical papers have high reproducible scores, which I use to support my arguments therefore my arguments are credible.* -P9

Specifically, P9 saw the tool as useful for addressing reviewers' concerns about study designs during the peer-review process, citing high replicability scores as justification for chosen research methods.

Given that publishing is a primary goal for researchers, many participants considered the tool beneficial for peer review and similar contexts. Although originally designed for evaluating published work, P13 suggested using it as a pre-submission checker, particularly benefiting researchers with less publication experience:

> *Especially like an early career researcher or grad student could pump their own papers in and then see what these synthetic agents think about my paper and (estimate) what other people are going to think about my paper.* -P13

This suggestion was seconded by P11, who wanted to use the tool to assess his own work and see how reproducible and replicable they were perceived to be. By contrast, P12 provided a different perspective as paper reviewers, and suggested that the system could help reviewers who were asked to review papers that they have conceptual and theoretical interests in, but don't have methodological sophistication:

> *So if I got asked to review a paper that involved machine learning, which was conceptually and theoretically involved in my area, and that's why I was chosen as a reviewer, I would probably use this tool as a kind of a cross-check against my thoughts on the paper.* -P11

Similarly, P15 proposed that funding institutions could use the score in evaluating grant proposals:

> *I think institutions such as the National Science Foundation, you know, they review grant proposals for funding, right? And replicability in itself is very important issue in science, regardless of hardcore science or social science. So institutions might be able to use this score to tell whether the proposed project has reproducibility.* -P15

Despite participants' interests in augmenting the peer review process with the estimated reproducibility, the functionalities of the tool need to be enhanced and expanded for unpublished papers to fulfill the usages suggested above. While participants were generally interested in using the estimated reproducibility to augment the peer review process, the prototype would require further enhancements to cater to unpublished papers. Most participants agreed that more information would at least be interesting, if not particularly helpful. However, one participant warned against the unintended consequences of relying too heavily on algorithmically generated scores. Researchers might make efforts to boost the score of their papers without actually enhancing the research quality:

> *I see it as helpful but also potentially harmful. Like if someone doesn't work out their paper in a certain way*

> *that is similar to other papers, but it's still reproducible? I can imagine there would be consequences for those papers that wrongfully hurt people's careers... With this competence score, I can see some strange reactions to it from scholars like trying to maximize the reproducibility score, where it's not really adding to the science, it's just adding to the perception of how people think of it.* -P3

In summary, while the estimated replicability score might not necessarily facilitate the literature search or review process for all, it could serve as a valuable tool for supporting research arguments, checking paper quality, and aiding in peer review. It should be noted that all suggestions are made based on the assumption that the estimation is fully reliable, which requires significant research efforts to improve the current tool. More importantly, further investigations on ethical implications are essential for future iterations of AI-estimated replicability.

7 DISCUSSION

Through qualitative interviews with 17 social science researchers, the current study explores design opportunities for technologies to support better literature review processes taking into consideration reproducibility and replicability. We present preliminary findings for three research questions. 1) We illustrate participants' approaches to literature search and review, highlighting the use of various indicators and utility-oriented information seeking (Section 6.1.1 & 6.1.2). 2) We identify several challenges participants encountered during the literature search, including the lack of expertise and limited capabilities of keywords-based search. 3) In terms of reproducibility, participants' understanding of the concept was blurred with perceptions of general research quality. Nonetheless, their feedback demonstrated the potential benefits of providing quantitative estimation of reproducibility for scientific publication. However, they also raised concerns about the system's explainability and interoperability and the ethical considerations of employing such a system. In the following sections, we contextualize these findings within relevant literature and discuss design implications as well as future research directions.

7.1 Future technologies for supporting literature review

Conducting effective literature search and review is foundational to research. However, the sheer volume of publications has made the task of identifying relevant literature increasingly challenging. While the HCI community has developed several design prototypes to potentially facilitate literature search and discovery, researchers predominantly rely on established bibliographic databases like Web of Science and Google Scholar. Consequently, understanding researchers' current experiences during literature search is crucial for developing effective and widely adopted technologies. Our findings about literature search align with prior research on user experiences with web search, highlighting the need for search tools that can handle more complex search queries [114] and the impact of domain expertise in effective search [82]. Yet, literature search brings a set of unique metrics which might inform, prioritize, and provide necessary context to search reseults. These findings also align with prior research emphasizing engineers' and technologists' use of publication venues and author names in search [5] and confirm some of the challenges encountered by computer scientists, such as keeping up to date with research and exploring unfamiliar topics [7]. These considerations can inform the design of better literature search tools, as discussed below.

7.1.1 Design implications for literature search tools. **Domain specific search interface:** Past research highlights Google Scholar's advantages, such as its user-friendly, easily navigable interface and extensive resource base [16, 83]. Our participants corroborated these points, stating a preference for Google Scholar. Although Google Scholar supports a smooth user experience, our participants expressed a need for more advanced features. One major limitation is that Google Scholar returns less specific and potentially lower-quality results compared to other databases. Some participants felt their expertise and familiarity with the topic helped mitigate this shortcoming. However, this presents an obstacle for novice researchers or those exploring an unfamiliar field, who may lack foundational knowledge about recognized venues and authors in that space.

Venue reputation and author reputation are used as indicators for good papers. Although impact factors and h-indexes provide some insight, participants largely relied on their own experience to assess these reputations. Some design prototypes have indirectly aided users in identifying key researchers through citation network visualization [30, 46, 109], but there is scant focus on venue and author reputation in existing literature search tool designs.

These findings highlight the benefits of incorporating domain-specific expertise in information search. Domain expertise can be used to present better results and query suggestions to users [115]. Therefore, the design of domain-specific search tools could be customized accordingly. For literature search tools, the user experience could be improved through better organization of search results based on expert knowledge: 1) enable users to filter results based on research areas; 2) recommend well-recognized publication venues; and 3) recommend well-recognized authors and allow users to filter results accordingly.

While these functionalities could help users find desired papers more efficiently, there are ethical considerations that should be further explored. Specifically, exclusive focus on well-known venues and authors could hinder the progress of science, which also relies on contributions from newcomers. Our participants have largely used citation count as an indicator of paper quality, with some mentioning the bias against newer publications, which naturally have fewer citations. Researchers have proposed alternative metrics such as weighted citation counts and Altmetrics [17, 117]. Future research on literature search should investigate how we can leverage these metrics to introduce emerging researchers while facilitating the discovery of foundational works.

Support complex user queries: Literature search tools and algorithms should also be designed to handle more complex user queries. First, the quality of keyword-based search results is highly dependent on the keywords entered. Choe et al. [30] observed that novice researchers often struggle with academic terminologies and proposed an interface that recommends relevant keywords based on the current search queries. Our study confirmed this challenge

but also underscored the usefulness of providing keyword alternatives that capture synonymous research concepts, as this was a common struggle among participants.

More crucially, keyword-based search methods alone may be insufficient for fulfilling the nuanced needs of the search process. Participants expressed a desire not just to locate papers on specific research topics but also to find research that employs particular methodologies, utilizes specific datasets, or even adopts a particular viewpoint on a research issue. While the exact algorithm behind the Google Scholar search engine remains proprietary, it appears incapable of handling such complex queries. One possible solution is to leverage the fast-advancing capabilities of large language models (LLMs) [15, 72]. Preliminary work has already been done on AI-based tools that can identify pertinent papers and summarize their key findings based on user-generated questions rather than mere keywords [73]. Our study suggests that this approach may better satisfy the needs of researchers. Therefore, future research should seek to evaluate the capacity of LLMs in handling complex search queries, and investigate the impact on user experience. In terms of technological design, the search tools should leverage LLMs to achieve the following: 1) identify and suggest alternative research terminologies that may increase the performance of user queries; and 2) perform deep analysis of literature to answer specific user queries on research questions, methods, and findings. These functionalities may accelerate the process of literature review, helping scholars to identify and organize papers that meet specific criteria without going through a long list of search results. Scholars can be empowered to understand theories, methodological approaches, and certain hypotheses more efficiently, which facilitates the process of scientific discovery. While promising, research is also needed to evaluate the verifiability of results from search engines empowered by LLMs [78].

7.1.2 Problems with scientific paper reading. After identifying relevant papers, researchers also invest considerable time in reading and understanding them. Prior studies have developed tools that aim to facilitate the reading process, using techniques such as tooltips and Natural Language Processing (NLP) to augment the text summarization[28, 61, 93]. While our study did not directly identify challenges in the reading process, it did reveal significant variability in participants' reading strategies. Specifically, they prioritized different paper sections based on their information-seeking needs.

This observation raises important questions. First, should this utility-oriented reading style be encouraged? Extant literature offers varied perspectives on effective approaches to reading and understanding a scientific paper [57, 68]. Some suggest deviating from the standard IMRaD (Introduction, Methods, Results, and Discussion) structure for a more efficient understanding [47]. While reading papers out of order or skipping sections may increase efficiency, there is a risk of neglecting essential information, potentially leading to misunderstandings and inappropriate referencing. The second question is more complicated: is it time to reconsider the IMRaD structure? The structure of scientific papers has been guided by long-standing traditions and norms, and despite criticisms regarding its rigidity [105], the research community continues to adhere to it. In light of the changes brought about by computer-mediated communication, some researchers advocate for new forms of scholarly communication [41, 97]. This exploratory study provides preliminary insights into this issue. Instead of concrete design implications, we believe that more thorough examinations of the following are required for supporting scientific paper reading: 1) how augmentative technologies can help readers locate information in a scientific report; 2) whether the use of such technologies would lead to lack-of-context and biased interpretations of a scientific report; and eventually 3) whether we need more modern and diverse formats for scientific reporting. The extension of reporting beyond IMRaD can accommodate the diversity of research methodologies and genres [49] and potentially facilitate the identification and reading of literature. In addition, the investigation of the misalignment between IMRaD and reading behaviors will inform teaching in scientific writing and help scholars to communicate their work in a manner that effectively captures attention from broad audiences [104].

7.2 Promises and risks of replicability estimation

Replicability and reproducibility are critical to the integrity of scientific progress. The replicability crisis has gained widespread recognition across disciplines, leading the scientific community to propose various initatives aimed at fostering reproducible research practices [42, 52]. This study contributes a unique perspective by examining the implications of quantitative metrics on the literature search and review process.

7.2.1 Promote understanding of key concepts. While participants in this study were generally aware of the concepts of reproducibility and reproducibility and acknowledged their importance, their definitions of these terms varied. This discrepancy underscores the potential need to address these concepts more rigorously in graduate education. For instance, P7's statement about p-values showed a common misconception about p-values [113] and, therefore, an incorrect definition of replication. Some participants confused the concepts of generalizability with reproducibility. This finding suggests that many researchers may have a nebulous understanding of replicability as a concept generally related to research quality. This general understanding does not undermine the validity of the study as we seek to reflect on participants' current understanding and daily practices. However, it also suggests that research is needed to help design curriculum and educational resources that promote researchers' understanding of key concepts, which can hardly be achieved through the prediction tool.

Participants P3 and P9 emphasized that specific characteristics of human behavior studies, such as qualitative interviews, are challenging to replicate, calling into question the applicability of replicability to these types of studies. Heightened awareness of the replicability crisis has primarily arisen from publications focusing on quantitative and statistical methods [9]. However, recent arguments from qualitative researchers suggest that even in studies not aimed at hypothesis testing, preregistration practices can enhance reproducibility by serving as a check on subjectivity [74]. The reproducibility of qualitative research remains a significantly underexplored area [33]. Our findings indicate a need to include

more qualitative researchers in discussions about reproducibility to broaden awareness and advocate for reproducible practices in qualitative studies.

Design opportunities for literature review: Contrary to the importance of replicability in current scientific dialogue, participants did not explicitly seek out indicators of replicability or reproducibility during their literature searches. Instead, they implicitly assessed these factors by scrutinizing the methodologies described in the papers they read. While such assessments can be somewhat subjective, objective indicators like data availability, code availability, and preregistration practices could bolster readers' confidence in a study. As such, literature search tool designers should consider highlighting these indicators in search interfaces and enable users to prioritize studies that adhere to reproducible practices. Previous studies have shown that open-access (OA) articles are cited slightly more frequently than non-OA articles [110]. Still, there is a noticeable gap in research exploring the impact of reproducible practices on citation rates. This study presents preliminary evidence that users pay attention to credibility markers when evaluating scholarly work. Future studies should examine how incorporating these credibility markers into the literature search experience may encourage researchers to prioritize publications adhering to recommended research practices.

7.2.2 Transparent AI estimation of replicability. In this study, we collected user feedback on a prototype AI to estimate replicability, building upon prior work [100]. Rather than assessing the accuracy of the prototype's estimations, our focus lied in understanding how a quantitative metric for reproducibility and reproducibility might impact the literature review process. Broadly speaking, participants expressed hesitancy to trust and use the AI-generated estimations, attributing their skepticism mainly to the interface's lack of explainability. Explainable AI (XAI) is crucial for facilitating user understanding in AI applications [103]. The current tool provides limited explanations of the decision-making process and related uncertainty. Specifically, they were unclear about which features most influenced a paper's estimated replicability, requesting quantification of the impact of each input on the output. This is a common technique used for developing more transparent systems [77]. More critically, participants also expressed concern over some of the metrics used by the tool, such as the number of authors and university ranking, which they considered counter-intuitive. Generally, researchers tend to rely on methodological rigor to assess reproducibility, rather than these auxiliary metrics. Machine learning models tend to capture correlations in data to make predictions; however, researchers advocate that XAI should aim to identify causal relationships to better align with human understanding [31]. This issue is especially important for replicability estimations. The decision of whether a publication is replicable often comes with significant ethical implications for both authors and readers. Our results suggest that the major reliance on non-causal relationships in replicability estimations may not be acceptable for researchers. Future research should try to address this issue from two aspects: 1) improve the transparency of model decision-making and enhance the presentation of decision uncertainty, which can facilitate human-AI collaboration by motivating users to think more vigilantly about the results; 2) seek features that can represent the actual causal relationship with replicability, striking a balance between model efficiency and socio-technical benefits.

Design implications for quantifying replicability: One important improvement suggested by participants is the contextualization of estimation scores. On the one hand, the state of reproducibility can vary significantly across disciplines, rendering a 0-100 range too general for universal application. On the other hand, a score like 50 can appear ambiguous to users, necessitating additional context for proper interpretation. As Gunning et al. [59] have posited, the definitions of interpretability and explainability in XAI may be domain-specific. Our study, as an initial exploration, suggests that within the realm of reproducibility, enhancing both transparency in decision-making and contextualization of results is critical for users to effectively interpret AI and decide whether and how they will use the output. In the design of AI and machine learning systems, explainability and interpretability should be emphasized in order to obtain improvement feedback from users and promote better human-AI collaboration.

7.2.3 Benefits and risks of estimating replicability. Replicability estimation is more than just providing a score to users. The wide applications of such systems may have a mixed impact on research activities and important ethical considerations. On the bright side, participants speculated that the score could serve as a subjective gauge of integrity for their literature review, support their methodological choices, or even supplement traditional peer review processes. This suggests that participants view reproducibility and replicability as important criteria for research quality and that a quantitative metric could facilitate their evaluation of scholarly works. While previous studies have associated research quality with reproducibility, they also highlight a lack of uniformity in evaluating reproducibility [66]. Our findings suggest that providing a quantitative metric could help researchers identify replicable work to serve as foundational premises for their own research. Therefore, introducing the replicability estimation tool can positively impact the research community, promoting awareness of and emphasis on replicable and reproducible work. Researchers could easily determine if findings from a prior publication can stand the test of scrutiny, which can be incorporated into literature review, peer review, and publication process. As the current research culture often disproportionately favors novel and affirmative results [88], such an impact aids in upholding the integrity of scientific progress but also fosters a more transparent and trustworthy research environment. However, it should noted that these potential benefits are based on the assumption that the metric is fully reliable and trustworthy. As discussed in Section 7.2.2, there are several issues to be addressed before the AI estimation can be accepted as a valid indicator for reproducibility.

More importantly, there are potential risks associated with quantifying replicability, as participants have warned against during the interview. Specifically, they were concerned that researchers might chase higher scores in ways that could be counterproductive or unethical. Indeed, it has been criticized by the research community that efforts to enhance reproducibility may not necessarily promote high-quality research [76] and the over-emphasize of reproducibility may hinder the progress of scientific discoveries to some extent [5]. Another fundamental issue with this tool is the validity and

reliability of its output, as its capability to handle the numerous, ever-increasing future publications remains uncertain. Developing safe and useful AI systems will require us to make progress on scalable oversight [19], which enables continuous supervision of AI performance. As such, further investigations are required to explore the practical and ethical ramifications of quantifying replicability, especially concerning how it may alter research practices. The current study serves as a starting point for examining replicability metrics. The preliminary findings suggest that providing indicators for replicability or research quality, in general, could potentially benefit researchers by supporting literature review, yet the design of such metrics must carefully consider transparency and explainability of the interface as well as the ethical implications of deploying the system. In addition to technical improvements, policy, and institutional initiatives are also desirable to ensure the quantification of replicability will lead to advancements in the quality and integrity of scientific research.

8 LIMITATIONS AND FUTURE WORK

This study has several limitations that future research should address. First, as we have discussed, even though open code and open data do not typically apply to qualitative studies, they can still benefit from pre-registration. The current study is not pre-registered due to its exploratory nature and lack of specific hypotheses. This is a limitation that will be addressed in future confirmatory studies. Second, we relied on a prototype built upon existing work. Although it offers advanced functionalities, there is scope for further refinement, and alternative designs for AI-based replicability estimation could present different advantages and drawbacks. Third, our study used a lab-based demonstration approach to gather feedback. Conducting a longitudinal study that involves the tool's usage in real research settings would provide a more in-depth understanding of its impact on the research life cycle.

Several pertinent questions arose from our findings, although they fall outside the scope of this study. For instance, participants exhibited a somewhat nebulous understanding of reproducibility and replicability, which was further complicated by other concerns related to research validity. While these concerns are intrinsically linked to research quality, the consequences of this overlapping understanding merit further investigation. Furthermore, participants questioned the rationale behind using metrics such as citation counts and author reputation to estimate replicability, despite admitting to using these very metrics in their own literature searches. Future research should explore to what extent this reliance on traditional metrics is a byproduct of community culture rather than an indicator of research quality.

9 CONCLUSION

This study adopted a qualitative approach to explore potential enhancements to the literature search, particularly through the integration of reproducibility and replicability estimation. By examining participants' existing literature search practices, we identified key metrics that influence search result filtering. These metrics—such as venue reputation, author reputation, citation counts, and publication recency—are only partially supported by current search tools. There is a clear need for more intelligent search tools that offer users flexibility in retrieving information that suits their requirements at various stages of research. While incorporating replicability metrics into search processes has the potential to support researchers, we observed mixed attitudes towards AI-driven replicability estimations. Issues related to explainability, interpretability, and causality within AI models should be addressed before users can assess the usefulness of such systems. Assuming that these estimations are trustworthy, they could be used to strengthen research design and even facilitate the peer-review process. However, this comes with ethical implications and an undue emphasis on reproducibility may not necessarily result in better research quality, demanding further investigations. Overall, this study offers valuable insights for enhancing the design of literature search and management tools. It also underscores the need for further reseach and discussion on integrating replicability metrics, addressing ethical considerations, and exploring the balance between reproducibility and research quality.

REFERENCES

[1] Ahmed Al-Zubidy and Jeffrey C Carver. 2019. Identification and prioritization of SLR search tool requirements: an SLR and a survey. *Empirical Software Engineering* 24 (2019), 139–169.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805.

[3] Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. 2019. Predicting the replicability of social science lab experiments. *PloS one* 14, 12 (2019), e0225826.

[4] Tawfiq Ammari and Sarita Schoenebeck. 2015. Understanding and supporting fathers and fatherhood on social media sites. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1905–1914.

[5] Alia Arshad and Kanwal Ameen. 2019. Scholarly information seeking of academic engineers and technologists. *International Information & Library Review* 51, 1 (2019), 1–8.

[6] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651.

[7] Kumaripaba Athukorala, Eve Hoggan, Anu Lehtiö, Tuukka Ruotsalo, and Giulio Jacucci. 2013. Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. *Proceedings of the American Society for Information Science and Technology* 50, 1 (2013), 1–11.

[8] Lauren Z. Atkinson and Andrea Cipriani. 2018. How to carry out a literature search for a systematic review: a practical guide. *BJPsych Advances* 24, 2 (March 2018), 74–82. https://doi.org/10.1192/bja.2017.3

[9] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (May 2016), 452–454. https://doi.org/10.1038/533452a Number: 7604 Publisher: Nature Publishing Group.

[10] Monya Baker. 2016. Reproducibility crisis. *Nature* 533, 26 (2016), 353–66.

[11] Michael J. Baker. 2000. Writing a literature review. *The marketing review* 1, 2 (2000), 219–247. ISBN: 1469-347X Publisher: Westburn Publishers Ltd.

[12] Robert E Bartholomew. 2014. Science for sale: the rise of predatory journals. , 384–385 pages.

[13] James Berry, Lucas C Coffman, Douglas Hanley, Rania Gihleb, and Alistair J Wilson. 2017. Assessing the rate of replication in economics. *American Economic Review* 107, 5 (2017), 27–31.

[14] Steven Bethard and Dan Jurafsky. 2010. Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, Toronto ON Canada, 609–618. https://doi.org/10.1145/1871437.1871517

[15] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics* (2023), 1–4. ISBN: 2522-5820 Publisher: Nature Publishing Group UK London.

[16] Martin Boeker, Werner Vach, and Edith Motschall. 2013. Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC medical research methodology* 13, 1 (2013), 1–12. ISBN: 1471-2288 Publisher: BioMed Central.

[17] Lutz Bornmann. 2014. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of informetrics* 8, 4 (2014), 895–903. ISBN: 1751-1577 Publisher: Elsevier.

[18] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. 2020. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582, 7810 (2020), 84–88.

[19] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540* (2022).

[20] Wichor M. Bramer, Gerdien B. de Jonge, Melissa L. Rethlefsen, Frans Mast, and Jos Kleijnen. 2018. A systematic approach to searching: an efficient and complete method to develop literature searches. *Journal of the Medical Library Association : JMLA* 106, 4 (Oct. 2018), 531–541. https://doi.org/10.5195/jmla.2018.283

[21] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[22] Jan vom Brocke, Alexander Simons, Bjoern Niehaves, Bjorn Niehaves, Kai Reimer, Ralf Plattfaut, and Anne Cleven. 2009. Reconstructing the giant: On the importance of rigour in documenting the literature search process. (2009).

[23] Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, and Taizan Chan. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351, 6280 (2016), 1433–1436. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

[24] Colin F Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351, 6280 (2016), 1433–1436.

[25] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature human behaviour* 2, 9 (2018), 637–644.

[26] Daniel N Cassenti and Lance M Kaplan. 2021. Robust uncertainty representation in human-AI collaboration. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, Vol. 11746. SPIE, 249–262.

[27] Julen Cestero, David Velásquez, Elizabeth Suescún, Mikel Maiza, and Marco Quartulli. 2022. Pysurveillance: A Novel Tool for Supporting Researchers in the Systematic Literature Review Process. *Advanced Intelligent Technologies for Industry* (2022), 239–248.

[28] Joseph Chee Chang, Amy X. Zhang, Jonathan Bragg, Andrew Head, Kyle Lo, Doug Downey, and Daniel S. Weld. 2023. CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3580847

[29] Catherine Chen and Carsten Eickhoff. 2022. Evaluating Search Explainability with Psychometrics and Crowdsourcing. *arXiv preprint arXiv:2210.09430* (2022).

[30] Kiroong Choe, Seokweon Jung, Seokhyeon Park, Hwajung Hong, and Jinwook Seo. 2021. Papers101: Supporting the Discovery Process in the Literature Review Workflow for Novice Researchers. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. 176–180. https://doi.org/10.1109/PacificVis52677.2021.00037 ISSN: 2165-8773.

[31] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. 2022. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion* 81 (May 2022), 59–83. https://doi.org/10.1016/j.inffus.2021.11.003

[32] Morris F Cohen. 2011. *An introduction to logic and scientific method.* Read Books Ltd.

[33] Nicki Lisa Cole, Sven Ulpts, Tony Ross-Hellauer, Agata Bochynska, and Thomas Klebel. 2023. Integrative review of conceptions and facilitators of and barriers to reproducibility of qualitative research. (July 2023). https://doi.org/10.17605/OSF.IO/Q4XWK Publisher: OSF.

[34] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

[35] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[36] Christian Collberg, Todd Proebsting, Gina Moraila, Akash Shankaran, Zuoming Shi, and Alex M Warren. 2014. Measuring reproducibility in computer systems research. *Department of Computer Science, University of Arizona, Tech. Rep* 37 (2014).

[37] David Colquhoun. 2017. The reproducibility of research and the misinterpretation of p-values. *Royal society open science* 4, 12 (2017), 171085.

[38] Chris Cooper, Andrew Booth, Jo Varley-Campbell, Nicky Britten, and Ruth Garside. 2018. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. *BMC Medical Research Methodology* 18, 1 (Aug. 2018), 85. https://doi.org/10.1186/s12874-018-0545-3

[39] Florian Cova, Brent Strickland, Angela Abatista, Aurélien Allard, James Andow, Mario Attie, James Beebe, Renatas Berniūnas, Jordane Boudesseul, Matteo Colombo, et al. 2021. Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology* 12 (2021), 9–44.

[40] John W Creswell and Dana L Miller. 2000. Determining validity in qualitative inquiry. *Theory into practice* 39, 3 (2000), 124–130.

[41] Harriet Dashnow, Andrew Lonsdale, and Philip E. Bourne. 2014. Ten Simple Rules for Writing a PLOS Ten Simple Rules Article. *PLOS Computational Biology* 10, 10 (Oct. 2014), e1003858. https://doi.org/10.1371/journal.pcbi.1003858 Publisher: Public Library of Science.

[42] Fabrice De Chaumont, Stéphane Dallongeville, Nicolas Chenouard, Nicolas Hervé, Sorin Pop, Thomas Provoost, Vannary Meas-Yedid, Praveen Pankajakshan, Timothée Lecomte, and Yoann Le Montagner. 2012. Icy: an open bioimage informatics platform for extended reproducible research. *Nature methods* 9, 7 (2012), 690–696. ISBN: 1548-7091 Publisher: Nature Publishing Group US New York.

[43] Nurullah Demir, Matteo Große-Kampmann, Tobias Urban, Christian Wressnegger, Thorsten Holz, and Norbert Pohlmann. 2022. Reproducibility and Replicability of Web Measurement Studies. In *Proceedings of the ACM Web Conference 2022*. ACM, Virtual Event, Lyon France, 533–544. https://doi.org/10.1145/3485447.3512214

[44] Berna Devezer, Luis G Nardin, Bert Baumgaertner, and Erkan Ozge Buzbas. 2019. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PloS one* 14, 5 (2019), e0216125.

[45] Xianghua Ding, Yubo Kou, Yiwen Xu, and Peng Zhang. 2022. "As Uploaders, We Have the Responsibility": Individualized Professionalization of Bilibili Uploaders. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[46] Cody Dunne, Ben Shneiderman, Robert Gove, Judith Klavans, and Bonnie Dorr. 2012. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology* 63, 12 (2012), 2351–2369. https://doi.org/10.1002/asi.22652 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22652.

[47] Charles G. Durbin. 2009. How to read a scientific research paper. *Respiratory care* 54, 10 (2009), 1366–1371. ISBN: 0020-1324 Publisher: Respiratory Care.

[48] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[49] David Eriksson. 2023. The art and science of scholarly writing: framing symmetry of specificity beyond IMRAD. *European Business Review* ahead-of-print (2023).

[50] Timothy M Errington, Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Joelle Lomax, and Brian A Nosek. 2014. An open investigation of the reproducibility of cancer biology research. *Elife* 3 (2014), e04333.

[51] Andres Fortino, Qitong Zhong, Luke Yeh, and Sijia Fang. 2020. Using Text Data Mining to Enhance the Literature Search Process for Novice STEM Researchers. In *2020 IEEE Integrated STEM Education Conference (ISEC)*. IEEE, 1–6.

[52] Erin D. Foster and Ariel Deardorff. 2017. Open science framework (OSF). *Journal of the Medical Library Association: JMLA* 105, 2 (2017), 203. Publisher: Medical Library Association.

[53] Jeremy Freese, Tamkinat Rauf, and Jan Gerrit Voelkel. 2022. Advances in transparency and reproducibility in the social sciences. *Social Science Research* 107 (2022), 102770.

[54] Darcy E Furlong and Jessica Nina Lester. 2023. Toward a Practice of Qualitative Methodological Literature Reviewing. *Qualitative Inquiry* 29, 6 (2023), 669–677.

[55] Niloy Ganguly, Dren Fazlija, Maryam Badar, Marco Fisichella, Sandipan Sikdar, Johanna Schrader, Jonas Wallat, Koustav Rudra, Manolis Koubarakis, Gourab K Patro, et al. 2023. A review of the role of causality in developing trustworthy ai systems. *arXiv preprint arXiv:2302.06975* (2023).

[56] Eugene Garfield. 1977. Proposal for a new profession-scientific reviewer. *Current contents* 14 (1977), 5–8. Publisher: INST SCI INFORM INC 3501 MARKET ST, PHILADELPHIA, PA 19104.

[57] Trisha Greenhalgh. 1997. How to read a paper: Assessing the methodological quality of published papers. *Bmj* 315, 7103 (1997), 305–308. ISBN: 0959-8138 Publisher: British Medical Journal Publishing Group.

[58] Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. Issue: 1.

[59] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (Dec. 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[60] Michael B. Harari, Heather R. Parola, Christopher J. Hartwell, and Amy Riegelman. 2020. Literature searches in systematic reviews and meta-analyses: A review, evaluation, and recommendations. *Journal of Vocational Behavior* 118 (April 2020), 103377. https://doi.org/10.1016/j.jvb.2020.103377

[61] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting Scientific Papers

with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445648

[62] Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. The extent and consequences of p-hacking in science. *PLoS biology* 13, 3 (2015), e1002106. ISBN: 1545-7885 Publisher: Public Library of Science.

[63] Elisa L Hill-Yardin, Mark R Hutchinson, Robin Laycock, and Sarah J Spencer. 2023. A Chat (GPT) about the future of scientific publishing. *Brain Behav Immun* 110 (2023), 152–154.

[64] Andreas Hinderks, Francisco José Domínguez Mayo, Jörg Thomaschewski, and María José Escalona. 2020. An SLR-tool: Search process in practice: A tool to conduct and manage systematic literature review (SLR). In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Companion Proceedings*. 81–84.

[65] Xiaoyun Huang, Jessica Vitak, and Yla Tausczik. 2020. " You Don't Have To Know My Past": How WeChat Moments Users Manage Their Evolving Self-Presentation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing systems*. 1–13.

[66] Martin E. Héroux, Annie A. Butler, Aidan G. Cashin, Euan J. McCaughey, Andrew J. Affleck, Michael A. Green, Andrew Cartwright, Matthew Jones, Kim M. Kiely, Kimberley S. van Schooten, Jasmine C. Menant, Michael Wewege, and Simon C. Gandevia. 2022. Quality Output Checklist and Content Assessment (QuOCCA): a new tool for assessing research quality and reproducibility. *BMJ Open* 12, 9 (Sept. 2022), e060976. https://doi.org/10.1136/bmjopen-2022-060976 Publisher: British Medical Journal Publishing Group Section: Communication.

[67] Sharon Favaro Ince, Christopher Hoadley, and Paul A Kirschner. 2018. A study of search practices in doctoral student scholarly workflows. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 245–248.

[68] Srinivasan Keshav. 2007. How to read a paper. *ACM SIGCOMM Computer Communication Review* 37, 3 (2007), 83–84. ISBN: 0146-4833 Publisher: ACM New York, NY, USA.

[69] Mallory C. Kidwell, Ljiljana B. Lazarević, Erica Baranski, Tom E. Hardwicke, Sarah Piechowski, Lina-Sophia Falkenberg, Curtis Kennett, Agnieszka Slowik, Carina Sonnleitner, and Chelsey Hess-Holden. 2016. Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS biology* 14, 5 (2016), e1002456. ISBN: 1544-9173 Publisher: Public Library of Science San Francisco, CA USA.

[70] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. " Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[71] Seth A. King, Douglas Kostewicz, Olivia Enders, Taneal Burch, Argnue Chitiyo, Johanna Taylor, Sarah DeMaria, and Milsha Reid. 2020. Search and Selection Procedures of Literature Reviews in Behavior Analysis. *Perspectives on Behavior Science* 43, 4 (Dec. 2020), 725–760. https://doi.org/10.1007/s40614-020-00265-9

[72] Sai Koneru, Jian Wu, and Sarah Rajtmajer. 2023. Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences. *arXiv preprint arXiv:2309.06578* (2023).

[73] Janice Y. Kung. 2023. Elicit. *The Journal of the Canadian Health Libraries Association* 44, 1 (April 2023), 15–18. https://doi.org/10.29173/jchla29657

[74] Tamarinde L. Haven and Dr. Leonie Van Grootel. 2019. Preregistering qualitative research. *Accountability in Research* 26, 3 (April 2019), 229–244. https://doi.org/10.1080/08989621.2019.1580147 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08989621.2019.1580147.

[75] Manoj Mathew Lalu, Larissa Shamseer, Kelly D Cobey, and David Moher. 2017. How stakeholders can respond to the rise of predatory journals. *Nature Human Behaviour* 1, 12 (2017), 852–855.

[76] Sabina Leonelli. 2018. Rethinking Reproducibility as a Criterion for Research Quality. In *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise*. Research in the History of Economic Thought and Methodology, Vol. 36B. Emerald Publishing Limited, 129–146. https://doi.org/10.1108/S0743-41542018000036B009

[77] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (Jan. 2021), 18. https://doi.org/10.3390/e23010018 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[78] Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848* (2023).

[79] Shengbo Liu, Chaomei Chen, Kun Ding, Bo Wang, Kan Xu, and Yuan Lin. 2014. Literature retrieval based on citation context. *Scientometrics* 101, 2 (Nov. 2014), 1293–1307. https://doi.org/10.1007/s11192-014-1233-7

[80] Matthew C. Makel, Jonathan A. Plucker, and Boyd Hegarty. 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science* 7, 6 (2012), 537–542. ISBN: 1745-6916 Publisher: Sage Publications Sage CA: Los Angeles, CA.

[81] Alan Maloney and Lettie Y Conrad. 2016. Expecting the unexpected: Serendipity, discovery, and the scholarly research process. *White Paper* (2016).

[82] Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How does domain expertise affect users' search interaction and outcome in exploratory search? *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 1–30.

[83] Alberto Martín-Martín, Rodrigo Costas, Thed Van Leeuwen, and Emilio Delgado López-Cózar. 2018. Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of informetrics* 12, 3 (2018), 819–841. ISBN: 1751-1577 Publisher: Elsevier.

[84] Erin C. McKiernan, Philip E. Bourne, C. Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A. Nosek, Karthik Ram, and Courtney K. Soderberg. 2016. How open science helps researchers succeed. *elife* 5 (2016), e16800. ISBN: 2050-084X Publisher: eLife Sciences Publications, Ltd.

[85] Blakeley B McShane, Jennifer L Tackett, Ulf Böckenholt, and Andrew Gelman. 2019. Large-scale replication projects in contemporary psychological research. *The American Statistician* 73, sup1 (2019), 99–105.

[86] Wondimagegn Mengist, Teshome Soromessa, and Gudina Legese. 2020. Method for conducting systematic literature review and meta-analysis for environmental science research. *MethodsX* 7 (2020), 100777.

[87] Marcin Miłkowski, Witold M. Hensel, and Mateusz Hohol. 2018. Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience* 45, 3 (Dec. 2018), 163–172. https://doi.org/10.1007/s10827-018-0702-z

[88] Marcus R. Munafò, Brian A. Nosek, Dorothy VM Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour* 1, 1 (2017), 1–9. ISBN: 2397-3374 Publisher: Nature Publishing Group.

[89] Nishanth Nakshatri, Arjun Menon, C Lee Giles, Sarah Rajtmajer, and Christopher Griffin. 2021. Design and Analysis of a Synthetic Prediction Market using Dynamic Convex Sets. *arXiv preprint arXiv:2101.01787* (2021).

[90] National Academies of Sciences, Engineering, and Medicine et al. 2019. *Reproducibility and replicability in science*. National Academies Press.

[91] Brian A Nosek, Tom E Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S Corker, Anna Dreber Almenberg, Fiona Fidler, Joseph Hilgard, Melissa Kline, Michèle B Nuijten, et al. 2021. Replicability, robustness, and reproducibility in psychological science. (2021).

[92] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery* 88 (2021), 105906.

[93] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3580841

[94] Samuel Pawel and Leonhard Held. 2020. Probabilistic forecasting of replication studies. *PloS one* 15, 4 (2020), e0231416.

[95] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving Reproducibility in Machine Learning Research. *Journal of Machine Learning Research* 22 (2021).

[96] Hans E. Plesser. 2018. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics* 11 (2018). https://www.frontiersin.org/articles/10.3389/fninf.2017.00076

[97] Diego Ponte and Judith Simon. 2011. Scholarly Communication 2.0: Exploring Researchers' Opinions on Web 2.0 for Scientific Knowledge Creation, Evaluation and Dissemination. *Serials Review* 37, 3 (Sept. 2011), 149–156. https://doi.org/10.1080/00987913.2011.10765376 Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00987913.2011.10765376.

[98] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 379–396.

[99] Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems* 32 (2019).

[100] Sarah Rajtmajer, Christopher Griffin, Jian Wu, Robert Fraleigh, Laxmaan Balaji, Anna Squicciarini, Anthony Kwasnica, David Pennock, Michael McLaughlin, Timothy Fritton, et al. 2021. A Synthetic Prediction Market for Estimating Confidence in Published Work. *arXiv preprint arXiv:2201.06924* (2021).

[101] Sarah Rajtmajer, Christopher Griffin, Jian Wu, Robert Fraleigh, Laxmaan Balaji, Anna Squicciarini, Anthony Kwasnica, David Pennock, Michael McLaughlin, Timothy Fritton, et al. 2022. A synthetic prediction market for estimating confidence in published work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 13218–13220.

[102] Melissa L Rethlefsen, Shona Kirtley, Siw Waffenschmidt, Ana Patricia Ayala, David Moher, Matthew J Page, and Jonathan B Koffel. 2021. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic reviews* 10 (2021), 1–19.

[103] Wojciech Samek and Klaus-Robert Müller. 2019. Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (Eds.). Springer International Publishing, Cham, 5–22. https://doi.org/10.1007/978-3-030-28954-6_1

[104] Frances Shiely, Kerrie Gallagher, and Seán R Millar. 2024. How, and why, science and health researchers read scientific (IMRAD) papers. *Plos one* 19, 1 (2024), e0297034.

[105] Luciana B. Sollaci and Mauricio G. Pereira. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the medical library association* 92, 3 (2004), 364. Publisher: Medical Library Association.

[106] Piotr Sorokowski, Emanuel Kulczycki, Agnieszka Sorokowska, and Katarzyna Pisanski. 2017. Predatory journals recruit fake editor. *Nature* 543, 7646 (2017), 481–483.

[107] Marek Sośnicki and Lech Madeyski. 2021. ASH: A New Tool for Automated and Full-Text Search in Systematic Literature Reviews. In *International Conference on Computational Science*. Springer, 362–369.

[108] Ayah Soufan, Ian Ruthven, and Leif Azzopardi. 2022. Searching the literature: an analysis of an exploratory search task. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 146–157.

[109] Nicole Sultanum, Christine Murad, and Daniel Wigdor. 2020. Understanding and Supporting Academic Literature Review Workflows with LitSense. In *Proceedings of the International Conference on Advanced Visual Interfaces*. ACM, Salerno Italy, 1–5. https://doi.org/10.1145/3399715.3399830

[110] Katherine A. Tamminen and Zoë A. Poucher. 2018. Open science in sport and exercise psychology: Review of current approaches and considerations for qualitative inquiry. *Psychology of Sport and Exercise* 36 (2018), 17–28. https://doi.org/10.1016/j.psychsport.2017.12.010 Place: Netherlands Publisher: Elsevier Science.

[111] Carol Tenopir, Lisa Christian, and Jordan Kaufman. 2019. Seeking, Reading, and Use of Scholarly Articles: An International Study of Perceptions and Behavior of Researchers. *Publications* 7, 1 (March 2019), 18. https://doi.org/10.3390/publications7010018 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[112] Xuanhui Wang and ChengXiang Zhai. 2007. Learn from web search logs to organize search results. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 87–94.

[113] Ronald L Wasserstein and Nicole A Lazar. 2016. The ASA statement on p-values: context, process, and purpose. , 129–133 pages.

[114] Ryen W White. 2018. Opportunities and challenges in search interaction. *Commun. ACM* 61, 12 (2018), 36–38.

[115] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*. 132–141.

[116] Jian Wu, Rajal Nivargi, Sree Sai Teja Lanka, Arjun Manoj Menon, Sai Ajay Modukuri, Nishanth Nakshatri, Xin Wei, Zhuoer Wang, James Caverlee, Sarah M Rajtmajer, et al. 2021. Predicting the Reproducibility of Social and Behavioral Science Papers Using Supervised Learning Models. *arXiv preprint arXiv:2104.04580* (2021).

[117] Erjia Yan and Ying Ding. 2010. Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology* 61, 8 (2010), 1635–1643. ISBN: 1532-2882 Publisher: Wiley Online Library.

[118] Yang Yang, Wu Youyou, and Brian Uzzi. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* 117, 20 (2020), 10762–10768.

[119] Yifan Zhu, Qika Lin, Hao Lu, Kaize Shi, Ping Qiu, and Zhendong Niu. 2021. Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks. *Knowledge-Based Systems* 215 (March 2021), 106744. https://doi.org/10.1016/j.knosys.2021.106744