Report on the Dagstuhl Seminar on Frontiers of Information Access Experimentation for Research and Education

Christine Bauer
Paris Lodron University Salzburg, Austria
christine.bauer@plus.ac.at

Ben Carterette
University of Delaware and Spotify, United States
carteret@acm.org

Nicola Ferro
University of Padua, Italy
nicola.ferro@unipd.it

Norbert Fuhr
University of Duisburg-Essen, Germany
norbert.fuhr@uni-due.de

Joeran Beel, Timo Breuer, Charles L. A. Clarke, Anita Crescenzi, Gianluca Demartini, Giorgio Maria Di Nunzio, Laura Dietz, Guglielmo Faggioli, Bruce Ferwerda, Maik Fröbe, Matthias Hagen, Allan Hanbury, Claudia Hauff, Dietmar Jannach, Noriko Kando, Evangelos Kanoulas, Bart P. Knijnenburg, Udo Kruschwitz, Meijie Li, Maria Maistro, Lien Michiels, Andrea Papenmeier, Martin Potthast, Paolo Rosso, Alan Said, Philipp Schaer, Christin Seifert, Damiano Spina, Benno Stein, Nava Tintarev, Julián Urbano, Henning Wachsmuth, Martijn C. Willemsen, Justin Zobel

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23031 "Frontiers of Information Access Experimentation for Research and Education", which brought together 38 participants from 12 countries. The seminar addressed technology-enhanced information access (information retrieval, recommender systems, natural language processing) and specifically focused on developing more responsible experimental practices leading to more valid results, both for research as well as for scientific education.

The seminar featured a series of long and short talks delivered by participants, who helped in setting a common ground and in letting emerge topics of interest to be explored as the main output of the seminar. This led to the definition of five groups which investigated challenges, opportunities, and next steps in the following areas: *reality check, i.e. conducting real-world studies*, *human–machine-collaborative relevance judgment frameworks*, *overcoming methodological challenges in information retrieval and recommender systems through awareness and education*, *results-blind reviewing*, and *guidance for authors*.

Date: 15–20 January 2023.

Website: https://www.dagstuhl.de/23031.

1 Introduction

Information access—which includes Information Retrieval (IR), Recommender Systems (RS), and Natural Language Processing (NLP)—has a long tradition of relying heavily on experimental evaluation, dating back to the mid-1950s, a tradition that has driven the research and evolution of the field. However, nowadays, research and development of information access systems are confronted with new challenges: information access systems are called to support a much wider set of user tasks (informational, educational, and entertainment, just to name a few) which are increasingly challenging, and as a result, research settings and available opportunities have evolved substantially (e.g., better platforms, richer data, but also developments within the scientific culture) and shape the way in which we do research and experimentation.

Therefore, we face two problems: Can we re-innovate how we do research and experimentation in the field by addressing emerging challenges in experimental processes to develop the next generation of information access systems? How can a new paradigm of experimentation be leveraged to improve education to give an adequate basis to the new generation of researchers and developers?

The Dagstuhl Seminar 23031 on "Frontiers of Information Access Experimentation for Research and Education" brought together experts from various sub-fields of information access to create a joint understanding of the problems and challenges presented above, to discuss existing solutions and impediments, and to propose next steps to be pursued in the area [Bauer et al., 2023a].

We started the seminar week with a series of long and short talks delivered by participants. This helped in setting a common ground and understanding and in letting emerge the topics and themes that participants wished to explore as the main output of the seminar.

This led to the definition of five groups which explored challenges, opportunities, and next steps in the following areas:

- **Reality check**: Ferwerda, Hanbury, Knijnenburg, Larsen, Michiels, Papenmeier, Said, Schaer, and Willemsen [2023] identified the main challenges in doing real-world studies in RS and IR research – and points to best practices and remaining challenges in both how to do domain-specific or longitudinal studies, how to recruit the right participants, using existing or creating new infrastructure including appropriate data representation, as well as how, why and what to measure.
- **Human-machine-collaborative relevance judgment frameworks**: Clarke, Demartini, Dietz, Faggioli, Hagen, Hauff, Kando, Kanoulas, Potthast, Soboroff, Stein, and Wachsmuth [2023] studied the motivation for using Large Language Models (LLMs) to automatically generate relevance assessments in information retrieval evaluation, and raises research questions about how LLMs can help human assessors with the assessment task, whether machines can replace humans in assessing and annotating, and what are the conditions under which human assessors cannot be replaced by machines.
- **Overcoming methodological challenges in IR and RS through awareness and education**: Given the potential limitations of today's predominant experimentation practices, Bauer, Fröbe, Jannach, Kruschwitz, Rosso, Spina, and Tintarev [2023b] discussed the need to better equip the various actors in the scientific ecosystem in terms of scientific methods, and identified a corresponding set of helpful resources and initiatives, which will allow them to adopt a more holistic perspective when evaluating such systems.

- **Results-blind reviewing**: The current review processes lead to undue emphasis on performance, rejecting papers focusing on insights in case they show no performance improvements. Beel, Breuer, Crescenzi, Fuhr, and Li [2023] proposed to introduce a results-blind reviewing process forcing reviewers to put more emphasis on the theoretical background, the hypotheses, the methodological plan and the analysis plan of an experiment, thus improving the scientific quality of the papers being accepted.
  This proposal has been then taken up by the CLEF 2023 conference which, as an experiment and a first attempt to put these ideas into practice, introduced a new submission and review model, based on two stages, i.e. a methodology and and experimental stage[1].
- **Guidance for authors**: The Information Retrieval community has over time developed expectations regarding papers, but these expectations are largely implicit. In contrast to adjacent disciplines, efforts in the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) community have been rather sparse and are mostly due to individuals expressing their own views. Drawing on materials from other disciplines, Di Nunzio, Maistro, Seifert, Urbano, and Zobel [2023] built a draft set of guidelines with the aim of them being understandable, broad, and highly concise. The working group believes that their proposal is general and uncontroversial, can be used by the main venues, and can be maintained with an open and continuous effort driven by, and for, the community.

In the following sections, we report one by one the outcomes and recommendations of the above working groups. For each working group, we discuss their motivations, their proposals, and some next steps and recommendations. Finally, the last section reports the complete list of participants who attended the seminar and contributed to the full report [Bauer et al., 2023a].

2 Reality Check—Conducting Real World Studies

As information retrieval and recommender systems are deployed in real world environments, we should study their characteristics in "real world studies". This raises the question: What does it mean for a study to be realistic? Does it mean the user has to be a real user of the system or can anyone participate in a study of the system? Does it mean the system needs to be perceived as realistic by the user? Does it mean the manipulations need to be perceived as realistic by the user?

Arguably, the most realistic users can be found on existing systems, which will typically have a sufficiently large user base. However, this raises some additional questions. Firstly, how to sample from this user base to obtain a representative sample. Secondly, these users may have some expectations of the system, which may make them somewhat resistant to (drastic) changes. On the other hand, recruiting new users comes with its own set of challenges.

In a similar vein, the largest degree of "system realism" would be achieved by studying real users of an existing system. For example, log-based studies have been considered the best examples of real world studies [Kelly, 2007] since they capture behavior in a real-life setting, with little chance of contamination or bias. However, this limits the amount of control we, as researchers, can exert, and thus the research questions we can pose and answer. On the other hand, highly controlled

[1] https://clef2023.clef-initiative.eu/index.php?page=Pages/cfp.html

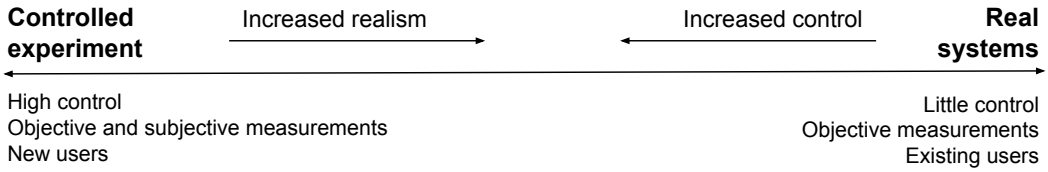

Figure 1. Control versus realism continuum

experiments might lack realism in terms of the system, the user experience (users knowing they are being studied) and the generalizability of the study. Realism in a study is a continuum, as illustrated in Figure 1, ranging from highly controlled experiments towards real systems with real users, and researchers need to identify the appropriate experiment type for their purpose [Zangerle and Bauer, 2022].

One central question in running real world studies is the influence of measurements on the behavior and experience of users. Following the Heisenberg principle [Heisenberg, 1927], it is impossible to measure without influencing. If we study existing users in an existing system, and only use behavioral measures and logs from the system we will not affect users much but it will be hard to answer our question, as the evaluation of our manipulation will be difficult. On the other hand, when we start collecting additional measures, like intermediate surveys, users will know they are part of a study and modify their behavior because of that (Hawthorne effect [Schwartz et al., 2013]). Also, longer surveys might break the actual flow of system usage and demotivate people. Survey questions might provide the users with insights into the underlying research questions, resulting in unwanted demand characteristics or socially desirable answer patterns.

However, triangulating objective (behavioral) data with subjective measures will be crucial to understand how users experience the system [Knijnenburg et al., 2012], so a careful development and usage of a combination of subjective and objective measures is going to be central to balancing realism with adequate measurement.

Then, we have the realism of the research question and experiment design. In any experiment, we manipulate the system, thus breaking some existing habits or patterns. Especially when studying users of an existing system, the realism of this manipulation is crucial. If users do not experience the manipulation as a realistic feature or implementation, the results may not be representative. Similarly, the degree of information given to the user may also influence the realism of the study. If we provide users with too much information, e.g., a very specific task and scenario to work from, users may perform actions they would not have in a realistic situation. On the other hand, if we provide too little information, e.g., when we introduce a new feature on an existing platform without any instruction, we require users to invest the time and effort to find out how the feature works before they can use it in the way we intended.

Another important consideration regarding experiment design is the assignment of users to different versions of a system. Should the experience of a single user be kept consistent throughout the entire study? Such between-subjects designs have the advantage of preventing any spill-over effects but users working side by side or communicating about the system might discover there are different versions of the system, accidentally revealing the experimental conditions and goals. Within-subject designs allow users to experience all experimental conditions, which increases statistical power (as we can control for participant variance) but ordering and spill-over effects have

to be considered. Moreover, to make a real world study sufficiently realistic and also understand how behavior changes over time and how habits are formed, we will need to consider longitudinal studies which come with their own set of challenges.

Even when we carefully design our experiments and research questions and select the appropriate participants, we may arrive at conclusions that do not necessarily generalize beyond the domain.

Finally, the cost of running a real world study is typically many times higher than performing offline evaluation [Zangerle and Bauer, 2022]. Therefore it is important to also consider the available research infrastructure, and promote the development of reusable research infrastructure and provide datasets in sufficiently general formats to promote reuse.

Next Steps and Recommendations

The following steps should be taken to carefully determine the **goals** of conducting real world studies:

- Classify domains by knowledge task types
- Establish context-specific evaluation targets
- Carefully consider users' information needs when conducting studies
- Develop a checklist of sample characteristics and user task details that should be collected and reported for each study

The following **resources** would expedite the design, execution and evaluation of real world studies:

- Provide researchers with access to flexible real world research infrastructure
- Obtain sufficiently large and rich content corpora that can be used in real world studies
- Create a repository of validated measurement scales
- Standardize practices for scale development
- Establish effective recruitment methods to find the "right" participants for a study
- Develop metrics that are as unobtrusive as possible to measure
- Design standardized but flexible ways to represent the data and meta-data collected in real world studies
- Study effective ways to limit attrition in longitudinal studies
- Produce best-practices guidelines for developing real world systems, getting infrastructures up and running, maintaining them and providing support for both users and researchers
- Establish guidelines to protect the privacy of research participants

The following steps must be taken to allow researchers to **integrate the findings of real world studies into generalizable knowledge**:

- Use theory to integrate domain-specific knowledge into a generalized knowledge
- Define a theoretical framework for measurement
- Develop an infrastructure for researchers to contribute analyses of and insights about real world datasets in a centralized manner

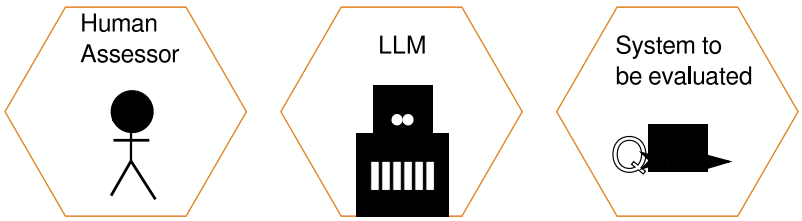

Figure 2. The three most relevant components in our system: the human assessor, the Large Language Model (LLM) that can help humans or replace them in annotating documents for relevance, and the system that we want to evaluate using the newly produced relevance judgements.

- Integrate research within specific domains as well as at the generalized knowledge level using systematic reviews, meta-analyses, task-specific workshops and domain-specific workshops
- Conduct studies to triangulate qualitative and quantitative insights, behavioral and subjective metrics, and short-term and long-term metrics

A more detailed discussion can be found in the full report [Ferwerda et al., 2023].

3 HMC: A Spectrum of Human–Machine-Collaborative Relevance Judgment Frameworks

IR evaluation traditionally needs human assessors to generate relevance judgements. Traditionally, human assessors are asked to judge the relevance of a document with respect to a topic [Harman, 2011]. Recently, work looking at preference judgements [Clarke et al., 2021; Potthast et al., 2019] has looked at research questions related to how to best evaluate IR systems by asking human assessors which of two results is the better given an information need. The recent availability of LLMs has opened the possibility to use them to automatically generate relevance assessments in the form of preference judgements. While the idea of automatically generated judgements has been looked at before [Büttcher et al., 2007], new-generation LLMs drive us to re-ask the question of whether human assessors are still necessary.

New models tend to fail in a different and more diverse way compared to traditional approaches. Failure points for old models were more uniform and clear, with new systems it is harder to predict in which ways the model will fail. In most cases, LLMs (especially for what concerns generative aspects) focus on entertainment tasks. Models tend to report false facts in such a convincing way that they need to be carefully read by some expert to identify lacking factuality (e.g., Michel Foucault simulation[2]).

Our motivation to investigate the possibility of using LLMs in order to provide automatic annotations stems from some fundamental research questions that can be summarized as follows.

- **RQ1**: In which way automatic approaches, and in particular LLMs, can help assessors with the assessment task to yield the most reliable annotations while improving the efficiency of the annotation process? This question raises other interesting related inquiries. For example, if we were to build such a mixed human-machine annotation paradigm, which held

[2] https://www.youtube.com/watch?v=L6c0xeAqEz4E

out (not provided to the IR system) supporting information about the topic would yield the best and fastest annotations? What weighting between human and LLMs and AI-assisted annotations is ideal?
- **RQ2**: Can machines (either in the form of LLMs or in general as Artificial Intelligence (AI) models) replace humans in assessing and annotating? This question raises also concerns about what annotation target (e.g., relevance labeling, summarization, paragraph highlighting, exam questions [Sander and Dietz, 2021]) would yield the best and fastest annotations.
- **RQ3**: What are the conditions under which human assessors cannot be replaced by machines? Alternatively, in which role can the Human assessor most productively provide relevance assessments?

Answering the questions mentioned would also require finding viable solutions for a set of additional questions and open issues that touch a number of IR evaluation process steps.

- Assessors And Collections:
  - How to use LLM to help assessors: some examples of possible usages include, summarising text, associating keywords and identifying the content of long podcasts to help assessors annotate the documents, for example by highlighting relevant fragments of text/podcast or segments with correct answers.
  - What is the effective role of the human assessor in annotating material for generative models? Should the annotator provide input at the beginning of the pipeline, by annotating the original documents, or are they more useful downstream, after the task has been carried out?
  - Generative models can be used to create new collections: corpora, conversations, queries, abstracts and so on.

- LLM and generative models to retrieve information in a broader sense:
  - IR tasks that employ LLMs have the means to provide more details: often a single answer is not satisfactory for the user. How to support the user in exploring the results further (for example via links and connected pages). Generative models can help, but is this helpful when the model simply generates the response without knowing where it comes from? In many cases, the user is not interested in receiving only the direct/short answer, but rather in seeing which documents contain it and related pieces of information to expand their knowledge.

- LLMs as an evaluation tool:
  - The model is biased: how can we use it to evaluate itself? If a model has been trained on biased data, then also the evaluation is prone to the same biases. How to detect and account for such biases?

- Evaluating LLMs and their trustworthiness:
  - Can we find a way to understand and measure to what level we can trust the results of a generative model?
  - How to carry out fact-checking, for example by identifying the source of information of a generative model and verifying that it is presented accurately.

Prompt	Select the passage that answers the question better. - If both answers are similar, select the one with the least extraneous information. - If both answers are still similar, select the one with the best formatting. Question: What are anemia's possible causes? Passage 1: If you have anemia, your blood does not carry enough oxygen to the rest of your body. The most common cause of anemia is not having enough iron. Your body needs iron to make hemoglobin. Hemoglobin is an iron-rich protein that gives the red color to blood. It carries oxygen from the lungs to the rest of the body. Anemia has three main causes: blood loss, lack of red blood cell production, and high rates of red blood cell destruction Passage 2: Broadly, causes of anemia may be classified as impaired red blood cell (RBC) production, increased RBC destruction (hemolytic anemias), blood loss and fluid overload (hypervolemia
ChatGPT	Passage 1: If you have anemia, your blood does not carry enough oxygen to the rest of your body. The most common cause of anemia is not having enough iron. Your body needs iron to make hemoglobin. Hemoglobin is an iron-rich protein that gives the red color to blood. It carries oxygen from the lungs to the rest of the body. Anemia has three main causes: blood loss, lack of red blood cell production, and high rates of red blood cell destruction.
YouChat	Passage 1 is the better answer, as it provides an explanation of what anemia is, and then lists the three main causes in detail, which is the most relevant information for the question. Passage 2, while providing some good background information, does not provide the same level of specificity as Passage 1, and includes extraneous information that is not directly relevant to the question.

Figure 3. Response of two well-known LLMs in comparing two passages as answer for the query 54_8 from CAsT '19: "What are anemia's possible causes?"

– Distinguish between human and machine-generated data: Important for many tasks, such as journalism, where it is of uttermost importance to verify the information. Human-generated data is more trusted.

We argue that the collaboration between humans and Machine Learning (ML), especially under the form of LLMs, could be abstracted in the form of a spectrum. On the two extremes of this spectrum, we have either the human or the machine entirely tasked to annotate documents for relevance with respect to a query. Within the spectrum, humans and LLMs interact to a different extent. Theoretically, such a spectrum corresponds also to moving from highly expansive annotations in terms of human effort, cost and time, but with high-quality annotations, to a much less expensive annotation procedure with also a decreased annotation quality. We also argue

Prompt	Select the passage that answers the question better. - If both answers are similar, select the one with the least extraneous information. - If both answers are still similar, select the one with the best formatting. Question: What is a DC half smoke? Passage 1: However, celebrities at Ben's are not limited to American citizens. In 1998, former DC mayor Marion Barry described having traveled to Ghana and meeting the Mayor of Accra, an alumnus of Howard University, whose greeting was, "Glad to have you in Accra. Is Ben's Chili Bowl still there?" And when French President Nicolas Sarkozy and his wife Carla Bruni-Sarkozy visited Washington, D.C. in March 2010, they reportedly each had two of Ben's half smokes during their visit to the restaurant. Passage 2: A half-smoke is a "local sausage delicacy" found in Washington, D.C. and the surrounding region. Similar to a hot dog, but usually larger, spicier, and with more coarsely-ground meat, the sausage is often half-pork and half-beef, smoked, and served with herbs, onion, and chili sauce.
ChatGPT	Passage 2: A half-smoke is a local sausage delicacy found in Washington, D.C. and the surrounding region. Similar to a hot dog, but usually larger, spicier, and with more coarsely-ground meat, the sausage is often half-pork and half-beef, smoked, and served with herbs, onion, and chili sauce.
YouChat	Passage 2 is the better answer as it more directly answers the question and does not include any extraneous information.

Figure 4. Response of two well-known LLMs in comparing two passages as answer for the query 67_8 from CAsT '19: "What is a DC half smoke?"

that something exists beyond the spectrum; it corresponds to the scenario in which the machine overcomes the human, by producing relevance judgments without any form of bias. We observed this phenomenon happening already in several tasks and scenarios, and therefore we can aspect this to happen also with respect to the construction of the relevance judgments.

Next Steps and Recommendations

Figures 3 and 4 report two examples of document annotation done with two well-known LLMs: YouChat[3] and ChatGPT[4]. It is interesting to notice that, in both cases, both models provided the correct answer, correctly identifying the passage which was annotated as more relevant. It is possible to observe that, while ChatGPT simply repeats the relevant passage, YouChat is capable of correctly identifying the reason why a passage is more relevant than the other.

[3] https://you.com/
[4] https://chat.openai.com/

To assess the feasibility of the proposed approaches, next steps would include an experimental comparison of the different Collaborative-Human-Machine paradigms. This should include multiple test collections (e.g., TREC-8 and TREC Deep Learning), multiple types of judgments (e.g., binary, graded, preference), and multiple models (e.g., GPT-2, GPT-3, chatGPT, etc.). Comparison between human-generated judgments and machine-generated judgments may be performed both using inter-assessor agreement metrics as well as IR system ranking correlation methods.

A more detailed discussion can be found in the full report [Clarke et al., 2023]. Since the workshop, a follow-up article about the study, conclusions, and expert opinions was published [Faggioli et al., 2023].

4 Overcoming Methodological Challenges in Information Retrieval and Recommender Systems through Awareness and Education

In recent years, we have observed a substantial increase in research in IR and RS. To a large extent, this increase is fueled by progress in ML (deep learning) technology. As a result, countless papers are nowadays published each year which report that they improved the state-of-the-art when adopting common experimental procedures to evaluate ML based systems. However, a number of issues were identified in the past few years regarding these reported findings and their interpretation. For example, both in IR and RS, studies point to methodological issues in *offline* experiments, where researchers for example compare their models against weak or non-optimized baselines or where researchers optimize their models on test data rather than on held-out validation data [Armstrong et al., 2009; Ferrari Dacrema et al., 2019; Sun et al., 2020; Yang et al., 2019].

Besides these issues in offline experiments, questions concerning the *ecological validity* of the reported findings are raised increasingly. Ecological validity measures how generalizable experimental findings are to the real world. An example of this problem in information retrieval is the known problem of mismatch between offline effectiveness measurement and user satisfaction measured with online experimentation [Chen et al., 2017; Hassan et al., 2010; Mao et al., 2016; Sanderson et al., 2010; Zhang et al., 2020] or when the definition of relevance does not consider the effect on a searcher and their decision-making. For example, the order of search results, and the viewpoints represented therein, can shift undecided voters toward any particular candidate if high-ranking search results support that candidate [Epstein and Robertson, 2015]. This phenomenon—often referred to as the *Search Engine Manipulation Effect (SEME)* —has been demonstrated for both politics [Epstein and Robertson, 2015; Epstein et al., 2017] and health [Allam et al., 2014; Pogacar et al., 2017]. By being aware of the phenomena, methods have been adapted to measure its presence [Draws et al., 2023, 2021a], and studies to evaluate when and how it affects human decision-makers [Draws et al., 2021b]. Similar questions of ecological validity were also raised in the RS field regarding the suitability of commonly used computational accuracy metrics as predictors of the impact and value such systems have on users in the real world. Several studies indeed indicate that the outcomes of offline experiments are often *not* good proxies of real-world performance indicators such as user satisfaction, engagement, or revenue [Beel and Langer, 2015; Gomez-Uribe and Hunt, 2016; Jannach and Bauer, 2020].

Overall, these observations point to a number of open challenges in how experimentation is predominantly done in the field of information access systems. Ultimately, this leads to the questions of *(i)* how much progress we really make despite the large number of research works that are published every year [Armstrong et al., 2009; Lin et al., 2021; Zobel, 2023] and *(ii)* how effective we are in sharing and translating the knowledge we currently have for doing IR and RS experimentation [Ferro and Sanderson, 2022; Sakai, 2018]. One major cause for the mentioned issues, for example, seems to lie in the somewhat narrow way we tend to evaluate information retrieval and recommender systems: primarily based on various computational effectiveness measures. In reality, information access systems are interactive systems used over longer periods of time, i.e., they may only be assessed holistically if the user's perspective (task and context) is taken into account, cf. [Lykke et al., 2022; White, 2016; Zangerle and Bauer, 2022]. Studies on long-term impact furthermore need to consider the wider scope of stakeholders [Bauer and Zangerle, 2019; Jannach and Bauer, 2020]. Moreover, for several types of information access systems, the specific and potentially competing interests of multiple stakeholders have to be taken into account [Bauer and Zangerle, 2019]. Typical stakeholders in a recommendation scenario include not only the consumers who receive recommendations but also recommendation service providers who for example want to maximize their revenue through the recommendations [Jannach and Adomavicius, 2017; Jannach and Bauer, 2020].

Various factors contribute to our somewhat limited view of such systems, e.g., the difficulties of getting access to real systems and real-world data for evaluation purposes. Unfortunately, the IR and RS research communities to a certain extent seem to have accepted to live with the limitations of the predominant evaluation practices of today. Even more worryingly, the described narrow evaluation approach has become more or less a standard in the scientific literature, and there is not much debate and—as we believe—sometimes even limited awareness of the various limitations of our evaluation practices.

There seems to be no easy and quick way out of this situation, even though some of the problems are known for many years now [Ekstrand et al., 2011; Hassan et al., 2010; Konstan and Adomavicius, 2013; Sanderson et al., 2010]. However, we argue that improved *education* of the various actors in the research ecosystem (including students, educators, and scholars) is one key approach to improve our experimentation practices and ensure real-world impact in the future. As will be discussed in the next sections, better training in experimentation practices is not only important for students, but also for academic teachers, research scholars, practitioners and different types of decision-makers in academia, business, and other organizations. This will, in fact, help address the much broader problems of reproducibility[5] and replicability[6] that we face in Computer Science [Cockburn et al., 2020; Freire et al., 2016] in general and in AI in particular [Gundersen and Kjensmo, 2018].

Finally, Combining and integrating resources and efforts in novel ways has the potential to reduce or even remove barriers between research and education, ultimately enabling Humboldt's ideal to combine teaching and research. Students who participate in shared tasks as part of their curriculum already go in this direction [Elstner et al., 2023]. Continuously maintaining and promoting the integration of test collections and up-to-date best practices for shared tasks into a

[5] https://www.wired.com/story/machine-learning-reproducibility-crisis/

[6] https://cacm.acm.org/magazines/2020/8/246369-threats-of-a-replication-crisis-in-empirical-computer-science/abstract

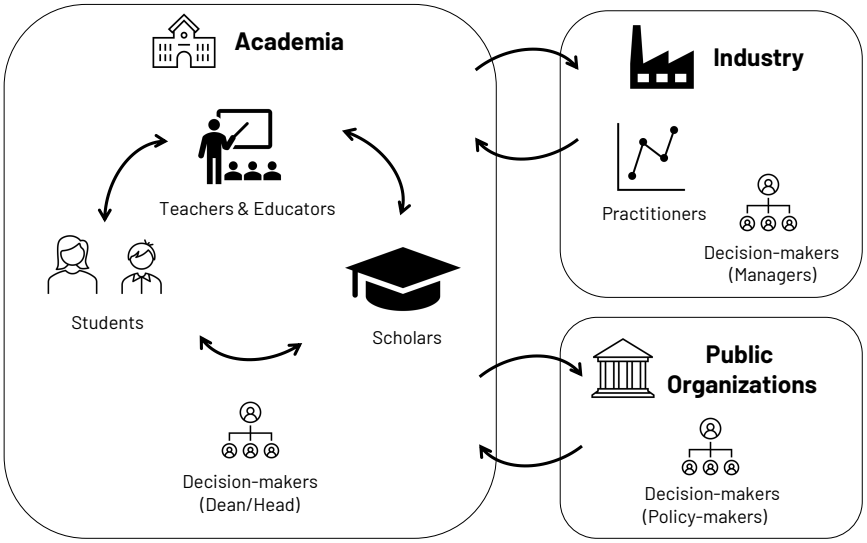

Figure 5. Interaction among actors involved in IR and RS experimental education.

shared resource might further foster student participants because it becomes easier to "stand on the shoulders of giants" yielding to the cycle of education, research, and evaluation that could be streamlined, e.g., by the European Conference on Information Retrieval (ECIR), Conference and Labs of the Evaluation Forum (CLEF), and European Summer School on Information Retrieval (ESSIR) [Ferro, 2023].

Next Steps and Recommendations

Given the importance of reliable and ecologically valid results, one may ask oneself which obstacles occur in the path of developing better education for experimentation and evaluation of information access systems. We see different potential barriers (and opportunities) for the different actors: students, educators, scholars, practitioners, and decision-makers (Figure 5).

Scholars. As has also been identified in a previous Dagstuhl seminar [Ferro et al., 2018], it is significantly harder to test the importance of assumptions in user-facing aspects of the system, such as the presentation of results or the task model, as it is prohibitively expensive to simulate arbitrarily many versions of a system and put them before users. User studies suffer from both the risk of being unable to reject a null hypothesis (non-significant results), as well as the difficulty of conclusively concluding that a hypothesis is rejected (rather than simply a failure to find an effect). User studies are, therefore, at higher risk of criticism and rejection from paper reviewers. Moves toward increased pre-registration (non-significance) have helped authors deal with potentially non-significant results. Furthermore, to test for 'a lack of an effect', there are some proponents of Equivalence Testing [Lakens, 2017][7] and Bayesian Analysis [van Doorn et al., 2021] in Psychology which may also be useful in Computer Science.

As LLMs are becoming a commodity, policies to educate and guide authors and reviewers in how different AI tools can (or cannot) be used for writing assistance should be discussed and

[7] See also https://cran.r-project.org/web/packages/TOSTER/TOSTER.pdf

defined.[8] These guidelines may inspire educators on how to characterize the role of these tools in learning & teaching environments, including assessment design and plagiarism policies.[9]

In addition, a current culture of 'publish or perish' incentivizes short-term and incremental findings,[10] over more holistic thinking and thoughtful comparative analysis. The problem of 'State-of-the-Art (SOTA)-chasing' has also been discussed in other research areas, e.g., in NLP [Church and Kordoni, 2022]. Change in academic incentive systems both within institutions and for conferences and journals change slowly but they do evolve.

Students and Educators. Thankfully, institutions are increasingly recognizing the need for reviewing studies before they are performed, such as Ethics and Data Management plans.[11] In Bachelor and Master education, in particular, this means that instructors may require training in writing such documents, and institutions appreciate and are equipped for timely review. Therefore, planning of education would benefit from allowing sufficient time for submission, review, and revision.

In that context, teaching evaluation methodologies may require some colleagues to retrain, in which case some resistance can be expected. Improving access to training initiatives and materials at post-graduate level can support colleagues who are willing but need additional support. Various forms of informal or even organized exchange between teachers may be a helpful instrument to grow the competency of educators.

Furthermore, certain evaluation concepts and methodologies cannot be taught before certain topics are covered in the curriculum. A student in recommender systems may need to understand the difference between a classification and regression problem; or the difference between precision and recall (for a given task and user it may be more important to retrieve accurate results, or to retrieve a wider range of results) before they can start thinking about the social implications.

Moreover, some students are prone to satisfice, thinking that "good enough is good enough": there are many methodologies available for evaluation, and the options are difficult to digest in a cost-effective way at entry-level—highlighting the need for availability of tutorials and low-entry level materials. Embedding participation to shared tasks and competitions (e.g., CLEF labs or Text REtrieval Conference (TREC) tracks) which provide a common framework for robust experimentation may help overcome this challenge—although the synchronization between the semester and participation timelines may not be straightforward.

Finally, there is a growing number of experiments in developing multi-disciplinary curricula – with the appreciation that different disciplines bring to such a program. Successful initiatives include group projects consisting of students in both Social Sciences and Humanities (SSH) and Computer Science. In fact, one of the underlying principles of the continuously growing *iSchools consortium*[12] is to foster such interdisciplinarity. The challenge here is not only the design of the content but also accreditation and support from the strategic level of institutions.

[8] For instance, see the Association for Computational Linguistics (ACL) 2023 Policy on AI Writing Assistance: https://2023.aclweb.org/blog/ACL-2023-policy/.

[9] https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/

[10] https://harzing.com/resources/publish-or-perish

[11] Further proposals for methodological review are also under discussion in Psychology, but will likely take longer to reach Computer Science: https://www.nature.com/articles/d41586-022-04504-8

[12] https://www.ischools.org

Practitioners. Maintenance of resources used to translate knowledge about models and methodologies for evaluation is challenging given the fast pace of the field. This can make it hard to compare results across studies and to keep up with the SOTA of best practices in experimentation. In this regard, lowering the entry barrier to participating in initiatives such as shared tasks/challenges [Ferro, 2019; Harman and Voorhees, 2005] and maintaining documentation of resources commonly used by non-experts are increasingly helpful.

Another issue is the homogeneity of actors. Often there is no active involvement of actors outside a narrow academic Computer Science sphere, who otherwise might have indicated assumptions or limitations early on. It can be challenging to set up productive collaborations between industry and academia, as well as across disciplines. Typical issues include, for instance, common terminology used in a different way, or different levels of knowledge of key performance indicators. Co-design in labs has set a good precedent in this regard. Examples are ICAI in the Netherlands[13], its extension in the new 10-year ROBUST initiative[14], and the Australian Centre of Excellence for Automated Decision-Making and Society (ADM+S)[15], where PhDs in multiple disciplines (Social Sciences & Humanities, Computer Science, Law, etc.) are jointly being trained in shared projects.

Research Advisory Boards are another effective instrument to draw in practitioners but here the challenge is to make the most of the little time that is usually available for the exchange of ideas between practitioners and academics.

Decision-makers. The output of evaluation and experimentation in IR and RS may be used to inform decision-making on the societal level. Consequently, if the evaluation is poorly done, or the results incorrectly generalized, the implications may also be poor decision-making with far-reaching impacts on society, e.g. [Kahneman, 2011, Ch. 10].

The ability of the other actors to support education on evaluation is constrained and shaped by decision-makers. Policy-makers in public organizations and program managers or deans in academia play a crucial role in curriculum design. Scholars and educators will have to communicate effectively the importance of experimental evaluation in information access in order to inform the decision-making process. The challenge here is to initiate change in the first place and to drive such changes. Any new initiative will necessarily involve not just a single decision-maker but more stakeholders and committees making this a more effortful but possibly also more impactful process than many of the other initiatives we have identified.

Additionally, decision-makers within academic institutions, namely libraries and career development centres, can play an important role towards developing the competency of students and educators. Making best practices in evaluation available as a commodity through these channels will require making resources more accessible for non-experts in IR and RS.

Concluding Remarks

Education and dissemination represent key pillars to overcoming methodological challenges in Information Retrieval and Recommender Systems. What we have sketched here can be interpreted as a general roadmap to create more awareness among and beyond the IR and RS communities.

[13] https://icai.ai/
[14] https://icai.ai/ltp-robust/
[15] https://www.admscentre.org.au/

We hope the recommendations—and the identified challenges to consider—on what we can do will help to support education for better evaluation in the different stages of the lifelong learning journey. We acknowledge that facets such as incentive mechanisms and processes in institutions are often slow-moving. The vision proposed in this section is therefore also aimed at a longer-term (5–10 years) perspective.

A more detailed discussion can be found in the full report [Bauer et al., 2023b].

5 Results-blind Reviewing

Campbell and Stanley [1963, p. 1] defined experiments as "that portion of research in which variables are manipulated and their effects upon other variables observed". Scientific experiments are used in confirmatory research to test a priori hypotheses as well as in exploratory research to gain new insights and help to generate hypotheses for future research [Shadish et al., 2002]. In information access research, the ultimate goal is to gain insights into cause and effect. Unfortunately, many reviewers of information access experiments place undue emphasis on performance, rejecting papers that contain insights if they fail to show improvements in performance. The focus on performance numbers not only leads to publication bias. It also puts additional pressure on early-career researchers who must publish or perish, thus being tempted to cheat if their proposed method does not yield the desired results. Moreover, reviewers pay little attention to the experimental methodology and analysis in case the results are impressive [Fuhr, 2017]. Focusing primarily on performance (and in particular aggregated performance) can lead to a neglect of insights; gaining insights is critical to move the information access field forward and essential to be able to make performance predictions [Ferro et al., 2018].

We think that one important step to change the situation is if we alter the review process such that there is more emphasis on the theoretical background, the hypotheses, the methodological plan and the analysis plan of an experiment, while improvement or decline of performance should play less of a role when deciding about the quality of a paper. It is hoped that this will lead to a higher scientific quality of publications, more insights, and improved reproducibility (as there is less incentive for beautifying results). As Woznyj et al. [2018] note in their survey of editorial board members, overall there are positive attitudes towards results-blind reviewing and advantages for the scientific community outweigh concerns.

In order to move the review focus away from performance improvement, appealing to reviewers alone will not be sufficient. A more drastic measure is the change of the review process such that reviewers decide about acceptance vs. rejection of a paper without knowing the outcome of the experiments described.

Next Steps and Recommendations

We propose several changes to the reviewing processes for information access papers to reduce publication biases

Recommendation 1: Pilot test of results-blind reviewing in conference(s) or journal(s). Our first and most important recommendation is that the information access research communities (i.e., IR and RS communities) adopt a results-blind approach to peer reviewing for

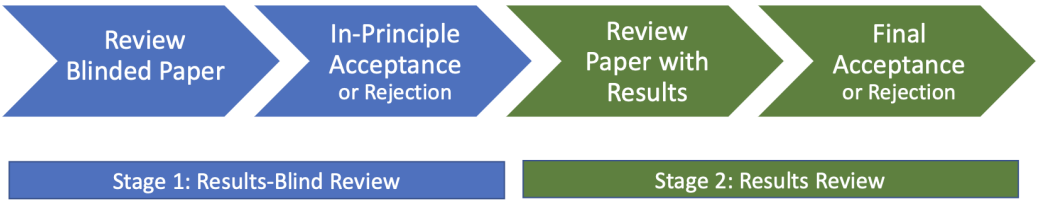

Figure 6. Proposed two-stage process for results-blind reviewing (figure adapted from BMC)

conference(s) and/or journal(s). We recommend that the community start with a pilot test of results-blind reviewing in an established conference track, perhaps with a new paper track with an earlier deadline to allow for a two-stage review process. In results-blind reviewing, the authors submit two versions of their manuscript: one version of the paper with the full results, and one version with the results blinded. The two submitted versions are the basis of a results-blind reviewing process with two major stages (see Figure 6).

Stage 1 consists of the Results-Blind Review. The results-blind version of the manuscript is reviewed and an in-principle acceptance (or rejection) is made. During Stage 1, as in the traditional reviewing process, the paper is reviewed by multiple reviewers who also make acceptance recommendations. In the case of conferences, the in-principle acceptance (or rejection) decision is made after discussion with the Senior Program Committee (SPC)/meta reviewer and in the Program Committee (PC) meeting. Papers that receive an in-principle acceptance proceed to Stage 2.

Stage 2 consists of the Results Review. The paper containing the results is reviewed by the same set of reviewers with a focus on the results. In the case of a conference, the final acceptance (or rejection) decision is made after a discussion period with the SPC and in the PC meeting.

Recommendation 2: Initial process recommendation for a results-blind reviewing pilot. Below, we recommend a high-level process for how a results-blind reviewing process pilot might be implemented and important considerations for conference organizers and reviewers as well as authors.

Once the decision for results-blind reviewing has been made, **conference organizers** would have to take the following steps:

- First, the Call for Papers (CfP) for the new track should be written. As the proposed results-blind reviewing process with two stages of review will take longer to complete, an earlier deadline for this track should be set.
- Criteria for both stages of the review (blinded and with results) should be defined. Special attention should be given to the criteria for changing an initial acceptance recommendation into a rejection.
- Author instructions for the results-blind reviewing track have to be formulated, describing not only the new reviewing criteria and process but also specific instructions on how to prepare the blinded version of an article. For the results-blind version of the paper, the authors will need to blind all mentions of the results (e.g., in the abstract, introduction, discussion, and conclusion in addition to in a results section) in a way that it is not technically

possible to recover the blinded text. There should be a way for reviewers to easily determine the differences between the results-blind version of the paper and the one with the results.
- Reviewers for the results-blind reviewing track have to be recruited. In the beginning, additional or different expertise will be required for this track. A special introduction of training for the reviewers might be necessary in order to make them familiar with the new process and criteria.
- The reviewing software will need to be configured for multiple stages of review for the results-blind reviewing. In the first stage of reviewing, only the blinded version of the papers should be distributed to reviewers (see below for the process for reviewers).
- After the final decision by the PC, the authors will be provided with the review and informed about the final accept or reject decision. In the case of a rejection decision, authors should also be notified at which stage the paper was rejected.
- The organizers should give special recognition to the PC member of the track (on the conference Web site and in the proceedings)
- The success of the new track and the process should be evaluated.

Once the **reviewers** are provided with instructions about the general process and received additional training, we recommend the following process:

- In the first stage, the reviewers are provided with the results-blind version of the submission and complete their review including a recommendation about the in-principle acceptance.
- Once the reviews are complete, a discussion phase with the SPC follows, leading to a recommendation for each paper.
- The PC for the track meets and makes an initial decision (in-principle acceptance or rejection) for each paper.
- For the second reviewing stage, only in-principle accepted papers are considered. Reviewers get the full versions of the papers they reviewed before. They add an additional part to their review focusing on the results which were previously blinded. Also, they make a second recommendation about acceptance.
- As for the first phase, a discussion phase with the SPC follows leading to a recommendation for each paper.
- The track PC meets for the second time and makes the final decision for each paper.

Authors will have to understand the new reviewing scheme, and possibly be trained for or educated on how to prepare manuscripts that satisfy the new reviewing criteria. They will have to prepare and submit two versions of a paper, a version with the results as in the traditional model as well as one in which the results are blinded.

Recommendation 3: Emphasize insights in papers. We recommend that authors, conference organizers, and reviewers place additional emphasis on communicating expected insights to be gained from experiments. Guidelines (and review forms) should ask the reviewers to comment on the theoretical background, the hypotheses, the methodological plan and the analysis plan of the experiment(s) described. Special attention should be given to the expected insights to be gained from experiments, i.e. regarding cause and effect.

Recommendation 4: Extra space for methods information. Another recommendation is for the community to consider explicitly allowing methodological appendices for authors to provide additional methodological details outside of page and/or word limits and to include these appendices with the text of the paper and not as supplementary materials. While not needed for all publications, this would be very beneficial for some types of studies so that the authors can include all study materials. For example, in user studies, researchers may administer multiple questionnaires, conduct a semi-structured interview, and read from a script. It is not uncommon for researchers to administer multiple questionnaires and conduct a semi-structured interview.

This would be especially important if adopting a results-blind reviewing process as careful scrutiny of the study design and all study materials is needed to ascertain whether the authors will be able to answer the research questions. For example, due to page limits, it is common for authors to describe the topics of an interview but uncommon to include the full text of an interview guide due to page limits.

In addition, this would have an additional benefit for other researchers who wish to replicate the study. While, for example, authors can currently make supplementary materials available in ACM Digital Library (ACM DL), these materials are not included in the downloadable version of the article or when reading online in the ACM DL in the eReader or HTML formats.

Recommendation 5: Consider a two-stage review process adapted from preregistered or registered reports. Although our primary recommendation is for conference organizers or journal editors to embrace a results-blind reviewing approach, we also recommend that they consider piloting a conference track or article type in which the study protocol undergoes peer review and is accepted in-principle before data collection or analysis begins. This may be more appropriate for certain types of research (e.g., user studies).

Concluding Remarks

At first glance, the new result-blind reviewing scheme might seem to be only attractive for papers describing failed experiments, while authors with successful results would go to the established tracks. In order to avoid this impression, it is essential that the new scheme is piloted as a highly visible and prestigious track in an established conference. Furthermore, it should be clearly communicated that the results-blind reviewing scheme aims at establishing high standards for the design, execution and analysis of experiments while shielding the reviewers from being blinded by shiny experimental results. Thus, it is our hope that papers published in this track will be regarded as high-quality publications which thoroughly address research questions and clearly demonstrate the insights that may be gained from the research.

A more detailed discussion can be found in the full report [Beel et al., 2023].

6 Guidance for Authors

The IR community has over time developed a strong shared culture of expectations of published papers, particularly in our leading venues. However, these expectations are not explicit and the evidence of submitted papers is that many authors are not aware of what elements, or omissions, are likely to be of concern to reviewers. While accepted papers do provide an indication of what

an author should do, they are, of course, uneven, and the small set of papers that an author is consulting in their new work could easily be unrepresentative of the best IR work as a whole.

In this section, our aim is to provide a basis for general guidance for authors and reviewers, with a focus on people who are new to the community. It should communicate to authors and reviewers a range of factors that the community regards as significant. Such guidance, if well designed, should help authors to lift the standard of their work and provide context should it not be accepted; for reviewers, especially those new to the task, it can provide checklists and (at a high level) advice about the field from beyond their immediate research environment.

Some elements in papers have attracted specific criticism in publications; this is particularly true of effectiveness measurement, where a long history of research on method has argued for and against a range of measures, forms of evidence for statistical validity, treatment of test collections, and so on. Such literature is critical to improving the quality of our research but does not necessarily represent a settled, shared view of best practice.

In our view, it is essential that general advice be constructive, readily understandable by new IR authors and reviewers, and—to the extent that is possible—not the subject of active debate. We used the following approach to create our guidelines: (1) search of existing guidelines; (2) brainstorming to identify common pitfalls; (3) categorization of the outcomes from the brainstorming exercise and comparison of these with existing guidelines; and (4) consolidation and integration with existing SIGIR guidelines. The existing guidelines include the ACM Special Interest Group on Information Retrieval (SIGIR) recommenda- tions to strengthen IR papers; Empirical Evaluation Guidelines from the ACM Special Interest Group on Programming Languages (SIGPLAN); the Special Interest Group on CHI SIGCHI69 guide for reviewing papers sub- mitted to the CHI conference; an ACL tutorial instructing reviewers on the ACL Rolling Review process, which includes common reasons for rejection: and Ulmer et al. [2022]'s list of best practices and guidelines for experimental standards within NLP.

Throughout each step of the process, we adhere to the principle of keeping only issues that we believe to be widely agreed upon within the community. How this work might develop over time is considered under "next steps".

Our proposed guidance is as follows.[16]

Motivation and claims

- The problem is well characterised and motivated, and the potential impact is discussed.
- The proposed application of the work is contextualised by pertinent knowledge from that domain, including potential ethical, social, or environmental impacts.
- The research goals and original contributions (that is, the elements that are a contrast to the prior art) are stated and are clearly distinguished from prior work.
- The claims are properly scoped and supported.
- There are explicit statements of what was done and what was not.

[16]This guidance has since been adopted by the SIGIR-AP 2023 conference.

Presentation

- The literature review considers competitive previous solutions for the problem, that is, it is not limited to consideration of other work on the same technology as that explored in the submission.
- There is a reasoned justification for each of the choices made in each step of the research and each element of the method.
- Results are presented in keeping with the norms in the field as exemplified in strong prior work.
- A substantive, focused, and insightful discussion accompanies the results taking into account limitations and scope of the work.

Experiments

- The experimental design and its scale are appropriate to the problem.
- In comparative studies, appropriate baselines are used; they are deployed and optimized in ways comparable to those used for the proposed method.
- The experimental results are reliable and generalizable, and preferably show illustrative individual cases as well as aggregated results.
- Where appropriate, a diversity of data sets are used, including public-domain data sets used in prior work.
- Sufficient details (with data and code where appropriate) are provided to enable other researchers to assess and reproduce the experiments; this includes the nature, source, and collection process for the data, and the data preparation steps.

Results and analysis

- The evaluation methods and measures address the research questions; the use of redundant or highly correlated measures should be avoided.
- Statistical analysis is used and reported appropriately.
- Development data, training data, and test data are distinguished from each other.
- User studies are based on adequately sized, representative cohorts; data is gathered in ways that meet ethical norms, or where appropriate in keeping with prescribed ethics practices.
- Final results were obtained after all development was complete, that is, not selected because they are the best outcomes amongst a larger set of experiments or hand-fitted to the data.

Common problems that lead to rejection

Issues with papers in relation to the recommendations above can lead to rejection. Other problems that can lead to rejection are as follows.

- Literature reviews that lack critical analysis of prior work or that largely consist of lists of papers, that is, do not have an insightful discussion.
- Contributions that consist of small modifications to established techniques, particularly where the contribution is a straightforward variation of the established technique or where there are numerous prior papers exploring similar variations.
- Methods that appear to be developed and hand-tuned on a specific data set without discussion or demonstration of their lessons for future work or of how the methods would be more generally applicable.
- Justification of a method solely by its score in experiments, lacking an a priori rationale for why the method is worth exploring.
- Experiments where the data volumes are too small to support the conclusions.
- Any form of academic fraud, misrepresentation, or dishonesty.

Next Steps and Recommendations

Guidance and lists of issues should be living documents that reflect a current and uncontroversial agreement in the community. Therefore, they should be open to change because there can always be some disagreements and expectations of authors can change over time, in some cases quite quickly, especially as the subjects of research shift to focus on new topics. For that reason, no set of advice should be regarded as fixed, but revision should be undertaken consultatively and with a spectrum of colleagues.

We suggest that the detailed list of issues of concern, such as those reported in Appendix 6.4 of the full report [Di Nunzio et al., 2023], be made available in some form as educative for reviewers. We stress here that it is not our intention that reviewers simply reject papers because of these issues. It could also provide a resource at forums such as doctoral consortia.

We thus believe that it would be valuable for the community to:

- Ensure that the guidelines are prominent in the calls-for-papers at our major conferences and journals, or otherwise disseminated.
- Encourage the SIGIR executive committee to take ownership of the guidelines and to occasionally convene a panel to produce an update.
- Use these resources educatively for new members of the community and for new reviewers.

In this exercise, we have not produced guidance for reviewers, which in other disciplines tends to consist of two parts: general advice on how to approach the task and specifics for the field. An example that we found was produced by the ACL, as discussed above; a particular strength of these guidelines in our view is the enumeration of unfair grounds for rejection. We believe that such guidance would be of value to our community, and could make use of the materials we have presented here.

A more detailed discussion can be found in the full report [Di Nunzio et al., 2023].

7 Participants

Below is a list of researchers who attended the seminar and contributed to the full report [Bauer et al., 2023a].

- Christine Bauer, Utrecht University, The Netherlands
- Joeran Beel, University of Siegen, Germany
- Timo Breuer, Technische Hochschule Köln, Germany
- Charles L. A. Clarke, University of Waterloo, Canada
- Anita Crescenzi, University of North Carolina at Chapel Hill, United States
- Gianluca Demartini, The University of Queensland, Australia
- Giorgio Maria Di Nunzio, University of Padua, Italy
- Laura Dietz, University of New Hampshire, United States
- Guglielmo Faggioli, University of Padua, Italy
- Nicola Ferro, University of Padua, Italy
- Bruce Ferwerda, Jönköping University, Sweden
- Maik Fröbe, Friedrich-Schiller-Universität Jena, Germany

- Norbert Fuhr, University of Duisburg-Essen, Germany
- Matthias Hagen, Friedrich-Schiller-Universität Jena, Germany
- Allan Hanbury, TU Wien, Austria
- Claudia Hauff, Spotify, The Netherlands
- Dietmar Jannach, University of Klagenfurt, Austria
- Noriko Kando, National Institute of Informatics, Japan
- Evangelos Kanoulas, University of Amsterdam, The Netherlands
- Bart P. Knijnenburg, Clemson University, United States
- Udo Kruschwitz, University of Regensburg, Germany
- Meijie Li, University of Duisburg-Essen, Germany
- Maria Maistro, University of Copenhagen, Denmark
- Lien Michiels, University of Antwerp, Belgium
- Andrea Papenmeier, University of Duisburg-Essen, Germany
- Martin Potthast, Leipzig University, Germany
- Paolo Rosso, Technical University of Valencia, Spain
- Alan Said, University of Gothenburg, Sweden
- Philipp Schaer, Technische Hochschule Köln, Germany
- Christin Seifert, University of Duisburg-Essen, Germany
- Ian Soboroff, National Institute of Standards and Technology, United States
- Damiano Spina, RMIT University, Australia
- Benno Stein, Bauhaus-Universität Weimar, Germany
- Nava Tintarev, Maastricht University, The Netherlands
- Julián Urbano, Delft University of Technology, The Netherlands
- Henning Wachsmuth, Leibniz Universität Hannover, Germany
- Martijn Willemsen, Eindhoven University of Technology & JADS, The Netherlands
- Justin Zobel, University of Melbourne, Australia

References

Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: Two experiments manipulating Google output. *Journal of Medical Internet Research*, 16(4):e100, 2014. doi: 10.2196/jmir.2642.

Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 601–610, 2009. doi: 10.1145/1645953.1646031.

Christine Bauer and Eva Zangerle. Leveraging multi-method evaluation for multi-stakeholder settings. In *1st Workshop on the Impact of Recommender Systems, co-located with 13th ACM Conference on Recommender Systems (ACM RecSys 2019)*, ImpactRS '19. CEUR-WS.org, 2019. URL http://ceur-ws.org/Vol-2462/short3.pdf.

Christine Bauer, Ben A. Carterette, Nicola Ferro, and Norbert Fuhr, editors. *Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education*, Dagstuhl Reports, Volume 13, Number 1, 2023a. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany.

Christine Bauer, Maik Fröbe, Dietmar Jannach, Udo Kruschwitz, Paolo Rosso, Damiano Spina, and Nava Tintarev. Overcoming methodological challenges in information retrieval and recommender systems through awareness and education. In Bauer et al. [2023a], pages 51–67.

Joeran Beel, Timo Breuer, Anita Crescenzi, Norbert Fuhr, and Meijie Li. Results-blind reviewing. In Bauer et al. [2023a], pages 67–73.

Jöran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *Proceedings of the 22nd International Conference on Theory and Practice of Digital Libraries*, TPDL '15, pages 153–168, 2015.

Stefan Büttcher, Charles L. A. Clarke, Peter C. K. Yeung, and Ian Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 63–70. ACM, 2007. doi: 10.1145/1277741.1277755.

Donald T. Campbell and Julian C. Stanley. *Experimental and quasi-experimental designs for research*. Houghton Mifflin Company, Boston, 1963.

Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 15–24. ACM, 2017. doi: 10.1145/3077136.3080804.

Kenneth Ward Church and Valia Kordoni. Emerging trends: Sota-chasing. *Natural Language Engineering*, 28(2):249–269, 2022. doi: 10.1017/S1351324922000043.

Charles L. A. Clarke, Gianluca Demartini, Laura Dietz, Guglielmo Faggioli, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Ian Soboroff, Benno Stein, and Henning Wachsmuth. Hmc: A spectrum of human–machine-collaborative relevance judgment frameworks. In Bauer et al. [2023a], pages 41–50.

Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. Assessing top-preferences. *ACM Transactions on Information Systems*, 39(3), 2021. doi: 10.1145/3451161.

Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8):70–79, 2020. doi: 10.1145/3360311.

Giorgio Maria Di Nunzio, Maria Maistro, Christin Seifert, Julián Urbano, and Justin Zobel. Guidance for authors. In Bauer et al. [2023a], pages 74–79.

Tim Draws, Nava Tintarev, and Ujwal Gadiraju. Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explorations Newsletter*, 23(1):50–58, 2021a. doi: 10.1145/3468507.3468515.

Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 295–305. ACM, 2021b. doi: 10.1145/3404835.3462851.

Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. Viewpoint diversity in search results. In *Advances in Information Retrieval*, pages 279–297. Springer, 2023. doi: 10.1007/978-3-031-28244-7_18.

Michael D. Ekstrand, Michael Ludwig, Joseph A. Konstan, and John T. Riedl. Rethinking the recommender research ecosystem: Reproducibility, openness, and LensKit. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 133–140, 2011.

Theresa Elstner, Frank Loebe, Yamen Ajjour, Christopher Akiki, Alexander Bondarenko, Maik Fröbe, Lukas Gienapp, Nikolay Kolyada, Janis Mohr, Stephan Sandfuchs, Matti Wiegmann, Jörg Frochte, Nicola Ferro, Sven Hofmann, Benno Stein, Matthias Hagen, and Martin Potthast. Shared tasks as tutorials: A methodical approach. In *37th AAAI Conference on Artificial Intelligence*, AAAI 2023. AAAI, 2023.

Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015. doi: 10.1073/pnas.1419828112.

Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017. doi: 10.1145/3134677.

Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. Perspectives on large language models for relevance judgment. In *Proceedings of the 13th International Conference on the Theory of Information Retrieval*, ICTIR '23, 2023.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pages 101–109. ACM, 2019. doi: 10.1145/3298689.3347058.

Nicola Ferro. What Happened in CLEF... For a While? In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, CLEF '19, pages 3–45. Springer, 2019.

Nicola Ferro. Coordinate research, evaluation, and education in information access: Towards a more sustainable environment for the community. In Bauer et al. [2023a], pages 13–16.

Nicola Ferro and Mark Sanderson. How do you test a test?: A multifaceted examination of significance tests. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pages 280–288. ACM, 2022. doi: 10.1145/3488560.3498406.

Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan, Pablo Castells, Elizabeth M. Daly, Thierry Declerck, Michael D. Ekstrand, Werner Geyer, Julio Gonzalo, Tsvi Kuflik, Krister Lindén, Bernardo Magnini, Jian-Yun Nie, Raffaele Perego, Bracha Shapira, Ian Soboroff, Nava Tintarev, Karin Verspoor, Martijn C. Willemsen, and Justin Zobel. From evaluating to forecasting performance: How to turn information retrieval, natural language processing and recommender systems into predictive sciences (Dagstuhl Perspectives Workshop 17442). *Dagstuhl Manifestos*, 7(1):96–139, 2018. doi: 10.4230/DagMan.7.1.96.

Bruce Ferwerda, Akkan Hanbury, Bart P. Knijnenburg, Birger Larsen, Lien Michiels, Andrea Papenmeier, Alan Said, Philipp Schaer, and Martijn C. Willemsen. Reality check—conducting real world studies. In Bauer et al. [2023a], pages 20–40.

Juliana Freire, Norbert Fuhr, and Andreas Rauber, editors. *Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science*, Dagstuhl Reports, Volume 6, Number 1, 2016. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany. doi: 10.4230/DagRep.6.1.108.

Norbert Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, 2017. doi: 10.1145/3190580.3190586.

Carlos A. Gomez-Uribe and Neil Hunt. The Netflix recommender system: Algorithms, business value, and innovation. *Transactions on Management Information Systems*, 6(4), 2016. doi: 10.1145/2843948.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 1644–1651, 2018. doi: 10.1609/aaai.v32i1.11503.

Donna K. Harman. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119, 2011. doi: 10.1007/978-3-031-02276-0.

Donna K. Harman and Ellen M. Voorhees, editors. *TREC. Experiment and Evaluation in Information Retrieval*, 2005. MIT Press.

Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond DCG: User behavior as a predictor of a successful search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 221–230. ACM, 2010. doi: 10.1145/1718487.1718515.

W. Heisenberg. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 43(3-4):172–198, 1927. doi: 10.1007/BF01397280.

Dietmar Jannach and Gediminas Adomavicius. Price and profit awareness in recommender systems. In *Proceedings of the ACM RecSys 2017 Workshop on Value-Aware and Multi-Stakeholder Recommendation*, 2017.

Dietmar Jannach and Christine Bauer. Escaping the McNamara fallacy: Towards more impactful recommender systems research. *AI Magazine*, 41(4):79–95, 2020. doi: 10.1609/aimag.v41i4.5312.

Daniel Kahneman. *Thinking, fast and slow*. Penguin, 2011.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2), 2007. doi: 10.1561/1500000012.

Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4–5):411–504, 2012. doi: 10.1007/s11257-011-9118-4.

Joseph A. Konstan and Gediminas Adomavicius. Toward identification and adoption of best practices in algorithmic recommender systems research. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 23–28, 2013.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Significant improvements over the state of the art? a case study of the MS MARCO document ranking leaderboard. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2283–2287. ACM, 2021. doi: 10.1145/3404835.3463034.

Marianne Lykke, Ann Bygholm, Louise Bak Søndergaard, and Katriina Byström. The role of historical and contextual knowledge in enterprise search. *Journal of Documentation*, 78(5):1053–1074, 2022. doi: 10.1108/jd-08-2021-0170.

Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. When does relevance mean usefulness and user satisfaction in Web search? In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 463–472. ACM, 2016. doi: 10.1145/2911451.2911507.

Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 209–216. ACM, 2017. doi: 10.1145/3121050.3121074.

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. Argument search: Assessing argument relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2019, pages 1117–1120. ACM, 2019. doi: 10.1145/3331184.3331327.

Tetsuya Sakai. Laboratory experiments in information retrieval. *The Information Retrieval Series*, 40, 2018. doi: 10.1007/978-981-13-1199-4.

David P. Sander and Laura Dietz. EXAM: how to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. In *Proceedings of the Second International Conference on Design of Experimental Search & Information Retrieval Systems*, pages 136–146. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-2950/paper-16.pdf.

Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 555–562. ACM, 2010. doi: 10.1145/1835449.1835542.

Daniel Schwartz, Baruch Fischhoff, Tamar Krishnamurti, and Fallaw Sowell. The Hawthorne effect and energy awareness. *Proceedings of the National Academy of Science*, 110(38):15242–15246, 2013. doi: 10.1073/pnas.1301687110.

William R Shadish, Thomas D Cook, and Donald T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company, New York, 2002.

Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Geng. Are we evaluating rigorously? Benchmarking recommendation for reproducible evaluation and fair comparison. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, pages 23–32, 2020. doi: 10.1145/3383313.3412489.

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. Experimental standards for deep learning in natural language processing research, 2022. URL https://aclanthology.org/2022.findings-emnlp.196.

Johnny van Doorn, Don van den Bergh, Udo Böhm, Fabian Dablander, Koen Derks, Tim Draws, Alexander Etz, Nathan J Evans, Quentin F Gronau, Julia M Haaf, et al. The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28(3):813–826, 2021.

Ryen W White. *Interactions with search systems*. Cambridge University Press, New York, NY, USA, 2016.

Haley M. Woznyj, Kelcie Grenier, Roxanne Ross, George C. Banks, and Steven G. Rogelberg. Results-blind review: a masked crusader for science. *European Journal of Work and Organizational Psychology*, 27(5):561–576, 2018. doi: 10.1080/1359432x.2018.1496081.

Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1129–1132, 2019. doi: 10.1145/3331184.3331340.

Eva Zangerle and Christine Bauer. Evaluating recommender systems: survey and framework. *ACM Computing Surveys*, 55(8), 2022. doi: 10.1145/3556536.

Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 379–388. ACM, 2020. doi: 10.1145/3397271.3401162.

Justin Zobel. When measurement misleads: The limits of batch assessment of retrieval systems. *SIGIR Forum*, 56(1), 2023. doi: 10.1145/3582524.3582540.