References

1 Stefan Büttcher, Charles L. A. Clarke, Peter C. K. Yeung, and Ian Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 63–70. ACM, 2007.

2 Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. Assessing top-k preferences. *ACM Trans. Inf. Syst.*, 39(3):33:1–33:21, 2021.

3 Donna Harman. Information retrieval evaluation. 2011.

4 Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. Argument search: Assessing argument relevance. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1117–1120. ACM, 2019.

5 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 Omar Alonso, Stefano Marchesin, Marc Najork, and Gianmaria Silvello, editors, *Proceedings of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15-18, 2021*, volume 2950 of *CEUR Workshop Proceedings*, pages 136–146. CEUR-WS.org, 2021.

6 Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2001.

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

Christine Bauer (Utrecht University, NL, c.bauer@uu.nl)
Maik Fröbe (Friedrich-Schiller-Universität Jena, DE, maik.froebe@uni-jena.de)
Dietmar Jannach (University of Klagenfurt, AT, dietmar.jannach@aau.at)
Udo Kruschwitz (University of Regensburg, DE, udo.kruschwitz@ur.de)
Paolo Rosso (Technical University of Valencia, ES, prosso@dsic.upv.es)
Damiano Spina (RMIT University, AU, damiano.spina@rmit.edu.au)
Nava Tintarev (Maastricht University, NL, n.tintarev@maastrichtuniversity.nl)

4.3.1 Background & Motivation

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 [4, 13, 48, 53].

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 [10, 5, 40, 46, 56] 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 [19]. This phenomenon – often referred to as the *Search Engine Manipulation Effect (SEME)* – has been demonstrated for both politics [19, 20] and health [2, 43]. By being aware of the phenomena, methods have been adapted to measure its presence [14, 15], and studies to evaluate when and how it affects human decision-makers [16]. 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 [7, 25, 30].

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 [4, 35, 57] and *(ii)* how effective we are in sharing and translating the knowledge we currently have for doing IR and RS experimentation [23, 45]. 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. [36, 51, 55]. Studies on long-term impact furthermore need to consider the wider scope of stakeholders [6, 30]. Moreover, for several types of information access systems, the specific and potentially competing interests of multiple stakeholders have to be taken into account [6]. 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 [29, 30].

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 [17, 5, 32, 46]. 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 problem of reproducibility[24] and replicability [25] we face in Computer Science [12, 1] in general and in AI in particular [26].

This chapter is organized as follows: Next, in Section 4.3.2 we briefly review which kinds of actors may benefit from better education in information access system experimentation. Afterwards, in Section 4.3.3, we provide concrete examples of what we can do in terms of concrete resources and initiatives to increase the awareness and knowledge level of the different actors. Finally, in Section 4.3.4, we sketch the main challenges that we may need to be aware of when implementing some of the described educational initiatives.

4.3.2 Actors

As in any process related to the advancement, communication, and sharing of knowledge, knowing how to properly design and carry out correct and robust experimentation concerns people with various different roles.

This covers a broad spectrum including academia, industry, and public organizations, e.g., from a lecturer in IR and RS introducing evaluation paradigms to undergrad students and data scientists – not necessarily experienced in IR and RS – choosing metrics aligned to business Key Performance Indicators (KPIs) by looking at textbooks and Wikipedia pages. We have identified a number of actors that are involved in the education to experimentation in information access, who are listed below. Note that this categorization is not exhaustive nor exclusive, as actors may have multiple roles.

Students

This category embraces the different stages of academic training. Starting from students enrolled in IR & RS courses [41], including, for instance, undergraduate students in Computer Science degrees and Master's students in Data Science, AI, and Human-Computer Interaction. It also includes students enrolled in a doctoral degree, i.e., PhD students, including those jointly co-supervised with industry.

Educators

Academic roles related to education, such as course coordinators, lecturers, teaching assistants, as well as research student supervisors.

Scholars

Researchers and academics involved in academic services, including reviewers, journal editors, program chairs, grant writers, etc.

[24] https://www.wired.com/story/machine-learning-reproducibility-crisis/
[25] https://cacm.acm.org/magazines/2020/8/246369-threats-of-a-replication-crisis-in-empirical-computer-science/abstract

Practitioners

Data scientists, developers, User Experience (UX) designers, and other practitioners outside academia, that may need support in their lifelong learning.

Decision-makers

People that make strategic decisions in processes, policies, products and/or human resources (e.g., managers in industry or policy-makers) that may benefit from having a better understanding of IR and RS core concepts in evaluation and experimentation.

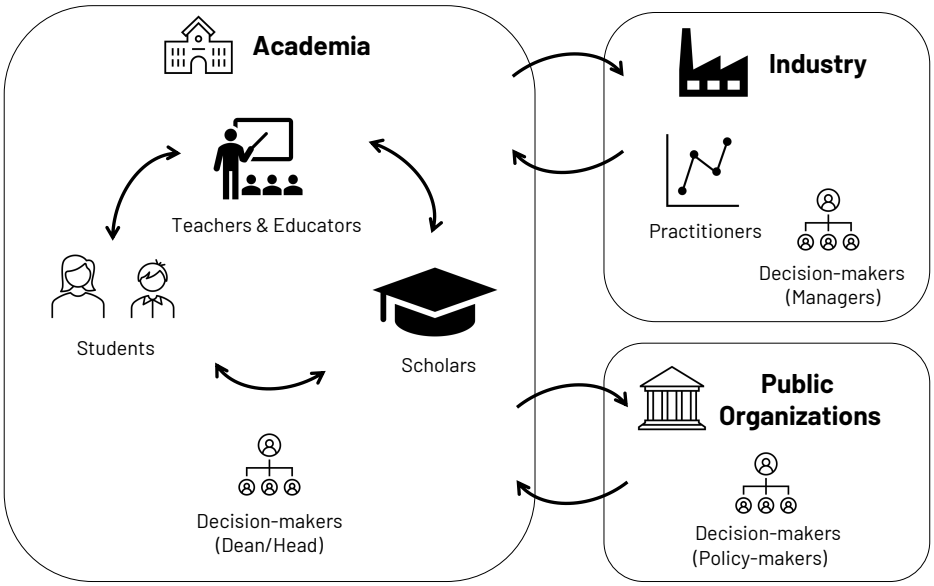

Figure 6 Interaction among actors involved in IR and RS experimental education.

Figure 6 shows the interaction among the identified actors. In academia, students, educators, and scholars are in continuous interaction through learning, teaching, and supervision processes, which are overseen and/or led by decision-makers such as deans, heads of departments, etc. In industry, decision-makers such as product and team managers, as well as practitioners, make use of training and education resources and initiatives to support experimentation in real-world domains. The cyclic arrows represent the active participation in the creation and development of those resources and initiatives. Decision-makers in public organizations, such as policy-makers, are also key actors in the definition of curricula, which has a direct impact on how and to which extent experimentation in IR and RS is included in Data Science, Computer Science, Computer-Human-Interaction (CHI), and AI programs.

4.3.3 What can we do?

In this section, we first provide examples of helpful *resources* to improve education in IR and RS evaluation. Then, we outline several possible *initiatives* that contribute to increasing awareness about current methodological issues and to disseminate knowledge about experimentation approaches.

4.3.3.1 Resources

The resources with which the actors interact are a way to share, maintain, and promote best practices while ensuring a low barrier of entry to the field. Given that those resources might be widely used in education, research (experimentation, etc.), and even production systems, resources have great potential to continuously grow the knowledge of future generations of scholars, practitioners, and decision-makers.

General Teaching Material. Textbooks quickly may become outdated,[26] but have the advantage that these typically reach a wide audience, whereas slides and tutorials that cover evaluation methodology in more depth might only reach smaller audiences. Often, today's online lectures primarily report on "mainstream" information retrieval (e.g., offline studies, common metrics), but foster reflection and discussion only to a very limited extent. More comprehensive resources should be made publicly available and shared across universities, summer schools, and meetups.[27] Finally, having the IR and RS community actively contribute to the curation of material in sources that are widely used by the general public – and, thus, also by students – as a starting point to get a basic understanding of a topic (e.g., Wikipedia) is advisable. Further, contributing to the documentation of software such as Apache Solr,[28] Elasticsearch,[29] Surprise,[30] Implicit,[31] etc. (see the report by Ferro et al. [22] for more that are widely used in practice), can help to make non-experts more aware of the best practices in IR and RS experimentation.

Apart from introducing modern information retrieval systems, **teaching material** should give more attention to a wider set of application fields of IR, including recommender systems and topics related to query and interaction mining and understanding, and online learning to rank [41]. To date, also online evaluation falls short in such resources although it is essential in the spectrum of evaluation types [41]. Students need to be introduced to concepts such as reproducibility and replicability, and it is essential that students understand what makes a research work impactful in practice. To lower the entry barrier to the field, students should be taught how to use available tools and environments that enable quick prototyping, and that have real-world relevance. Teaching fairness, privacy, and ethical aspects, both in designing experiments and also in how to evaluate them, is also important.[32]

Moreover, the participation in **shared tasks (challenges or competitions)** of evaluation campaigns in IR (e.g., TREC,[33] CLEF,[34] NTCIR,[35] or FIRE[36]) and RecSys (e.g., the yearly ACM RecSys challenges[37]) should be fostered. To facilitate the participation of

[26] In contrast to that, the main textbook in the area of natural language processing has for years only been available as an online draft and is continuously being updated: `https://web.stanford.edu/~jurafsky/slp3/`

[27] For instance, Sebastian Hofstätter released Open-Source Information Retrieval Courses: `https://github.com/sebastian-hofstaetter/teaching`.

[28] `https://solr.apache.org/`

[29] `https://www.elastic.co/es/elasticsearch/`

[30] `https://surpriselib.com/`

[31] `https://implicit.readthedocs.io`

[32] Cyprus Center for Algorithmic Transparency (CyCAT) project: `https://sites.google.com/view/biasvisualizationactivity/home`

[33] `https://trec.nist.gov/`

[34] `https://www.clef-initiative.eu/`

[35] `https://research.nii.ac.jp/ntcir/`

[36] `https://fire.irsi.res.in/fire/`

[37] `https://recsys.acm.org/challenges/`

students, it is worthwhile to make the timelines of such challenges and competitions compatible with the academic (teaching) schedules (e.g., in terms of semesters). Students will be provided with the datasets used in the benchmarks and will be able to learn more on evaluation methodologies (for instance, students from Padua, Leipzig, and Halle participated in Touché [8, 9] hosted at CLEF). At the same time, it is important to critically reflect with students on the limitations and dangers of competitions [11] and encourage them to go beyond leaderboard State Of The Art (SOTA) chasing culture – e.g., only optimizing on one metric or a limited set of metrics without reflection of the suitability of these metrics in a given application context [50, 30]. Hence, it is important that a student's (or student group's) grade does not depend on their rank in the leaderboard but to a large degree on their approach, reasoning, and reflection to counteract SOTA chasing and help students to focus on insights. Inspired by result-blind reviewing in Section 4.4, we might refer to this as "result-blind grading".

Test collections[38] and **runs/submissions** – typically combined with novel evaluation methodologies – are the main resources resulting from shared tasks or evaluation campaigns. Integrating the resulting test collections into tools such as `Hugging Face datasets` [34], `ir_datasets` [38] or `EvALL` [3] allows for unified access to a wide range of datasets. Furthermore, some **software components** such as `Anserini` [52], `Capreolus` [54], `PyTerrier` [39], `OpenNIR` [37], etc., can directly load test collections integrated into `ir_datasets` which substantially simplifies data wrangling for scholars of all levels. For instance, PyTerrier allows for defining end-to-end experiments, including significance tests and multiple-test correction, using a declarative pipeline and is already used in research and teaching alike (e.g., in a master course with 240 students [39]). Other resources for performance modeling and prediction in RS, IR, and NLP can also be found in the manifesto of a previous Dagstuhl Perspectives Workshop [22]. The broad availability of such resources makes it tremendously easier to replicate and reproduce approaches that were submitted to a shared task (challenge) before. Further, it lowers the entry barrier to experiment with a wider set of datasets and approaches across domains as switching between collections will be easy. New test collections can be added with limited effort. Still, further promoting the practice of sharing code and documentation,[39] or using software submissions with tools such as TIRA [24, 44] in shared tasks is important.

Combining and integrating the resources listed above 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 [18]. Continuously maintaining and promoting the integration of test collections and up-to-date best practices for shared tasks into a 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 is streamlined by ECIR, CLEF, and ESSIR (see Section 3.14).

4.3.3.2 Initiatives

We have identified a range of actors, and we argue that addressing the problems around education requires a number of different initiatives some of which target one particular type of actor but more commonly offer benefits for different groups. These initiatives should not

[38] In IR, an offline test collection is typically composed of a set of topics, a document collection, and a set of relevance judgments.

[39] `https://www.go-fair.org/fair-principles/`

be seen in isolation as our vision is in line with what has been proposed in Section 3.14 which calls for coordinated action around education, evaluation, and research. Here we will discuss instruments we consider to be essential on that path. There is no particular order in this discussion other than starting with well-established popular concepts.

Summer schools are a key instrument primarily aimed at graduate students. ESSIR[40] is a prime example of a summer school focusing on delivering up-to-date educational content in the field of IR; the Recommender Systems Summer School is organized in a similar manner focusing on RS. Beyond the technical content, summer schools do also serve the purpose of community-building involving different actors, namely students and scholars. Annually organized summer schools appear most effective as they make planning easier by integrating them into the annual timeline of IR- and RS-related events. This is in line with the *flow-wise* vision discussed earlier in Section 3.14.

Summer schools also provide a good setting to embed (research-focused) **Mentoring** programs and **Doctoral Consortia**. This allows PhD students as well as early-career researchers to learn from experts in the field outside their own institutions. Both instruments are well-established in the field. However, even though the established summer schools are repeatedly organized, these often happen on an irregular basis (sometimes yearly, sometimes with longer breaks) and using different formats. This irregular setting makes it difficult to integrate it into a PhD student's journey from the outset. Currently, Mentoring is often merely a by-product of other initiatives such as Summer Schools and Doctoral Consortia. It may be a fruitful path to see mentoring programs as an independent (yet, not isolated) initiative. For instance, the "Women in Music Information Retrieval (WiMIR) Mentoring program"[41] sets an example of a sustainable initiative that is organized independently of other initiatives and on yearly basis. A similar format seems a fruitful path to follow in the IR and RS communities, where it is advisable to facilitate exchange across (sub-)disciplines and open up the initiative to the entire community. We note that – similar to the WiMIR – mentoring may not only address PhD students but is well suited also for later-career stages.

While the IR and RS communities have a tradition of research-topic-driven **Tutorials** as part of the main conferences, **Courses** that address skills and practices beyond research topics (similar to courses hosted by the CHI conference[42]) would be an additional fruitful path to follow. Such courses may, for instance, address specific research and evaluation methods on an operational level[43] or how to write better research papers for a specific outlet or community[44]. With regard to support in writing better papers, see also Section 4.5.

In Bachelor and Master education, more resources in the form of Formal Educational Materials could be developed. For example, students could benefit from The Black Mirror Writers' Room exercise[45] which helps convey ethical thinking around the use of technology. Participants choose current technologies that they find ethically troubling and speculate about what the next stage of that technology might be. They work collaboratively as if they were science fiction writers, and use a combination of creative writing and ethical speculation to consider what protagonist and plot would be best suited to showcase the potential negative

[40] https://www.essir.eu
[41] https://wimir.wordpress.com/mentoring-program/
[42] https://chi2023.acm.org/for-authors/courses/accepted-courses/
[43] See, e.g., CHI 2023's C12: Empirical Research Methods for Human-Computer Interaction https://chi2023.acm.org/for-authors/courses/accepted-courses/#C12, C18: Statistics for CHI https://chi2023.acm.org/for-authors/courses/accepted-courses/#C18
[44] See, e.g., CHI 2021's C02: How to Write CHI Papers [42]
[45] https://discourse.mozilla.org/t/the-black-mirror-writers-room/46666

consequences of this technology. They plot episodes, but then also consider what steps they might take now (in regulation, technology design, social change) that might result in *not* getting to this negative future. More experienced Bachelor students and Master students could have assessments similar to paper reviews as part of their curriculum to practice critical thinking.

Topically relevant **Meetups** ranging from informal one-off meetings to more regular thematically structured events offer a much more flexible and informal way to learn about the field. Unlike summer schools they bring together the community for an evening and cater for a much more diverse audience involving *all* actors with speakers as well as attendees from industry, academia and beyond. Talks range from specific use cases of IR in the industry (e.g., search at Bloomberg), to the latest developments in well-established tools (such as Elasticsearch) to user studies in realistic settings. There is a growing number of information-retrieval-related and recommender-systems-related Meetups[46] and many of which have become more accessible recently as they offer virtual or hybrid events. Meetups offer a low entry barrier in particular for students at all levels of education and they help participants obtain a more holistic view of the challenges of building and evaluating IR and RS applications. Loosely incorporating Meetups in the curriculum, in particular when there is alignment with teaching content (e.g., **joint seminars**), has been demonstrated to be effective in our own experience. These joint initiatives may go beyond the dissemination of content, but also involve practitioners as well as decision-makers in terms of facilitating (or hindering) strategic alliances or setting strategic themes.

Knowledge Transfer through **collaboration between industry and academia** is another instrument offering a mutually beneficial collaboration between three key actors: PhD students, academic scholars, and practitioners in the industry. By tackling real-world problems (as defined by the industrial partner) using state-of-the-art research approaches in the fields of IR and RS (as provided by the academic partner) knowledge does not just flow in one direction but both ways. In the context of our discussion, this is an opportunity to gain insights into evaluation methods and concerns in the industry. There are well-established frameworks to foster knowledge transfer such as Knowledge Transfer Partnerships[47] in the UK with demonstrated impact in IR[48] and beyond.

Knowledge transfer should also be facilitated and supported at a higher level at conferences and workshops. This is where the RS community is particularly successful in attracting industry contributions to the RecSys conference series. In IR, there is still an observable gap between key academic conferences such as SIGIR and practitioners' events like Haystack (*"the conference for improving search relevance"*[49]). The annual Search Solutions conference is an example of a successful forum to exchange ideas between all different actors.[50]

With a view to improving evaluation practices in the long-term, the reviewing process and practices play an important role. Hence, **addressing reviewers and editors** is essential. Reviewers are important actors in shaping what papers will be published and which not. And it is essential that good evaluation is acknowledged and understood while poorly evaluated

[46] See, e.g., `https://opensourceconnections.com/search-meetups-map/`, `https://recommender-systems.com/community/meetups/`
[47] `http://ktp.innovateuk.org`
[48] `https://www.gov.uk/government/news/media-tracking-firm-wins-knowledge-transfer-partnership-2015`
[49] `https://haystackconf.com`
[50] `https://www.bcs.org/membership-and-registrations/member-communities/information-retrieval-specialist-group/conferences-and-events/search-solutions/`

papers are not let through. Similarly, it is crucial to have reviewers who acknowledge and understand information retrieval and recommendation problems in their broader context (e.g., tasks, users, organizational value, user interface, societal impact) and review papers accordingly. Hence, it is essential to develop educational initiatives concerning evaluation that address current and future reviewers (and editors) accordingly. Promising initiatives include the following:

- Clear reviewer guidelines acknowledging the wide spectrum of evaluation methodology and the holistic view on information retrieval and recommendation problems. For example, CHI[51] and Association for Computational Linguistics (ACL)[52] provide detailed descriptions of what needs to be addressed and considered in a review and what steps to take.[53] Care has to be taken, though, that such guidelines are kept concise to not overwhelm people before even starting to read. Further suggestions on results-blind reviewing and guidance for authors can be found in Sections 4.4 and Section 4.5 respectively.
- Next to reviewers, meta-reviewers and editors is another entity to address, which can be done in a similar manner as addressing reviewers. These senior roles can have strong momentum in inducing change – but have a strong power position in preventing it. Stronger resistance might be expected on that (hierarchical) level. Seemingly, only a few conferences and journals – for instance, ACL[54] – seem to offer clear guidelines for the meta-reviewing activity.
- Similar to courses on research methods or addressing paper-writing skills, it is advisable to provide courses that specifically address how to peer review.[55]
- Mentored reviewing is another promising initiative to have better reviews that, on the one hand, better assess submitted papers and, on the other hand, are more constructive to induce better evaluation practices for future research. Mentored reviewing programs are, for instance, established in Psychology[56]. The MIR community[57] has a New-to-ISMIR mentoring program[58] that mainly addresses paper-writing for people who are new to the community but will likely also have an impact on reviewing practices. Similar programs could be established in the IR and RS communities with a particular focus on evaluation aspects. It is worthwhile to note that a recent study (in ML and AI) indicates that novice reviewers provide valuable contributions in the reviewing process [47].
- Summer schools mainly address (advanced) students and are also a good opportunity to include initiatives addressing reviewing.

General Public Dissemination is another important aspect that needs to be addressed. Communication in the lay language of our field is very important. Editing and curating better relevant Wikipedia pages on evaluation measures for information retrieval[59] and recommender systems[60] will increase the potential of reaching a wider audience, including potential future students. Other actions can concern publishing papers in magazines

[51] ACM CHI Conference on Human Factors in Computing Systems
[52] Association for Computational Linguistics
[53] CHI 2023 Guide to reviewing papers `https://chi2023.acm.org/submission-guides/guide-to-reviewing-papers/`; ACL's How to Review for ACL Rolling Review `https://aclrollingreview.org/reviewertutorial`; Ken Hinckley's comment on what excellent reviewing is [28].
[54] ACL's Action Editor Guide to Meta-Reviewing `https://aclrollingreview.org/aetutorial`
[55] `https://chi2023.acm.org/for-authors/courses/accepted-courses/#C16`
[56] `https://www.apa.org/pubs/journals/cpp/reviewer-mentoring-program`
[57] `https://www.ismir.net`
[58] `https://ismir2022.ismir.net/diversity/mentoring`
[59] `https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)` [Accessed: 20-Jan-2023]
[60] `https://en.wikipedia.org/wiki/Recommender_system#Evaluation` [Accessed: 20-Jan-2023]

Table 3 Actors generating or consuming resources and initiatives related to education in evaluation for IR and RS. ✓and (✓) indicate primary and secondary actors, respectively.

Actors:	Students	Educators	Scholars	Practitioners	Decision-makers
			Resources		
Teaching Materials	✓	✓			(✓)
Shared tasks/challenges/competitions	✓	✓	✓	✓	
Test collections & runs/submissions	✓	✓	✓	✓	
Software (components)	✓	✓	✓	✓	
			Initiatives		
Mentoring: Summer schools and Doctoral Consortia	✓		✓	(✓)	
Tutorials and courses	✓		✓	✓	
Meetups	(✓)	(✓)	✓	✓	✓
Joint seminars	✓	✓		✓	(✓)
Collaboration between industry and academia	✓		✓	✓	
Reviewing	(✓)		✓		
General public dissemination	(✓)	(✓)	✓	✓	✓

with a wider and differentiated audience, such as *Communications of the ACM*[61], *ACM Inroads*[62], *ACM XRDS: Crossroads*[63], *IEEE Spectrum*[64]. One of the final goals is to make IR and RS more popular to both attract students to the field and grow a healthy ecosystem of professionals at various levels.

We have described actors, resources, and initiatives that we think are worth considering in moving forward as a community towards creating more awareness, as well as sharing and transferring knowledge on experimental evaluation for IR and RS. We summarize the participation (either primary or secondary actors) in generating and consuming these resources and initiatives in Table 3. This is not intended as a definitive list but aimed to represent the primary and secondary actors which are involved.

4.3.4 Challenges & Outlook

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 possibilities) for the different actors: students, educators, scholars, practitioners, and decision-makers. We will investigate each actor in turn.

Scholars. As has also been identified in a previous Dagstuhl Seminar [22], 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 are therefore

[61] `https://cacm.acm.org/`
[62] `https://inroads.acm.org/`
[63] `https://xrds.acm.org/`
[64] `https://spectrum.ieee.org/`

also at higher risk of resulting in hypotheses that cannot be clearly rejected (non-significant results), leading to fear of criticism and rejection from paper reviewers. There are some proponents of Equivalence Testing [33][65] and Bayesian Analysis [49] 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 defined.[66] These guidelines may inspire educators on how to characterize the role of these tools in learning & teaching environments, including assessment design and plagiarism policies[67].

In addition, a current culture of 'publish or perish' incentivizes short-term and incremental findings[68], over more holistic thinking and thoughtful comparative analysis. The problem of 'SOTA-chasing' has also been discussed in other research areas, e.g., in NLP [11]. 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 plan[69]. 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 as indicated earlier in Section 4.3.3. Embedding participation to shared tasks and competitions (e.g., CLEF labs or 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.

---

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

[66] For instance, see the ACL 2023 Policy on AI Writing Assistance: https://2023.aclweb.org/blog/ACL-2023-policy/.

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

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

[69] 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

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*[70] 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.

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 [21, 27] 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[71], its extension in the new 10-year ROBUST initiative[72], and the Australian Centre of Excellence for Automated Decision-Making and Society (ADM+S)[73], 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. [31, 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.

[70] https://www.ischools.org
[71] https://icai.ai/
[72] https://icai.ai/ltp-robust/
[73] https://www.admscentre.org.au/

4.3.5 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. 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.

References

1 *Reproducibility of Data-Oriented Experiments in e-Science (Dagstuhl Seminar 16041)*, volume 6, 2016.

2 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.

3 Enrique Amigó, Jorge Carrillo de Albornoz, Mario Almagro-Cádiz, Julio Gonzalo, Javier Rodríguez-Vidal, and Felisa Verdejo. Evall: Open access evaluation for information access systems. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1301–1304. ACM, 2017.

4 Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy Lin, editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 601–610. ACM, 2009.

5 Ahmed Hassan Awadallah, Rosie Jones, and Kristina Lisa Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 221–230. ACM, 2010.

6 Christine Bauer and Eva Zangerle. Leveraging multi-method evaluation for multi-stakeholder settings. In Oren Sar Shalom, Dietmar Jannach, and Ido Guy, editors, *Proceedings of the 1st Workshop on the Impact of Recommender Systems co-located with 13th ACM Conference on Recommender Systems, ImpactRS@RecSys 2019), Copenhagen, Denmark, September 19, 2019*, volume 2462 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

7 Jöran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Research and Advanced Technology for Digital Libraries – 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings*, volume 9316 of *Lecture Notes in Computer Science*, pages 153–168. Springer, 2015.

8 Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of touché 2022: Argument retrieval. In Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 13th International Conference of the CLEF Association, CLEF 2022, Bo-*

logna, Italy, September 5-8, 2022, Proceedings, volume 13390 of *Lecture Notes in Computer Science*, pages 311–336. Springer, 2022.

9 Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of touché 2021: Argument retrieval. In K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings*, volume 12880 of *Lecture Notes in Computer Science*, pages 450–467. Springer, 2021.

10 Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and offline web search evaluation metrics. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 15–24. ACM, 2017.

11 Kenneth Ward Church and Valia Kordoni. Emerging trends: Sota-chasing. *Nat. Lang. Eng.*, 28(2):249–269, 2022.

12 Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. Threats of a replication crisis in empirical computer science. *Commun. ACM*, 63(8):70–79, 2020.

13 Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk, editors, *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 101–109. ACM, 2019.

14 Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. Viewpoint diversity in search results. In *ECIR*, 2023.

15 Tim Draws, Nava Tintarev, and Ujwal Gadiraju. Assessing viewpoint diversity in search results using ranking fairness metrics. *SIGKDD Explor.*, 23(1):50–58, 2021.

16 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 Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 295–305. ACM, 2021.

17 Michael D. Ekstrand, Michael Ludwig, Joseph A. Konstan, and John Riedl. Rethinking the recommender research ecosystem: reproducibility, openness, and lenskit. In Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius, editors, *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, pages 133–140. ACM, 2011.

18 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.

19 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.

20 Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. Suppressing the search engine manipulation effect (SEME). *Proc. ACM Hum. Comput. Interact.*, 1(CSCW):42:1–42:22, 2017.

21 Nicola Ferro. What happened in CLEF \ldots for a while? In Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9-12, 2019, Proceedings*, volume 11696 of *Lecture Notes in Computer Science*, pages 3–45. Springer, 2019.

22 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.

23 Nicola Ferro and Mark Sanderson. How do you test a test?: A multifaceted examination of significance tests. In K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang, editors, *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 – 25, 2022*, pages 280–288. ACM, 2022.

24 Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, 2023. Springer.

25 Carlos Alberto Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manag. Inf. Syst.*, 6(4):13:1–13:19, 2016.

26 Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1644–1651. AAAI Press, 2018.

27 D. K. Harman and E. M. Voorhees, editors. *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge (MA), USA, 2005.

28 Ken Hinckley. So you're a program committee member now: On excellence in reviews and meta-reviews and championing submitted work that has merit, 2016.

29 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*, Como, Italy, 2017.

30 Dietmar Jannach and Christine Bauer. Escaping the mcnamara fallacy: Towards more impactful recommender systems research. *AI Mag.*, 41(4):79–95, 2020.

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

32 Joseph A. Konstan and Gediminas Adomavicius. Toward identification and adoption of best practices in algorithmic recommender systems research. In Alejandro Bellogín, Pablo Castells, Alan Said, and Domonkos Tikk, editors, *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, RepSys 2013, Hong Kong, China, October 12, 2013*, pages 23–28. ACM, 2013.

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

34 Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In Heike Adel and Shuming Shi, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics, 2021.

35 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 Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2283–2287. ACM, 2021.

36 Marianne Lykke, Ann Bygholm, Louise Bak Søndergaard, and Katriina Byström. The role of historical and contextual knowledge in enterprise search. *J. Documentation*, 78(5):1053–1074, 2022.

37 Sean MacAvaney. Opennir: A complete neural ad-hoc ranking pipeline. In James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang, editors, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 845–848. ACM, 2020.

38 Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. Simplified data wrangling with ir_datasets. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2429–2436. ACM, 2021.

39 Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. Pyterrier: Declarative experimentation in python from BM25 to dense retrieval. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 – 5, 2021*, pages 4526–4533. ACM, 2021.

40 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 Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel, editors, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 463–472. ACM, 2016.

41 Ilya Markov and Maarten de Rijke. What should we teach in information retrieval? *SIGIR Forum*, 52(2):19–39, 2018.

42 Lennart E. Nacke. How to write CHI papers, online edition. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, and Takeo Igarashi, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8-13, 2021, Extended Abstracts*, pages 126:1–126:3. ACM, 2021.

43 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 Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz, editors, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017*, pages 209–216. ACM, 2017.

44 Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. TIRA integrated research architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 123–160. Springer, 2019.

45 Tetsuya Sakai. Laboratory experiments in information retrieval – sample sizes, effect sizes, and statistical power. 40, 2018.

46 Mark Sanderson, Monica Lestari Paramita, Paul D. Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 555–562. ACM, 2010.

47 Ivan Stelmakh, Nihar B. Shah, Aarti Singh, and Hal Daumé III. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *CoRR*, abs/2011.14646, 2020.

48 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 Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura, editors, *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, pages 23–32. ACM, 2020.

49 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.

50 Ellen M. Voorhees. Coopetition in IR research. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, page 3. ACM, 2020.

51 Ryen W. White. *Interactions with Search Systems*. Cambridge University Press, 2016.

52 Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1253–1256. ACM, 2017.

53 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 Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1129–1132. ACM, 2019.

54 Andrew Yates, Siddhant Arora, Xinyu Zhang, Wei Yang, Kevin Martin Jose, and Jimmy Lin. Capreolus: A toolkit for end-to-end neural ad hoc retrieval. In James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang, editors, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 861–864. ACM, 2020.

55 Eva Zangerle and Christine Bauer. Evaluating recommender systems: Survey and framework. *ACM Comput. Surv.*, 55(8):170:1–170:38, 2023.

56 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 Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 379–388. ACM, 2020.

57 J. Zobel. When measurement misleads: The limits of batch assessment of retrieval systems. *SIGIR Forum*, 56(1), 2022.

4.4 Results-blind Reviewing

Joeran Beel (University of Siegen, DE, joeran.beel@uni-siegen.de)
Timo Breuer (Technische Hochschule Köln, DE, timo.breuer@th-koeln.de)
Anita Crescenzi (University of North Carolina at Chapel Hill, US, amcc@unc.edu)
Norbert Fuhr (University of Duisburg-Essen, DE, norbert.fuhr@uni-due.de)
Meijie Li (University of Duisburg-Essen, DE, meijie.li@uk-essen.de)

4.4.1 Motivation

Campbell and Stanley defined experiments as "that portion of research in which variables are manipulated and their effects upon other variables observed" (p. 1 in [1])." 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 [7]. 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 [4] in case the results are impressive. 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 [2].

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. [8] 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.