Overcoming Methodological Challenges in Information Retrieval and Recommender Systems through Awareness and Education[*]

Christine Bauer, Maik Fröbe, Dietmar Jannach, Udo Kruschwitz,
Paolo Rosso, Damiano Spina, Nava Tintarev

Contributing authors: c.bauer@uu.nl; maik.froebe@uni-jena.de; dietmar.jannach@aau.at; udo.kruschwitz@ur.de; prosso@dsic.upv.es; damiano.spina@rmit.edu.au; n.tintarev@maastrichtuniversity.nl;

Keywords: Information Retrieval, Recommender Systems, Methodology, Evaluation

1 Background & Motivation

In recent years, we have observed a substantial increase in research in information retrieval (IR) and recommender systems (RS). To a large extent, this increase is fueled by progress in machine learning (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 machine learning 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 were published that 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 [1–4].

Besides these issues in offline experiments, increasingly questions concerning the *ecological validity* of the reported findings are raised. Ecological validity measures how generalizable experimental findings are to the real world. An example of this problem

[*] Part of the report of Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education (https://arxiv.org/pdf/2305.01509)

in information retrieval is the known problem of mismatch between offline effectiveness measurement and user satisfaction measured with online experimentation [5–9] 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 [10]. This phenomenon—often referred to as the *search engine manipulation effect*—has been demonstrated for both politics [10, 11] and health [12, 13]. 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 [17–19].

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 [1, 20, 21] and *(ii)* how effective we are in sharing and translating the knowledge we currently have for doing IR and RS experimentation [22, 23]. 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. [24–26]. Studies on long-term impact furthermore need to consider the wider scope of stakeholders [19, 27]. Moreover, for several types of information access systems, the specific and potentially competing interests of multiple stakeholders have to be taken into account [27]. 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 [19, 28].

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 [6, 8, 29, 30]. 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 addressing the much broader problem of reproducibility[1] and replicability [2] we face in Computer Science [31, 32] in general and in AI in particular [33].

This chapter is organized as follows: Next, in Section 2 we briefly review which kinds of actors may benefit from better education in information access system experimentation. Afterwards, in Section 3, we provide concrete examples of what we can do in terms of concrete resources and initiatives to increase the awareness and knowledge level for the different actors. Finally, in Section 4, we sketch main challenges that we may need to be aware of when implementing some of the described educational initiatives.

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, to 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 education of experimentation in information access, who are listed below. Note that this categorization is not exhaustive nor exclusive, as actors may have multiple roles.

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

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

Students

This category embraces the different stages of the academic training. Starting from students enrolled in IR & RS courses [34], including, for instance, undergraduate students in Computer Science degrees and Master's students in Data Science, Artificial Intelligence, 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.

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 of having a better understanding of IR and RS core concepts in evaluation and experimentation.

Figure 1 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 direct impact on how and to which extent experimentation in IR and RS is included in Data Science, Computer Science, Human-Computer Interaction, and Artificial Intelligence programs.

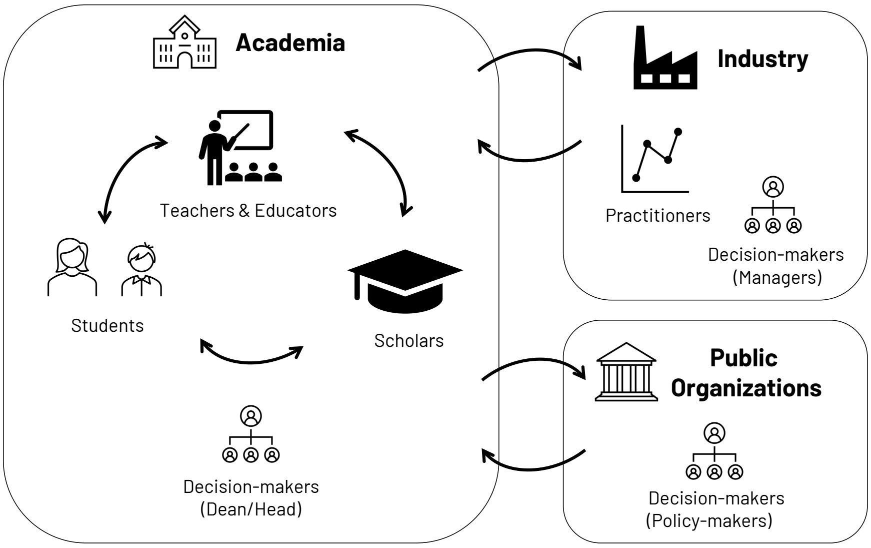

Fig. 1 Interaction among actors involved in IR and RS experimental education.

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.

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,[3] 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

[3] 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/

available and shared across universities, summer schools, and meetups.[4] Finally, having the IR and RS community actively contribute to the curation of material in sources that 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,[5] Elasticsearch,[6] Surprise,[7] Implicit,[8] etc. (see the report by Ferro et al. [35] 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 [34]. To date, also online evaluation falls short in such resources although it is essential in the spectrum of evaluation types [34]. 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 ethics aspects, both in designing experiments and also in how to evaluate them, is also important.[9]

Moreover, the participation in **shared tasks (challenges or competitions)** of evaluation campaigns in IR (e.g., TREC,[10] CLEF,[11] NTCIR,[12] or FIRE[13]) and RS (e.g., the yearly ACM RecSys challenges[14]) should be fostered. To facilitate the participation of 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 [36, 37] hosted at CLEF). At the same time, it is important to critically reflect with students on the limitations and dangers of competitions [38] and encourage 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 [19, 39]. 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'.

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

[5] https://solr.apache.org/

[6] https://www.elastic.co/es/elasticsearch/

[7] https://surpriselib.com/

[8] https://implicit.readthedocs.io

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

[10] https://trec.nist.gov/

[11] https://www.clef-initiative.eu/

[12] https://research.nii.ac.jp/ntcir/

[13] https://fire.irsi.res.in/fire/

[14] https://recsys.acm.org/challenges/

Test collections[15] 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` [40], `ir_datasets` [41] or `EvALL` [42] allows for unified access to a wide range of datasets. Furthermore, some **software components** such as `Anserini` [43], `Capreolus` [44], `PyTerrier` [45], `OpenNIR` [46], 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 [45]). 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 [47]. 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,[16] or using software submissions with tools such as TIRA [48, 49] 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 [50]. 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.

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 targeting one particular type of actor but more commonly offering benefits for different groups. These initiatives should not be seen in isolation as our vision is in line with what has been proposed in Section 3.14 which calls for a 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[17] is a prime example of a summer school focusing on delivering up-to-date educational content in the field of Information Retrieval; the Recommender Systems Summer School is organized in a similar manner focusing on recommender systems.

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

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

[17] https://www.essir.eu

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 in 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 in 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"[18] sets an example of a sustainable initiative that is organized independent 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 opening up the initiative to the entire community. We note that—similar to the WiMIR—mentoring may not only address PhD students but is a 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[19]) would be an additional fruitful path to follow. Such courses may, for instance, address specific research and evaluation methods on an operational level[20] or how to write better research papers for a specific outlet or community[21]. 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[22] 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 potential negative 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 review as part of their curriculum to practice critical thinking.

[18] https://wimir.wordpress.com/mentoring-program/

[19] https://chi2023.acm.org/for-authors/courses/accepted-courses/

[20] 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 HCI https://chi2023.acm.org/for-authors/courses/accepted-courses/#C18

[21] See, e.g., CHI 2021's C02: How to Write CHI Papers [51]

[22] https://discourse.mozilla.org/t/the-black-mirror-writers-room/46666

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 industry (e.g., search at Bloomberg), to 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[23] 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 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 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 industry. There are well-established frameworks to foster knowledge transfer such as Knowledge Transfer Partnerships[24] in the UK with demonstrated impact in IR[25] 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"*[26]). The annual Search Solutions conference is an example of a successful forum to exchange ideas between all different actors.[27]

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 papers are not let through. Similarly, it is crucial

[23] See, e.g., https://opensourceconnections.com/search-meetups-map/, https://recommender-systems.com/community/meetups/

[24] http://ktp.innovateuk.org

[25] https://www.gov.uk/government/news/media-tracking-firm-wins-knowledge-transfer-partnership-2015

[26] https://haystackconf.com

[27] https://www.bcs.org/membership-and-registrations/member-communities/information-retrieval-specialist-group/conferences-and-events/search-solutions/

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[28] and ACL[29] provide detailed descriptions on what needs to be addressed and considered in a review and what steps to take.[30] Care has to be taken, though, that such guidelines are kept concise to not overwhelm people before even starting to read.
- Next to reviewers, meta-reviewers and editors are 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[31]—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.[32]
- 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[33]. The MIR community[34] has a New-to-ISMIR mentoring program[35] 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 the ML and AI) indicates that novice reviewers provide valuable contributions in the reviewing process [53].
- 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 lay language of our field is very important. Editing and curating better relevant Wikipedia pages on evaluation measures for information retrieval[36] and recommender systems[37] will increase the potential of reaching a wider audience, including potential future students. Other action can concern publishing papers in magazines with a wider and differentiated audience, such as *Communications*

[28] ACM CHI Conference on Human Factors in Computing Systems

[29] Association for Computational Linguistics

[30] 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 [52].

[31] ACL's Action Editor Guide to Meta-Reviewing https://aclrollingreview.org/aetutorial

[32] https://chi2023.acm.org/for-authors/courses/accepted-courses/#C16

[33] https://www.apa.org/pubs/journals/cpp/reviewer-mentoring-program

[34] https://www.ismir.net

[35] https://ismir2022.ismir.net/diversity/mentoring

[36] https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval) [Accessed: 20-Jan-2023]

[37] https://en.wikipedia.org/wiki/Recommender_system#Evaluation [Accessed: 20-Jan-2023]

Table 1 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	(✓)	(✓)	✓	✓	✓

of the ACM[38], *ACM Inroads*[39], *ACM XRDS: Crossroads*[40], *IEEE Spectrum*[41]. 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 1. This is not intended as a definitive list, but aimed to represent the primary and secondary actors which are involved.

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 [35], 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 also at higher risk of resulting in hypotheses that cannot be clearly rejected (non-significant results), leading to fear of criticism and rejection

[38] https://cacm.acm.org/
[39] https://inroads.acm.org/
[40] https://xrds.acm.org/
[41] https://spectrum.ieee.org/

from paper reviewers. There are some proponents of Equivalence Testing [54][42] and Bayesian Analysis [55] in Psychology which may also be useful in Computer Science.

As large language models (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.[43] These guidelines may inspire educators on how to characterize the role of these tools in learning & teaching environments, including assessment design and plagiarism policies[44].

In addition, a current culture of 'publish or perish' incentivizes short-term and incremental findings[45], 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 [38]. 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[46]. In Bachelor and Masters 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 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 overcoming this challenge—although the synchronization between the semester and participation timelines may not be straightforward.

Finally, there is a growing number of experiments of developing multi-disciplinary curricula—with the appreciation that different disciplines bring to such a program.

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

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

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

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

[46] 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

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*[47] 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 participate to initiatives such as shared tasks/challenges [56, 57] 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[48], its extension in the new 10-year ROBUST initiative[49], and the Australian Centre of Excellence for Automated Decision-Making and Society (ADM+S)[50], 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 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. [58, 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 centers, can play an important role towards developing competency of students and educators. Making best practices in evaluation available as a

[47] https://www.ischools.org
[48] https://icai.ai/
[49] https://icai.ai/ltp-robust/
[50] https://www.admscentre.org.au/

commodity through these channels will require making resources more accessible for non-experts in IR and RS.

5 Concluding Remarks

Education and dissemination represent key pillars to overcome 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] Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: 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, pp. 601–610 (2009)

[2] Ferrari Dacrema, M., Cremonesi, P., Jannach, D.: Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In: Proceedings of the 2019 ACM Conference on Recommender Systems (RecSys 2019), Copenhagen (2019)

[3] Sun, Z., Yu, D., Fang, H., Yang, J., Qu, X., Zhang, J., Geng, C.: Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison. In: Proceedings of the 14th ACM Conference on Recommender Systems. RecSys '20, pp. 23–32 (2020)

[4] Yang, W., Lu, K., Yang, P., Lin, J.: 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, pp. 1129–1132 (2019)

[5] Chen, Y., Zhou, K., Liu, Y., Zhang, M., Ma, S.: 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, pp. 15–24. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3077136.3080804 . https://doi.org/10.1145/3077136.3080804

[6] Hassan, A., Jones, R., Klinkner, K.L.: 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, pp. 221–230. Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1718487.1718515 . https://doi.org/10.1145/1718487.1718515

[7] Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., Ma, S., Sun, J., Luo, H.: 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, pp. 463–472. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2911451.2911507 . https://doi.org/10.1145/2911451.2911507

[8] Sanderson, M., Paramita, M.L., Clough, P., Kanoulas, E.: 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, pp. 555–562. Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1835449.1835542 . https://doi.org/10.1145/1835449.1835542

[9] Zhang, F., Mao, J., Liu, Y., Xie, X., Ma, W., Zhang, M., Ma, S.: 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, pp. 379–388. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401162 . https://doi.org/10.1145/3397271.3401162

[10] Epstein, R., Robertson, R.E.: The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proceedings of the National Academy of Sciences **112**(33), 4512–4521 (2015) https://doi.org/10.1073/pnas.1419828112 . Accessed 2021-07-13

[11] Epstein, R., Robertson, R.E., Lazer, D., Wilson, C.: Suppressing the Search Engine Manipulation Effect (SEME). Proceedings of the ACM on Human-Computer Interaction **1**(CSCW), 1–22 (2017) https://doi.org/10.1145/3134677 . Accessed 2021-07-13

[12] Allam, A., Schulz, P.J., Nakamoto, K.: 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), 100 (2014) https://doi.org/10.2196/jmir.2642 . Accessed 2021-07-13

[13] Pogacar, F.A., Ghenai, A., Smucker, M.D., Clarke, C.L.A.: 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, pp. 209–216. ACM, Amsterdam The Netherlands (2017). https://doi.org/10.1145/3121050.3121074 . https://dl.acm.org/doi/10.1145/3121050.3121074 Accessed 2021-07-13

[14] Draws, T., Roy, N., Inel, O., Rieger, A., Hada, R., Yalcin, M.O., Timmermans, B., Tintarev, N.: Viewpoint diversity in search results. In: ECIR (2023)

[15] Draws, T., Tintarev, N., Gadiraju, U.: Assessing viewpoint diversity in search results using ranking fairness metrics. ACM SIGKDD Explorations Newsletter **23**(1), 50–58 (2021)

[16] Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., Timmermans, B.: 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, pp. 295–305 (2021)

[17] Beel, J., Langer, S.: 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, pp. 153–168 (2015)

[18] Gomez-Uribe, C.A., Hunt, N.: The Netflix recommender system: Algorithms, business value, and innovation. Transactions on Management Information Systems **6**(4), 13–11319 (2015)

[19] Jannach, D., Bauer, C.: Escaping the mcnamara fallacy: Towards more impactful recommender systems research. AI Magazine **41**(4), 79–95 (2020)

[20] Lin, J., Campos, D., Craswell, N., Mitra, B., Yilmaz, E.: 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, pp. 2283–2287. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463034 . https://doi.org/10.1145/3404835.3463034

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

[22] Ferro, N., Sanderson, M.: 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, pp. 280–288 (2022)

[23] Sakai, T.: Laboratory experiments in information retrieval. The information retrieval series **40** (2018) https://doi.org/10.1007/978-981-13-1199-4

[24] Lykke, M., Bygholm, A., Søndergaard, L.B., Byström, K.: The role of historical and contextual knowledge in enterprise search **78**(5), 1053–1074 (2022) https://doi.org/10.1108/JD-08-2021-0170

[25] White, R.W.: Interactions with Search Systems. Cambridge University Press, New

York, ??? (2016)

[26] Zangerle, E., Bauer, C.: Evaluating recommender systems: survey and framework. ACM Computing Surveys **55**(8) (2022) https://doi.org/10.1145/3556536

[27] Bauer, C., Zangerle, E.: Leveraging multi-method evaluation for multi-stakeholder settings. In: Shalom, O.S., Jannach, D., Guy, I. (eds.) 1st Workshop on the Impact of Recommender Systems, Co-located with 13th ACM Conference on Recommender Systems (ACM RecSys 2019). ImpactRS '19, vol. 2462 (2019). http://ceur-ws.org/Vol-2462/short3.pdf

[28] Jannach, D., Adomavicius, G.: 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)

[29] Ekstrand, M.D., Ludwig, M., Konstan, J.A., Riedl, J.T.: Rethinking the recommender research ecosystem: Reproducibility, openness, and lenskit. In: Proceedings of the Fifth ACM Conference on Recommender Systems. RecSys '11, pp. 133–140 (2011)

[30] Konstan, J.A., Adomavicius, G.: 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, pp. 23–28 (2013)

[31] Cockburn, A., Dragicevic, P., Besançon, L., Gutwin, C.: Threats of a replication crisis in empirical computer science. Commun. ACM **63**(8), 70–79 (2020) https://doi.org/10.1145/3360311

[32] Freire, J., Fuhr, N., Rauber, A. (eds.): Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. Dagstuhl Reports, Volume 6, Number 1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany, ??? (2016)

[33] Gundersen, O.E., Kjensmo, S.: State of the art: Reproducibility in artificial intelligence. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), pp. 1644–1651 (2018)

[34] Markov, I., Rijke, M.: What should we teach in information retrieval? SIGIR Forum **52**(2), 19–39 (2019) https://doi.org/10.1145/3308774.3308780

[35] Ferro, N., Fuhr, N., Grefenstette, G., Konstan, J.A., Castells, P., Daly, E.M., Declerck, T., Ekstrand, M.D., Geyer, W., Gonzalo, J., et al.: From evaluating to forecasting performance: How to turn information retrieval, natural language processing and recommender systems into predictive sciences. Dagstuhl manifestos (2018)

[36] Bondarenko, A., Fröbe, M., Kiesel, J., Syed, S., Gurcke, T., Beloucif, M., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2022: Argument Retrieval. In: Barrón-Cedeño, A., Martino, G.D.S., Esposti, M.D., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022). Lecture Notes in Computer Science. Springer, Berlin Heidelberg New York (2022)

[37] Bondarenko, A., Gienapp, L., Fröbe, M., Beloucif, M., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2021: Argument Retrieval. In: Candan, K., Ionescu, B., Goeuriot, L., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021). Lecture Notes in Computer Science, vol. 12880, pp. 450–467. Springer, Berlin Heidelberg New York (2021). https://doi.org/10.1007/978-3-030-85251-1_28 . https://link.springer.com/chapter/10.1007/978-3-030-85251-1_28

[38] Church, K.W., Kordoni, V.: Emerging trends: Sota-chasing. Nat. Lang. Eng. **28**(2), 249–269 (2022) https://doi.org/10.1017/S1351324922000043

[39] Voorhees, E.M.: Coopetition in IR research. In: Huang, J.X., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (eds.) Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, p. 3. ACM, ??? (2020). https://doi.org/10.1145/3397271.3402427 . https://doi.org/10.1145/3397271.3402427

[40] Lhoest, Q., Moral, A.V., Jernite, Y., Thakur, A., Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Sasko, M., Chhablani, G., Malik, B., Brandeis, S., Scao, T.L., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A.M., Wolf, T.: Datasets: A community library for natural language processing. In: Adel, H., Shi, S. (eds.) 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, pp. 175–184. Association for Computational Linguistics, ??? (2021). https://doi.org/10.18653/v1/2021.emnlp-demo.21 . https://doi.org/10.18653/v1/2021.emnlp-demo.21

[41] MacAvaney, S., Yates, A., Feldman, S., Downey, D., Cohan, A., Goharian, N.: Simplified data wrangling with ir_datasets. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, pp. 2429–2436. ACM, ??? (2021). https:

//doi.org/10.1145/3404835.3463254 . https://doi.org/10.1145/3404835.3463254

[42] Amigó, E., Carrillo-de-Albornoz, J., Almagro-Cádiz, M., Gonzalo, J., Rodríguez-Vidal, J., Verdejo, F.: EvALL: Open access evaluation for information access systems. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17, pp. 1301–1304. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3077136.3084145 . https://doi.org/10.1145/3077136.3084145

[43] Yang, P., Fang, H., Lin, J.: Anserini: Enabling the use of lucene for information retrieval research. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17, pp. 1253–1256 (2017)

[44] Yates, A., Arora, S., Zhang, X., Yang, W., Jose, K.M., Lin, J.: Capreolus: A toolkit for end-to-end neural ad hoc retrieval. In: Proceedings of the 13th International Conference on Web Search and Data Mining. WSDM '20, pp. 861–864 (2020)

[45] Macdonald, C., Tonellotto, N., MacAvaney, S., Ounis, I.: Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM '21, pp. 4526–4533 (2021)

[46] MacAvaney, S.: OpenNIR: A complete neural ad-hoc ranking pipeline. In: WSDM 2020 (2020)

[47] Ferro, N., Fuhr, N., Grefenstette, G., Konstan, J.A., Castells, P., Daly, E.M., Declerck, T., Ekstrand, M.D., Geyer, W., Gonzalo, J., Kuflik, T., Lindén, K., Magnini, B., Nie, J., Perego, R., Shapira, B., Soboroff, I., Tintarev, N., Verspoor, K., Willemsen, M.C., Zobel, J.: 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) https://doi.org/10.4230/DagMan.7.1.96

[48] Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., Potthast, M.: 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. Springer, Berlin Heidelberg New York (2023)

[49] Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World. The Information Retrieval Series. Springer, Berlin Heidelberg New York (2019). https://doi.org/10.1007/978-3-030-22948-1_5

[50] Elstner, T., Loebe, F., Ajjour, Y., Akiki, C., Bondarenko, A., Fröbe, M., Gienapp, L., Kolyada, N., Mohr, J., Sandfuchs, S., Wiegmann, M., Frochte, J., Ferro, N.,

Hofmann, S., Stein, B., Hagen, M., Potthast, M.: Shared Tasks as Tutorials: A Methodical Approach. In: 37th AAAI Conference on Artificial Intelligence (AAAI 2023). AAAI, ??? (2023)

[51] Nacke, L.E.: How to write chi papers, online edition. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21 (2021)

[52] Hinckley, K.: So You're a Program Committee Member Now: On Excellence in Reviews and Meta-Reviews and Championing Submitted Work That Has Merit (2016). https://www.microsoft.com/en-us/research/wp-content/uploads/2016/10/Excellence-in-Reviews-MobileHCI-2015-Web-Site.pdf

[53] Stelmakh, I., Shah, N.B., Singh, A., III, H.D.: Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. CoRR **abs/2011.14646** (2020) 2011.14646

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

[55] Doorn, J., Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N.J., Gronau, Q.F., Haaf, J.M., *et al.*: The jasp guidelines for conducting and reporting a bayesian analysis. Psychonomic Bulletin & Review **28**(3), 813–826 (2021)

[56] Ferro, N.: What Happened in CLEF... For a While? In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF '19, pp. 3–45. Springer, Cham (2019)

[57] Harman, D.K., Voorhees, E.M. (eds.): TREC. Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge (MA), USA, ??? (2005)

[58] Kahneman, D.: Thinking, Fast and Slow. Penguin, ??? (2011)