

Cite this document as follows:

Beel, Joeran, Timo Breuer, Anita Crescenzi, Norbert Fuhr, and Meijie Li. "Results-blind Reviewing." In: Report from Dagstuhl Seminar 23031 -- Frontiers of Information Access Experimentation for Research and Education (2023), pp. 67-73.

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)

License  Creative Commons BY 4.0 International license

© Joeran Beel, Timo Breuer, Anita Crescenzi, Norbert Fuhr, Meijie Li

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.

4.4.2 Current Situation and Gaps

As part of IR or RS conferences, the peer-reviewing process usually involves the review of the full paper using double-blinded reviewing, i.e., both authors and reviewers remain anonymous to each other. Before submission, authors are informed about possible reviewing criteria and areas of interest in the Call for Papers (CfP) that can be found on the conference website. Upon submission, the paper should contain all of the relevant information regarding the motivation, the research methodology or study design, the experimental results, and finally, a discussion that puts the results into context.

For each submission, usually, a group of three reviewers is assigned. All of them should align their reviews to those criteria mentioned in the CfP and, depending on the submission system, express their opinion in written text or by pre-defined answers regarding particular aspects. In addition, they can assign (overall) scores. The final decision is based on a discussion among reviewers, which is governed by an additional meta-reviewer, and consolidation with the program chairs.

Even though this traditional review model has been established for several years, it can imply negative impacts on the stakeholders or the scientific community as a whole. Under the assumption that reviewers overemphasize positive outcomes, the authors might be inclined to “search for” performance gains in system-oriented experiments at the cost of scientific rigor and reasoning. Even more, there is the danger of fraud or selecting positive outcomes, considering the need to publish in order to proceed in an academic career.

Alternatives to the traditional review process have emerged with an initial round of peer review of a manuscript with the results blinded or a study protocol and a subsequent round of peer review of the full paper including results. Table 4 shows the traditional peer review model with our recommended results-blind reviewing and two other variants, each of which we describe below. The Center for Open Science notes that, as of January

■ **Table 4** Comparison of traditional and emerging approaches to peer review: results-blind, preregistered reports, and registered reports.

	Traditional	Results-Blind	Preregistered	Registered Report
protocol preregistration	optional	optional	yes (in journal repository)	no
protocol publication (separate from research article)	no	no	no	yes
peer review of research protocol before data collection	no	no	yes	yes
peer review of paper with blinded results	no	yes	no	no
peer review of full paper	yes	yes (if in-principle acceptance)	yes with focus on results (if in-principle acceptance)	yes (if in-principle acceptance)
Example publication(s)	ACM SIGIR, ACM CHIIR	BMC Psychology	PLOS	Bio-logy PLOS ONE

2023, over 300 journals have adopted one or more variants of this approach.⁷⁴ In addition, several preliminary analyses of their implementation have been conducted and published (e.g., [3, 5, 8]).

A results-blind review involves an in-principle acceptance or rejection decision based on peer review of the paper *with the results blinded* from the reviewers (see the third column of Table 4). The reviewers can put more emphasis on judging the merits of the general motivation, the study design, and what kinds of scientific insights could be gained from the experiments. If the paper is accepted in-principle, it proceeds to a second stage of peer review of the *paper with the results* included for reviewers. The final decision about the acceptance is based on the second stage of the review in which the reviewers have access to the experimental outcomes.

Other peer-reviewing models have emerged in recent years as part of the growing awareness of preregistration^{75,76} and its adoption [6]. One such approach to peer review involves the review and in-principle acceptance of the study protocol including the methods and analysis plan before data is collected or analysis begins. Variants of this approach include preregistered research articles and registered reports for confirmatory research⁷⁷. Although preregistered reports and registered reports are typically used for confirmatory research, there are variants for exploratory research and some journals also use a separate approach for exploratory research projects which do not have a confirmatory component (e.g., an Exploratory Report article type in journal *Cortex*).

Preregistered research articles involve researchers submitting a research study protocol including the rationale and hypotheses, methodology including analysis plan, and materials

⁷⁴ <https://www.cos.io/initiatives/registered-reports>

⁷⁵ <https://www.cos.io/initiatives/prereg>

⁷⁶ <https://plos.org/open-science/preregistration/>

⁷⁷ For examples of how preregistered research articles and research reports have been implemented, see the summary provided by PLOS. <https://plos.org/open-science/preregistration/>

to a journal for review and simultaneous depositing into a repository often associated with the journal (see the fourth column of Table 4). The preregistered protocol is peer-reviewed with a focus on methods and the analytic approach, and a provisional in-principle acceptance conditional upon the execution of the study as designed. The researchers execute the study, analyze the results, and submit a full manuscript. After peer review of the new sections, the completed manuscript is published.

Registered Reports also involve submission and peer review of a study protocol (see the third column of Table 4). A key difference from preregistered articles is that accepted protocols are published immediately and a future article with the results of the study is given an in-principle acceptance. After the study execution, the full manuscript is submitted and reviewed.

4.4.3 Next Steps

We propose several changes to the reviewing processes for information access papers to reduce publication biases. Our recommendations are that information access scholarly community:

1. adopts a pilot test of results-blind reviewing for a conference or journal,
2. considers starting from our initial process recommendation for results-blind reviewing,
3. ask authors, conference organizers, and reviewers to place more emphasis within papers on the insights that can be gained from their research,
4. considers allowing additional space for additional details about study methodology, and
5. considers whether to implement a two-stage review process in which research proposals and/or preregistered research reports are reviewed with a tentative acceptance decision before data collection and analysis are conducted.

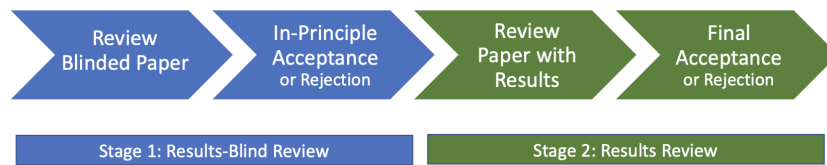
Each of these is described in more detail below.

Recommendation 1: Pilot test of results-blind reviewing in conference(s) or journal(s)

Our first and most important recommendation is that the information access research communities (i.e., IR and RS communities) adopt a results-blind approach to peer reviewing for conference(s) and/or journal(s). We recommend that the community start with a pilot test of results-blind reviewing in an established conference track, perhaps with a new paper track with an earlier deadline to allow for a two-stage review process. In results-blind reviewing, the authors submit two versions of their manuscript: one version of the paper with the full results, and one version with the results blinded. The two submitted versions are the basis of a results-blind reviewing process with two major stages (see Figure 7).

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

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



■ **Figure 7** Proposed two-stage process for results-blind reviewing (figure adapted from BMC⁷⁸)

Recommendation 2: Initial process recommendation for a results-blind reviewing pilot

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

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

- First, the CfP for the new track should be written. As the proposed results-blind reviewing process with two stages of review will take longer to complete, an earlier deadline for this track should be set.
- Criteria for both stages of the review (blinded and with results) should be defined. Special attention should be given to the criteria for changing an initial acceptance recommendation into a rejection.
- Author instructions for the results-blind reviewing track have to be formulated, describing not only the new reviewing criteria and process but also specific instructions on how to prepare the blinded version of an article. For the results-blind version of the paper, the authors will need to blind all mentions of the results (e.g., in the abstract, introduction, discussion, and conclusion in addition to in a results section) in a way that it is not technically possible to recover the blinded text. There should be a way for reviewers to easily determine the differences between the results-blind version of the paper and the one with the results.
- Reviewers for the results-blind reviewing track have to be recruited. In the beginning, additional or different expertise will be required for this track. A special introduction of training for the reviewers might be necessary in order to make them familiar with the new process and criteria.
- The reviewing software will need to be configured for multiple stages of review for the results-blind reviewing. In the first stage of reviewing, only the blinded version of the papers should be distributed to reviewers (see below for the process for reviewers).
- After the final decision by the PC, the authors will be provided with the review and informed about the final accept or reject decision. In the case of a rejection decision, authors should also be notified at which stage the paper was rejected.
- The organizers should give special recognition to the PC member of the track (on the conference Web site and in the proceedings)
- The success of the new track and the process should be evaluated.

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

- In the first stage, the reviewers are provided with the results-blind version of the submission and complete their review including a recommendation about the in-principle acceptance.

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

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

Recommendation 3: Emphasize insights in papers

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

Recommendation 4: Extra space for methods information

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

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

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

Recommendation 5: Consider a two-stage review process adapted from preregistered or registered reports

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

4.4.4 Conclusion

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

References

- 1 Donald T. Campbell and Julian C. Stanley. *Experimental and quasi-experimental designs for research*. Houghton Mifflin Company, Boston, 1963.
- 2 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.
- 3 Michael G. Findley, Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky. Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study. *Comparative Political Studies*, 49(13):1667–1703, 2016. Publisher: SAGE Publications Inc.
- 4 Norbert Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, 2017.
- 5 Daniel M. Maggin, Rachel E. Robertson, and Bryan G. Cook. Introduction to the special series on results-blind peer review: An experimental analysis on editorial recommendations and manuscript evaluations. *Behavioral Disorders*, 45(4):195–206, 2020.
- 6 Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.
- 7 William R Shadish, Thomas D Cook, and Donald T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company, New York, 2002.
- 8 Haley M. Woznyj, Kelcie Grenier, Roxanne Ross, George C. Banks, and Steven G. Rogelberg. Results-blind review: a masked crusader for science. *European Journal of Work and Organizational Psychology*, 27(5):561–576, 2018.