Report from Dagstuhl Seminar 22442

Toward Scientific Evidence Standards in Empirical Computer Science

Timothy Kluthe*[1], Brett A. Becker†[2], Christopher D. Hundhausen†[3], Ciera Jaspan†[4], Andreas Stefik†[5], and Thomas Zimmermann†[6]

1 **University of Nevada – Las Vegas, US.** tjkluthe@gmail.com
2 **University College Dublin, IE.** brett.becker@ucd.ie
3 **Oregon State University – Corvallis, US.** hundhaus@wsu.edu
4 **Google – Mountain View, US.** ciera@google.com
5 **University of Nevada – Las Vegas, US.** stefika@gmail.com
6 **Microsoft Corporation – Redmond, US.** tzimmer@microsoft.com

Abstract

Many scientific fields of study use formally established evidence standards during the peer review and evaluation process, such as Consolidated Standards of Reporting Trials (CONSORT) in medical research, the What Works Clearinghouse (WWC) used in education in the United States, or the APA Journal Article Reporting Standards (JARS) in psychology. The basis for these standards is community agreement on what to report in empirical studies. Such standards achieve two key goals. First, they make it easier to compare studies, facilitating replications, through transparent reporting and sharing of data, which can provide confidence that multiple research teams can obtain the same results. Second, they establish community agreement on how to report on and evaluate studies using different methodologies. The discipline of computer science does not have formalized evidence standards, even for major conferences or journals. This Dagstuhl Seminar has three primary objectives:

1. To establish a process for creating or adopting an existing evidence standard for empirical research in computer science.
2. To build a community of scholars that can discuss what a general standard should include.
3. To kickstart the discussion with scholars from software engineering, human-computer interaction, and computer science education.

In order to better discuss and understand the implications of such standards across several empirical subfields of computer science and to facilitate adoption, we brought together participants from a range of backgrounds; including academia and industry, software engineering, computer-human interaction and computer science education, as well as representatives from several prominent journals.

Funding: This material is based upon work supported by the National Science Foundation under Grant Numbers NSF HCC: 2106392 and NSF I-TEST: 2048356.

Seminar October 30–November 4, 2022 – http://www.dagstuhl.de/22442

2012 ACM Subject Classification General and reference → Empirical studies; Human-centered computing → Empirical studies in HCI; Social and professional topics → Computing education; General and reference → Reliability

Keywords and phrases Community evidence standards, Human factors

Digital Object Identifier 10.4230/DagRep.12.10.225

* Editorial Assistant / Collector
† Editor / Organizer

Toward Scientific Evidence Standards in Empirical Computer Science, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 225–240
Editors: Timothy Kluthe, Brett A. Becker, Christopher D. Hundhausen, Ciera Jaspan, Andreas Stefik, and Thomas Zimmerman

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Timothy Kluthe University of Nevada – Las Vegas, US, tjkluthe@gmail.com
Andreas Stefik University of Nevada – Las Vegas, US, stefika@gmail.com

The goals of the seminar *Toward Scientific Evidence Standards in Empirical Computer Science* were to establish a process for introducing evidence standards in computer science, build a community of scholars that discuss what a general standard would include and have enough diversity of background to have a good basis for the breadth of community needs across a range of computer science-related venues.

Over the first few days, we conducted a series of breakout groups and larger group discussions. In these, to introduce people to evidence standards, we reviewed several, including: APA JARS[1], WWC[2], and CONSORT[3]. The purpose was introductory and to scaffold for discussions on what could work across the breadth of computer science or in subareas. We also conducted a session looking at existing papers and noted the changes that would need to be made to fit the APA JARS standards. This exercise in particular was found to be useful by participants, as it made it clear that the the conversion is not particularly difficult, although it is aided by advanced planning for what might need to be collected during a study.

During the Dagstuhl, we also had several talks. These included an introductory talk by Andreas Stefik on evidence standards as a whole, telling the story of the well-known Tolbutamide drug and its influence on the medical field in regard to evidence standards. Christopher Hundhausen provided a talk on his experience with introducing reporting standards at ACM's Transactions on Computing Education (TOCE) (Section 3.2). Paul Ralph presented on the problems in scholarly peer review and how evidence standards could be a solution, along with a reviewing tool that he has developed (Section 3.3). Neil Ernst covered registered reports, their benefits to the transparency and quality of research, and his experience with introducing them at Mining Software Repositories (MSR) and Empirical Software Engineering (EMSE) (Section 3.4). Lastly, Kate Sanders et al. discussed a review on reviews, which spanned a variety of the computer science subfields. This included their observations on review criteria, ethical concerns in the peer review process and excerpts from interviews with conference chairs and journal editors that were relevant to the subject of the seminar (Section 3.5). Each of these gave insights into the process of adopting an evidence standard and some of the potential impacts of the status quo and potential changes (positive or negative).

Finally, after discussion, we identified four topics for breakout groups to brainstorm potential avenues toward actionable progress on goals: a deeper dive into how to write guidelines for more complex experiments like mixed-methods studies (Section 4.5), how can we measure the effects that evidence standards have both in reference in paper quality and community satisfaction (Section 4.6), what are the first steps towards community engagement as far as introducing the topic and adopting it (Section 4.7) and how to operationalize these standards in a way that is open source to allow for community control (Section 4.8). A final working group session went through some of the first steps could be made at conferences and a dissemination plan for how to start information the community about the topic (Section 4.9).

Overall, the seminar brought a range of computer science stakeholders up to speed on the state of evidence standards in the field, what could be gained by moving towards a domain-wide guidelines and started a discussion on how to spark the conversation in

various communities. A set of next steps on where and what to recommend and talk about with communities were set in motion, as well as plans for a collaborative position paper to introduce the topic to a wider audience.

References

1 American Psychological Association. APA Style Journal Article Reporting Standards (APA Style JARS). Accessed on December 12, 2022 from `https://apastyle.apa.org/jars`.
2 National Center for Education Evaluation and Regional Assistance. WWC | Find What Works. Accessed on December 12 , 2022 from `https://ies.ed.gov/ncee/wwc/`.
3 The CONSORT Group. CONSORT Transparent Reporting of Trials. Accessed on December 12, 2022 from `https://www.consort-statement.org/`.

2 Table of Contents

3 Overview of Talks

3.1 Seminar Goals and a Brief Introduction to Evidence Standards

Andreas Stefik (University of Nevada – Las Vegas, US)

The concept of checking our assumptions through the use of independently verifiable evidence has a long tradition in the sciences. While this is well known, throughout the 20th century, especially, external stakeholders to science have pushed the community toward using increasingly rigorous evidence. This is in part because science can have implications for the public or for public policy. In this talk, we will briefly introduce the topic of the Dagstuhl: evidence standards. In doing so, we will review an exemplar of why evidence standards came about and provide an overview of the week's activities.

3.1.1 Discussion

In this talk, there was a large focus on CONSORT[1], which is the set of evidence standards used in the medical community. The first point of discussion was to note that the standards required in the medical community are often centered around life or death problems which may not be quite the level of what is faced in computer science. A counterpoint was made that while computer science may not be facing life or death decisions, often times the papers and research can have an impact on public policy and economics (e.g., do we adopt a new programming language). Bad software design can be life threatening (e.g., self-driving cars), and social media algorithms can impact worldviews, which in turn impacts politics. Thus, unless we conclude computer science research does not matter for the real-world, any work that actually matters should plausibly be held to a more scientifically rigorous set of guidelines. Put succinctly, the greater the impact of the research, the more rigorous it needs to be to ensure the impact is desirable.

References

1 The CONSORT Group. CONSORT Transparent Reporting of Trials. Accessed on December 12, 2022 from `https://www.consort-statement.org/`.

3.2 TOCE's Journey into Reporting Standards (and a Flirtation with Evidence Standards)

Christopher D. Hundhausen (Oregon State University – Corvallis, US)

The ACM Transactions on Computing Education (TOCE) is a premier journal for computing education research. In 2019, the journal convened a task force to explore the possibility of adopting evidence standards for the journal. In 2021, ACM TOCE became the first journal in the field of computing to adopt reporting standards. In this talk, I chronicle TOCE's two year development of an evidence standards proposal; its discussion with the TOCE editorial board; and the TOCE editorial board's ultimate decision to recommend, but not require, the use of the APA JARS reporting standard for new submissions. Based on TOCE's experience, I identify barriers to the adoption of evidence standards in empirical computer science and consider possible ways forward.

3.2.1 Discussion

Some of the key lessons learned from introducing reporting standards at TOCE were:

- Include more people in the task force.
- Do not call it an evidence standard (most prefer reporting standard).
- Focus on how it makes the papers easier to review, read and search.

A lot of the pushback when introducing the CONSORT, and to some extent APA JARS, standards at TOCE was the fear that some might be excluded. Both of these standards have a heavy focus on quantitative research, and therefore left some concern that qualitative research may start to be rejected. It was noted that qualitative research is especially common in education, as compared to medical and psychology where these standards were established, and it may have been a closer fit to introduce them at a computer science conference or journal that has a heavier percentage of quantitative research, such as ACM's Conference on Human Factors in Computing Systems (CHI). Ultimately, the choice to introduce the standards at an education-focused journal was The fear of rejection is very real, but in reality, rejection already happens and the exclusion / inclusion criteria is not explicitly talked about. The counterpoint was made that by having a set of guidelines, there is more transparency in what gets excluded and it can be left to the community to adjust that line in the sand rather than leave it to the opinions of a few people behind closed doors.

3.3 Revolutionizing Peer Review with Empirical Standards

Paul Ralph (Dalhousie University – Halifax, CA)

Main reference Paul Ralph, Sebastian Baltes, Domenico Bianculli, Yvonne Dittrich, Michael Felderer, Robert Feldt, Antonio Filieri, Carlo Alberto Furia, Daniel Graziotin, Pinjia He, Rashina Hoda, Natalia Juristo, Barbara A. Kitchenham, Romain Robbes, Daniel Méndez, Jefferson Seide Molléri, Diomidis Spinellis, Miroslaw Staron, Klaas-Jan Stol, Damian A. Tamburri, Marco Torchiano, Christoph Treude, Burak Turhan, Sira Vegas: "ACM SIGSOFT Empirical Standards", CoRR, Vol. abs/2010.03525, 2020.
URL https://arxiv.org/abs/2010.03525

Scholarly peer review is the practice of subjecting a scholarly work (e.g. a manuscript) to the scrutiny of one or more experts (e.g. to decide whether to accept the manuscript for publication). Empirical research consistently demonstrates that scholarly peer review is ineffective, unreliable, and prejudiced. In principle, the solution is to move from contemporary, unstructured, essay-like reviewing to evaluating an artifact against an unambiguous, standard using a checklist. Therefore, a task force of over 50 leading scholars created natural-language models–called "empirical standards"–of the software engineering research community's expectations for different popular methodologies (e.g. case study, controlled experiment). These living documents, which should be continuously revised to reflect evolving consensus around research best practices, will improve research quality and make peer review more effective, reliable, transparent and fair. This talk will include a demonstration of reviewing tools developed based on the empirical standards.

3.3.1 Discussion

There are many problems with peer review as it exists today; many of which revolve around individuals making their own set of criteria or introducing biases to the process. The seemingly random nature of acceptance and rejection can impact on people's lives and even

sway how decisions are made in public policy. This can plausibly be be detrimental if a potentially impactful paper is rejected or less rigorous paper is accepted, but came to a fallacious conclusion.

The presented reviewing tools act as a sort of "checklist" for authors and reviewers to determine if the paper meets a set of guidelines. The attributes within the guidelines are broken down into essential and desirable attributes which can be established on the backend and it has the flexibility to change based on the type of methodologies being used. Much of the discussion was centered around the flexibility of the system and potential features and changes.

Overall, this was a fairly central talking point throughout the seminar, as the tool provides an open source option for introducing customizable guidelines in a straightforward manor that could be fit into a review process.

3.4 Registered Reports in Computer Science: Why Bother?

Neil Ernst (University of Victoria, CA)

Main reference Neil A. Ernst, Maria Teresa Baldassarre: "Registered Reports in Software Engineering", arXiv, 2023.
URL https://doi.org/10.48550/ARXIV.2302.03649

Registered reports are scientific publications which begin the publication process by first having a detailed research protocol, including key research questions, reviewed and approved by peers. Subsequent analysis and results are published with minimal additional review, even if there was no clear support for the underlying hypothesis, as long as the approved protocol is followed. Registered reports can prevent several questionable research practices and give early feedback on research designs. In software engineering research, registered reports were first introduced in the International Conference on Mining Software Repositories (MSR) in 2020. They are now established in three conferences and two pre-eminent journals. We explain the motivation for registered reports, outline the way they have been implemented in software engineering, and outline some ongoing challenges for addressing high quality software engineering research.

3.4.1 Discussion

Registered reports are a common part of the submission and review process in the medical and psychology fields, and has recently started to be established in a few computer science-related conferences and journals: MSR / EMSE, Transactions on Software Engineering and Methodology (TOSEM) and a special issue of Computer Science Education (CSE)). This method starts before any data has been collected and has the benefit of getting some feedback at an early stage of the research which can save time, money and be very helpful in getting insights and opinions outside of the authors' research groups. The other benefit is in reducing or eliminating under-powered or selectively reported findings. The current design involves two phases: Phase 1 where you declare what you would like to do and then Phase 2 where you report on the findings.

Some of the raised concerns with this process involved problems like a paper being published as Phase 1 and then being scooped before the results are published or a common sentiment is that only positivist philosophies matter (e.g., significance testing, falsification). This is not true in fields like medicine and is specifically in place so that results cannot just

be cherry picked. Similarly, without a requirement to declare methodological design and analysis ahead of time, there is a risk of focusing on novelty and significance over whether the research is sound. Computer science as a field is currently at risk of various questionable research practices and these are somtimes common at many venues.

In the presented case, the Phase 1 review process happened at MSR and then Phase 2 reviews were done in the EMSE system. Some pitfalls that came up during this were:

- Reviewers at MSR had the burden of accepting for a top journal (EMSE).
- Hard to manage reviewer discussion between multiple systems.
- Reviewer continuity between phases.

Something similar is being introduced at TOSEM, but both phases would be done under the same journal which would simplify many of the above issues. One of the major concerns is the additional burden this places on reviewers and whether or not there is the proper bandwidth to keep up with the increased capacity of reviews. This would effectively introduce two rounds of review per paper. On the other hand, much of what is in Phase 2 was already written and reviewed in Phase 1. For example, in the medical community, often times the pre-registration paper simply leaves the results section blank. This highlights the importance of reviewer continuity, as it would be quicker to review the Phase 2 submission without having to catch back up about all of the details that were already reviewed.

Another potential problem is with judgements on if a paper is important early in the research design phase. It becomes a balancing act where one side of the scale is to only let in "important" papers, which are subjective opinions and can lead to problems with bias or gatekeeping, and the other side is to allow in papers based on "soundness", which has potential to create a flood of "sound" but uninteresting, via some subjective criterion, research. For conferences, there are much harder limits on the number of submissions and page count, which makes this a problem somewhat unique to computer science's conference journal model as compared to some of the other fields that have implemented evidence standards.

3.5 Criteria and Scrutiny: Computing Education Research

Kate Sanders (Rhode Island College – Providence, US), Joseph Maguire (University of Glasgow, GB), Monica McGill (CSEdResearch.org – Peoria, US)

Main reference Marian Petre, Kate Sanders, Robert McCartney, Marzieh Ahmadzadeh, Cornelia Connolly, Sally Hamouda, Brian Harrington, Jérémie O. Lumbroso, Joseph Maguire, Lauri Malmi, Monica M. McGill, Jan Vahrenhold: "Mapping the Landscape of Peer Review in Computing Education Research", in Proc. of the Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE-WGR 2020, Trondheim, Norway, June 15-19, 2020, pp. 173–209, ACM, 2020.
URL https://doi.org/10.1145/3437800.3439207

In 2020, a working group was convened at the Innovation and Technology in Computer Science Education (ITiCSE) conference, led by Marian Petre, Kate Sanders and Robert McCartney on Mapping the Landscape of Peer Review in Computing Education Research (CER). The working group considered 17 venues, including CER conferences and journals, as well as overlapping conferences in Software Engineering and Human Factors. In this presentation, we consider some of the common review criteria observed across venues as well as some of the ethical concerns that emerged in peer-review and the process itself. In the

present talk, these elements are considered through the lens of excerpts and vignettes drawn from conference chairs and journal editors interviewed by the working group that reflect aspects of the conversations and debates that have happened during week at the seminar.

3.5.1 Discussion

After a retrospective on the criteria, ethical concerns and scrutiny involved in the review, much of the discussion was about continued topics like importance versus soundness, what types of comments should and should not be allowed and should we have a system for reviewing reviewers. The general sentiment is that more of the review process with evidence standards in place should be focused on the soundness, rather than off the cuff opinions on importance from reviewers with varying degrees of expertise and bias. If reviews were done in a fashion closer to Paul Ralph's work (Section 3.3), there would be more emphasis on determining if a paper contained the essential and desirable attributes for acceptance and less room for potentially ethically interactions between reviewers and authors. Lastly, it was discussed whether that reviewing tool could be enhanced to include a way to review reviewers, as that may provide some valuable data and create transparency in the quality of reviews.

4 Working Groups

4.1 APA JARS – Quantitative

In the first breakout session, participants formed small groups and reviewed the APA JARS guidelines for quantitative research. APA JARS provides a concise guide on what needs to be reported in each section. The overall goal is to improve the scientific rigor of peer-reviewed papers by providing requirements that support clear and transparent research. They can also work as a useful learning tool for novice researchers.

4.1.1 Discussion

After going through the APA JARS sections on quantitative guidelines in breakout groups, there was a larger discussion with the whole group to tie everything together.

First, it was discussed the types of audiences that would be impacted by the inclusion of reporting standards. These include, but are not limited to: authors, decision makers (e.g., reviewers and editors), novice researchers, readers, machine readers, lawmakers and the public. While there may be some initial turbulence in training those closely working on the writing and reviewing process to adhere to the guidelines, there are benefits to a wide audience by producing work that has a higher degree of clarity and transparency.

Overall, it was felt that there were pros and cons to this type of reporting standard. Some of the concerns were in needing to expand the standards further to account for the various types of research done in computer science and computer science education. APA JARS is centered around human studies, but standards in CS would need to be broadened to include things like graphics papers with timings or benchmarks. Others disagreed, as benchmarks are clearly quantitative, so differences with existing standards may not be large or valuable.

Furthermore, a major concern is that reporting standards will be used as a checklist to reject a paper. There may still be value in a paper even if it does not follow the standards perfectly. For example, in the medical community, they track the impact that CONSORT

has had overtime; analyzing what got better, rather than cutting off strict requirements. Introducing standards at a venue as optional with a plan in place to collect metrics on its adoption could be a good way to ease the community into using standards and down the road report on how and what it has changed.

The structured aspect of APA JARS was a hot topic of discussion. There were a handful of pros and cons mentioned about strict structure in the standards. Some participants felt that having strict requirements would hinder the style of the paper. APA JARS has guidelines on sections and what should be included in each, as well as requirements on a structured abstract which has very specific requirements on labeling and what is reported in it. The discussion on structured abstracts noted that it is very efficient and having the findings clearly labeled upfront is useful, but there was some concern that you miss out on telling a story that is helpful in attracting the reader. Some journals have gone the route of having both structured and general public abstracts so that detailed and easy to read options are available, which can help to broaden the audience that can consume the information. Some of the other benefits of having requirements on structure include the ease of finding information between different papers, as it is always reported in a specific section. Additionally, requirements on structure can be helpful in mining information across a larger set of papers.

4.1.2 Conclusions

Overall, the consensus seemed to be that reporting standards could be a beneficial addition to the field, although they may need to be adapted to better align with each community. The main takeaway was the importance of how it is introduced to the community. This will be vital in not ostracizing anyone while still uplifting the quality of reporting for the community as a whole.

4.2 APA JARS – Qualitative

Next, participants worked with their breakout groups to go over the qualitative and mixed-methods sections of the APA JARS guidelines before having a larger discussion with the whole group.

4.2.1 Discussion

The first note was the difference in the qualitative guidelines providing separate sections for guidance to reviewers and authors. Many interpreted this as sort of a "defensive mechanism" to instruct reviewers to not just reject a paper. There seems to be more fear of not meeting the standards, and when compared to the quantitative guidelines, the depth of detail was much shallower. It could be that the qualitative side is still being built out, which is plausible given the qualitative standards are new, and in the interim it was easier to ask reviewers to take more consideration. Another note was that there may be a problem with cross-paradigm reviewing where reviewers with a background in quantitative research were making judgements about qualitative submissions without fully understanding it.

While the quantitative side was broken down into many different methodology guidelines, the qualitative standards were condensed to just one. Qualitative research is very common in computer science, especially in computer science education, and more care will be necessary

when developing the standards than what is in APA JARS. In particular, there are a wide variety of approaches to inquiry, some include:

- Positivist: Makes hypotheses and gathers evidence to support or refute them. Restricts yourself to empirical/observable data. No cognitive processes because you cannot observe them. Rely on behavioral data.
- Post-positivist: Accepts that reality can only be known imperfectly. There's observed bias.
- Feminist: Capturing perspectives of marginalized peoples. Acknowledges complexity of social life.
- Intersectionality: People's experiences are different with combinatorial composition of their demographic attributes.
- Postmodern: Personal perspectives over truth.
- Constructivist: People construct their own reality through inquiry.
- Critical: Recognizing that the default system in society is biased toward able-bodied, white males. The system must be dismantled before we can be equal.

Each of these have foundations, some participants claimed, are different and the quality criteria for judging them differ significantly. From this perspective, the standards will need to be broken down into a set of guidelines that more closely matches each, and there needs to be more guidelines on how a reviewer should approach criticism for these papers. For example, they will need to have the expertise to recognize the philosophy being used and not make judgements on the paper based on a different philosophy. Others, however, disagreed. Just because a paper meets a philosophical, and subjective category, has no bearing on whether that paper should have an introduction or state research questions. Many commonalities can and do exist, despite paradigmatic claims.

If many templates are necessary to accommodate such a variety, there is a problem with who controls the standards. It could be an issue that those in charge could be seen as "gatekeeping" if they do not have standards for a particular philosophy. There needs to be a mechanism for introducing additional templates and allowing the community to give feedback and adjust the guidelines. In contrast, if paradigms exist at different levels of rigor, free discussion needs to be had for whether that rigor is sufficient for publication. For example, if a paper simply claimed to be part of a paradigm that did not require evidence at all, most scientists would find this lacking. How to manage this natural balance and tension was not clear.

4.2.2 Conclusions

Overall, the APA JARS standards for qualitative research were useful in providing a gap analysis of what it was missing compared to what the computer science subfields would need. They were a bit more lenient than the quantitative standards, but that may have been due to being relatively new and still under construction. A lot of good points were made about how these standards would need to, arguably, be broken down and maintained if they were to be adopted at computer science venues.

4.3 What Works Clearinghouse

The next breakout session consisted of forming small groups, similar to the previous sessions, and going through various sections of the What Works Clearinghouse (WWC) standards. These were established by U.S. Department of Education as a way to identify studies that

meet specific thresholds of evidence. This helps educators, policymakers, researchers and the public to understand the effectiveness of education programs and interventions, and ultimately plays a role in determining grant funding.

4.3.1 Discussion

When comparing APA JARS to WWC, it becomes clear that they are meant for different purposes. One is reporting standards for publications in general and the other is related to public policy and grants. The degree of detail and breakdown of tiers of evidence is not a part of computer science culture.

Most of the discussion centered around how would we introduce a set of standards like these into the field. Some of the suggestions included forming a working group or workshop just to give people a place to try it out and get feedback. Introducing standards as required, but not a cause of rejection during a startup period is another option. This would give time for authors to get feedback on their papers to understand what they didn't meet on the requirements and improve, as well as giving reviewers time to get used to the guidelines before having an impact on acceptance. Some discussed the route of making the standards optional, which is the direction that TOCE went. Both have their benefits, completely optional at the start is easier to get the community on board with, and then the expectation is that as more people choose to go in that direction, if people like that style of paper they will adopt it into their practices as well. The requirement with no impact at the start is probably going to be harder to get approval from the community and risks pushing people away, but it also forces authors and reviewers to become familiar with the standards while not rejecting them initially.

4.4 Take a Paper and Convert It

In this session, seminar attendees split up into breakout groups and were tasked with selecting one of their already published works and going through the relevant APA JARS guidelines. While doing this, they were tasked with making several classifications:

1. Information that was required in the guidelines was present and in the required section.
2. Information that was required in the guidelines was present but in a different section.
3. Information that was required in the guidelines was not present but was recorded and could have been included.
4. Information that was required in the guidelines was not present and was not recorded.

4.4.1 Discussion

This activity gave a more hands on interaction beyond reading through the guidelines. By putting the requirements in the APA JARS guidelines through the lens of existing works that had been published by the attendees, it became a bit more apparent which parts were essential and which parts might not be necessary for inclusion. This is similar to how the guidelines for APA JARS were formed by first looking at CONSORT as a model and fitting it to the needs of the psychology field. Some of the main takeaways were:

- It simplifies the process of having to remember what needs to be included because there's a list to remind you.
- It would not be too difficult to have written the papers with these standards if they were asked to.

- The standards would work well for teaching novice researchers about writing for publication.
- Not everything listed in APA JARS seems necessary to be in the actual narrative of the paper.
- The requirements on section content and section headers may be met with argument of reducing freedom of expression.

The paper structure requirements were heavily discussed. The two competing viewpoints were largely that the structure will take away academic freedom, which was also a common argument historically in other fields. The other is the benefit of consistency between papers. While the heavy structural requirements of a standard like CONSORT may seem a bit jarring (e.g., some subsections breakdown to headers with one sentence), it can have the benefit of increasing the speed and comprehensibility when reading through many papers if the relevant content is exactly where you expect it to be. Further, standardized sections ease automated analysis. Using this activity as an example, many found it that it was time consuming as a reviewer to take the checklist of requirements and sift through the paper looking for each piece, whereas with a structured requirement it would have been simple to run down the list.

4.5 Breakout: Mixed-Methods

While focus has been placed on quantitative and qualitative studies, there's the additional problem of how to handle mixed-methods studies. There are several types of mixed-methods studies that were taken into consideration. For example, there are mixed-methods studies that collect quantitative data in an A/B experiment followed by a survey and interview to gather qualitative data about a participant's experience. Also, there are triangulation studies which are common in the formative design of tools which follow a waterfall approach of interviews, surveys, and validation with user studies.

The main concerns were about the amount of page bloat that would come from reporting on every detail in a series of experiments in just one paper and the complexity of the guidelines becoming too much of a hurdle or barrier to entry. The group felt that a structured appendix or supplemental material might be a good location for the required information while keeping the page length short and still allowing the author to have some freedom with the narrative style of the paper. It was suggested to build upon this with a fork of Paul Ralph's work (Section 3.3); this would allow for the flexibility to create minimum viable templates based on the community's unique needs. That system could then be used by the author to input the information that was not included in the paper itself and then the structured appendix could be auto-generated. While this would create another set of documents that reviewers would need to go through, they would already need to verify that the information is included and it would all be in one place.

4.6 Breakout: Measuring Effects

The topic of this breakout was on how to measure the effect of implementing evidence standards. If a venue were to make use of a checklist like from Paul Ralph's talk (Section 3.3), even if the standards were introduced as an optional recommendation, we could measure the percentage of papers that passed that test of rigor and compare that year to year to see if the optional recommendations were sufficient to increase adoption. Another suggestion was to

try to get a measurement of "happiness factor" or sense of belonging in the community from authors and reviewers to see how that changes over time. If part of the reason for enforcing standards is to improve people's trust in the system, then getting a measure before and after could be beneficial. This could range from satisfaction with the submissions process, their experience with the guidelines or whether they felt the reviews they received were fair. Another metric for reviews could be to have conferences release review data and report on inter-rater reliability. This would provide transparency to the community and there could be a bit more retrospective year to year on what changes were made at the conference and the impact it may have had.

4.7 Breakout: Community Engagement

Getting the community on board is a key piece of successfully implementing evidence standards. Compensation and motivation were a big factor. Providing cash incentives for training reviewers and paying them for the work they put in. One option for training is Designing Empirical Education Research Studies (DEERS) [1]. For authors, it could help motivate them to adhere to optional standards if a badge system was put in place to mark papers that follow the standards. And for venues, one path to motivating them to change could be to focus efforts at highest-quality venues first and show whether it has an impact on things like quality, clarity or transparency of work, and if the community likes the changes. Smaller venues may pick up on that and be motivated to change as well.

Reducing opposition is important so that members of the community do not feel like they no longer belong. It is important to be transparent in the messaging and get information about the changes out in various ways, such as running panels at venues and writing op-eds. Beyond spreading information, allowing the community to be able to contribute to and have some control over the standards is essential in making sure they work well for the types of papers at their venue. Deadlines for adoption should be far enough in the future and are stair-stepped to allow for adaptation, so that year to year, the standards can be contributed to and changed to be a proper fit.

References

1 Carver, Jeffrey C. and Heckman, Sarah and Sherriff, Mark. Designing Empirical Education Research Studies (DEERS). Accessed on December 12, 2022 from `http://empiricalcsed.org/`.

4.8 Breakout: Operationalizing

There needs to be a focus on implementing things like peer review into technology. Ideally, this could be done in an existing technology, like OpenReview[1], as long as it is open sourced. Unfortunately, many of the existing systems that are commonly used, like EasyChair[2], are closed and do not support the types of things from Paul Ralph's talk (Section 3.3). An open source system would allow for easier adoption between different communities and provide the ability to include data gathering elements such as the previously discussed "happiness factors" or inter-rater reliability measures for reviews (Section 4.6).

References

1 OpenReview. Accessed on December 12, 2022 from `https://github.com/orgs/openreview/repositories`.
2 EasyChair. Accessed on December 12, 2022 from `https://easychair.org/`.

4.9 Next steps: What Form of Actions Will We Take?

The expectation is that evidence standards will not be something that can be put into place across the entire field quickly, nor should that be the goal. Instead, one goal is to make some movement toward positive changes that may help with acceptance incrementally. Toward this goal, the group discussed a variety of conferences in computer science, software engineering and computer science education where some of these changes could be brought up to program chairs, steering committees or town hall meetings. Following a similar path to what was done at TOCE, most recommendations will involve offering optional guidelines as a first step to get authors', reviewers' and editors' a chance to trial the changes and adjust the requirements. Additionally, a dissemination plan was discussed to try to give some information on evidence standards to a broader audience.

4.9.1 Conferences and Journals

- Koli: The information discussed at the seminar will be brought up at the next PC meeting. Koli has two tracks, systems and tools, and the tools track might be a good candidate for introducing a rubric using a fork of Paul Ralph's work. This could help to build up this style of review in the community and get some feedback.
- ETRA: Similarly, this information will be brought up at a town hall next year to see what the community thinks. This could be a good way to introduce it here first, and then use it as an example when bringing it up at other CHI sponsored conferences that might be interested in this style of standards.
- TOSEM: This will be presented at a board meeting with the intention of proposing a special topic on Human Factors in Software Engineering which could be used to test the reporting standards.
- ICER: It will be proposed to adapt the structure of the reviewing form to use Paul Ralph's work and have a subset of papers go through this process.
- CSE: Currently has structured abstracts and recently added a track for registered reports. As a next step, there will be a suggestion added to the instructions for authors to look at JARS and encourage them to consider following it where appropriate.

4.9.2 Dissemination

- Position Papers: All of the participants at the seminar were interested in contributing on a position paper that would be submitted to Communications of the ACM (CACM).
- Panels: Plans were discussed to set up a CHI panel on this subject. Initial ideas include presenting something similar to the activity from Section 4.4, which converted an already published paper into APA JARS format, and then let the audience look through the differences then discuss concerns with the panel.

Participants

- Brett A. Becker
University College Dublin, IE
- Andrew Begel
Carnegie Mellon University – Pittsburgh, US
- Michelle Craig
University of Toronto, CA
- Andrew Duchowski
Clemson University, US
- Neil Ernst
University of Victoria, CA
- Arto Hellas
Helsinki University of Technology, FI
- Christopher D. Hundhausen
Oregon State University – Corvallis, US
- Ciera Jaspan
Google – Mountain View, US
- Timothy Kluthe
University of Nevada – Las Vegas, US
- Juho Leinonen
Aalto University, FI
- Joseph Maguire
University of Glasgow, GB
- Monica McGill
CSEdResearch.org – Peoria, US
- Brad Myers
Carnegie Mellon University – Pittsburgh, US
- Andrew Petersen
University of Toronto Mississauga, CA
- Mauro Pezzè
University of Lugano, CH
- Paul Ralph
Dalhousie University – Halifax, CA
- Kate Sanders
Rhode Island College – Providence, US
- Andreas Stefik
University of Nevada – Las Vegas, US
- Claudia Szabo
University of Adelaide, AU
- Jan Vahrenhold
Universität Münster, DE
- Titus Winters
Google – New York, US
- Aman Yadav
Michigan State Universit – East Lansing, US