Extracting Hypothesis Formalizations to Support Reproducible Research

Madeleine Grunde-McLaughlin[1], Weixuan Liu[1], Ria Patil[1], Nino Migineishvili[1], Ranjay Krishna[1], Daniel S. Weld[1] and Jeffrey Heer[1]

[1]*University of Washington, 185 E Stevens Way NE, Seattle, WA 98195, United States*

Abstract

Ambiguity in how a hypothesis is formalized and reported leaves degrees of freedom that contribute to questionable research practices and the reproducibility crisis. Drawing from work in psychology that defines these degrees of freedom, we design an LLM-backed system to support researchers in interpreting experimental results of HCI papers. The system inputs scientific text to judge if and how its hypotheses and variables are specified and operationalized, and how that formalization impacts the generalizability of research. With such a system, we aim to support researchers in understanding how concepts have been operationalized in the past and in judging the specification quality of their own or others' work. In this position paper, we propose the motivation and design of such a system. We then express our intentions for the workshop.

Keywords

reproducibility, preregistration, hypothesis formalization, large language models

1. Introduction and background

Many research disciplines, including HCI, use statistics to understand, interpret, and present findings. The takeaways that statistics support can then be read and built upon by other researchers for the collective pursuit of knowledge. To be effective at generalizing to future work, findings should be robust enough that reproducing the experiment finds similar results, a concept known as mechanistic reproducibility [1].

However, across many fields there has been a realization of a "reproducibility crisis," in which findings cannot be reproduced [2, 3, 4, 5]. This crisis has many sources, including insufficient reporting and questionable research practices that report significant results within the existing data, but the results of which cannot be assumed to generalize to new data [6, 7, 8]. These include practices like p-hacking, which (intentionally or not) attempts different analyses until a significant difference is found [9], and HARKing (Hypotheses After Results are Known) in which researchers review the data's findings before determining their proposed hypothesis [10].

Improving reporting and research practices is a challenging problem. Statistics can provide useful, but different, findings when conducted in exploratory analyses, confirmatory analyses, and analyses that have both exploratory and confirmatory elements. Furthermore, experimental setups are a complex, multi-faceted endeavor in which underreporting of methodology and overclaiming of findings are common [6, 11]. How can we support HCI researchers in interpreting experimental design and its impact on statistical findings?

In this paper, we consider what HCI can draw upon from research on study design and reporting from psychology. Researchers in the field of psychology have built a growing body of work to guide study design and reporting for improving statistical robustness in response to the reproducibility crisis [6, 12, 11]. These include concepts that apply to statistical methods already used in HCI research, such as Null Hypothesis Significance Testing (NHST), but only a small subset of HCI researchers follow these practices [13]. For instance, consider the practice of preregistration, in which authors declare their study design on a public forum before executing the study. Preregistration can be a powerful tool for reducing the chance of p-hacking and HARKing [14, 15]. Preregistration is common in psychology,

Meta-HCI '25: First Workshop on Meta-Research in HCI, April 26, 2025, Yokohama, Japan

with some journals incentivizing their inclusion to publish [16]. Despite recent calls to incorporate pre-registration into HCI research, the rate of preregistration in HCI remains relatively low [14, 17, 18]. As of 2018-2021, only 1.11% of CHI papers referenced a preregistration [17].

In this position paper, we propose a system to scaffold interpretation of the results of both confirmatory and exploratory research in HCI using a related body of work in psychology on defining Researcher Degrees of Freedom (RDoFs) [6, 19]. RDoFs reflect a researcher's ability to choose among multiple options during analysis, such as testing out different confounders. More flexibility at the time of study execution and analysis increases the ability for questionable research practices to occur, reducing the robustness of statistical outcomes. However, pre-specifying choices along all RDoFs into a medium like a pre-registration is often not sufficiently completed [17, 12, 11, 20, 21, 19]. Furthermore, some RDoFs should always appropriately specified (e.g., "operationalizing non-manipulated variables in different ways"), while others can be left flexible for exploratory research (e.g., "choosing to include different measured variables as covariates, independent variables, mediators, or moderators"). Prior work in HCI has used these RDoF guidelines to investigate HCI research, finding that researchers have used these RDoFs opportunistically [22].

We focus on supporting people in reading existing scientific studies to understand the content and limitations of their hypothesis and variable operationalizations. We begin with three user scenarios about researchers doing a literature review, a paper review, and a paper submission. We will then propose a system to address those users' needs by processing a paper and judging it along the RDoFs. Finally, we will discuss our intentions for the Meta-HCI workshop.

2. User scenario

We introduce three imagined researchers who need to better understand the statistical robustness of the results described in the text.

User 1. PhD student performing a literature review

As a... As a PhD student in Human Computer Interaction, I am completing a literature review on trust in AI to inform my study. My study investigates the effect of the trustworthiness of the way of presenting an AI's outputs on the change in user beliefs.

I want to... I have collected literature on previous studies measuring conditions that improve trust in AI, but I want to know if I can generalize those findings to my study design. I first want to be able to review the ways that these current papers operationalize trust in AI and the confounders they use. For instance, one study claims that people who are given text explanations of AI behavior are significantly more likely to overrely on the AI's output, which is sometimes used as a proxy for trust. I additionally want to understand if this research is sufficiently confirmatory to generalize to my case. Making these judgements is a labor-intensive process, as this information is given throughout long papers. Additionally, although I have taken statistics classes several years ago, I do not have expertise in judging if the authors are overclaiming results.

so that... By better understanding the robustness of research findings, I have more confidence that my study design will isolate the underlying effect that I care about, and I am less likely to create a study with a faulty assumption. I will also find knowledge gaps where my study fits into existing literature, brainstorm how my analyses might differ from the related work, and consider new confounders.

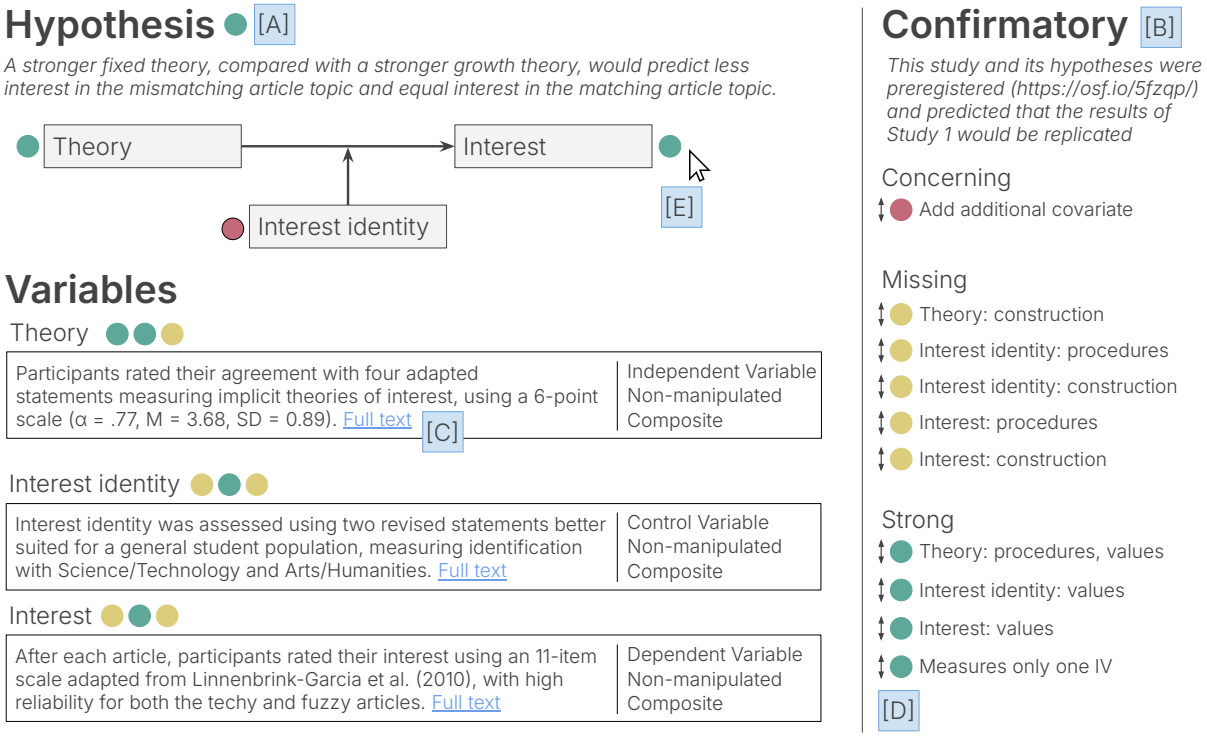

Figure 1: We propose a system to help researchers understand the robustness and generalizability of statistical results in scientific papers. The system inputs a scientific paper and synthesizes a report about the hypothesis formalization and whether it is appropriately specified. The system provides [A] flags judging the quality of RDOF specifications, [B] the categorization of the study as exploratory or confirmatory, [C] operationalization summaries with reference to the original text, [D] interactive capabilities for the users to make corrections if the classifications are wrong, and [E] interactivity of flags to reveal an explanation of the categorization. This figure references a psychology paper [23] as annotated by a prior work in psychology [21]

User 2. Conference reviewer

As a... As a reviewer for CHI, I am reviewing a paper that runs a user study and uses NHST, finding statistically significant results to the claim that "artificially generated quizzes improve retention over expert-generated quizzes." The discussion generalizes to conclude that if teachers want to optimize for information retention, artificially generated quizzes are superior. Although I am an HCI researcher, this is not my exact area of expertise.

I want to... I want to understand if the submission is missing information from the methodology to make it reproducible and if the strength of the claims matches the methodology.

so that... I can better educate myself and provide more helpful and targeted feedback.

User 3. Researcher writing a paper submission

I am... I am an HCI researcher who has completed a research study testing the effects of my system's design on comprehension. I have been through months of study conceptualization, design, and execution. Our methodology included testing different model formulations based on the data we collected to find the best fit, such as whether the user's Need for Cognition should be a confounder. I have compiled my notes and procedures into a draft I am planning to submit to a conference. I have written my methodology with the final model we used, but I do not know how many of the changes throughout the study development and analyses should be reported.

I want to... I want to know what is unclear, misspecified, and missing in my current reporting.

so that... I can rework my paper draft prior to submission to improve its reproducibility, ensure I do not overclaim, and contribute quality scientific practices.

3. System proposal

We propose a system to support researchers in interpreting scientific papers through the lens of hypothesis formalization and its effect on the generalizability of findings. In this section, we give a system overview, discuss our design goals and their manifestation in the system design, and demonstrate how our imagined users could benefit from the system.

3.1. System overview

This system takes as input a paper or preregistration text and a hypothesis from that text, then outputs a report of the hypothesis formalization as seen in Figure 1. The system uses a subset of guidelines from psychology that relate to RDoFs in the hypothesis specification and variable operationalization parts of experiment development [6]. Figure 1 displays an output from a dataset we plan to use to evaluate our system. This dataset's annotations were written by researchers in psychology on psychology preregistrations and papers [21].

3.2. Design goals

[D1] Errors are easily noticed. As the goal of the system is to communicate the study's specification quality, this quality judgement should be easily reviewed at all times. The system uses flags, separated by color, to focus visual attention on the RDoFs. The flags exist spatially by the elements of the hypothesis or variable descriptions towards which they are relevant (Figure 1 [A]). The system also provides a summarization and categorization of the errors on a sidebar for easy access. The categorizations of the flags update depending on if the investigation is presented as confirmatory or exploratory (Figure 1 [B]). When hovering over flags on the left, the associated description on the right is highlighted.

[D2] The user's interpretation takes precedence. The system should provide guidance for interpreting this complicated set of specification requirements while still giving the user control to verify and interpret the content within their own contexts. The system includes references to verbatim text from the paper as well as definitions of the degrees of freedom for the user to review (Figure 1 [C]). The system also has interactive capabilities for users to make corrections if it is wrong (Figure 1 [D]).

[D3] Choices between different levels of detail. The report begins as a summarization, but allows users to explore different levels of detail for further understanding as they desire. The variable information is presented as a summary, with the option to explore full text (Figure 1 [C]). Additionally, the flags can be clicked in order to get an explanation for the categorization, the related degree of freedom, and why it is concerning (Figure 1 [E]).

3.3. Support for user scenarios

This system provides support for each of the anticipated users.

User 1. With our system, User 1 inputs papers and their relevant hypotheses into the system. Getting reports for multiple papers enables comparison of the papers' relative strength of confirmatory analysis to filter which papers serve User 1's purposes. The associated explanations of RDoFs as well as the contextualization of the flags as within a confirmatory or an explanatory study, provide User 1 with context for deciding if the paper's findings are relevant for their study (Figure 1 [B]). Then, User 1 uses the operationalization summaries of variables for trust in AI to see fine-grained differences in operationalization (Figure 1 [C]).

User 2. The system flags what details are missing for reproducibility, which the reviewer can filter by what they consider reasonable to expect to include within the standards of their field (Figure 1 [A]). User 2 goes through each of the flags marked as concerning individually, matching them with the verbatim text from the paper. By employing the drill-down capabilities of the system to ground themselves into the extracted text from the paper, they can review each flag marked as concerning

(Figure 1 [C,E]). The system will categorize the flags based on whether the research is confirmatory or exploratory, supporting User 2 in deciding and justifying if the paper is overclaiming (Figure 1 [B]). The reviewer can interact with the different flags and categorizations such that the report visually reflects their subjective understanding of the importance of any omissions (Figure 1 [D]).

User 3. User 3 uses the system similarly to a reviewer. For each flag marked as concerning, they can determine if it is related to underreporting their own methodology, which can be improved by more detailed descriptions, or if it is an inherent limitation of the study, which can be addressed by ensuring an appropriate strength of the claim (Figure 1 [A,B]).

4. Intention for the Meta-HCI Workshop

The Meta-HCI workshop provides a valuable venue to further develop this proposal. In regard to this proposal, we welcome feedback on whether there would be a desire for this type of scaffolding in the HCI community, if the user scenarios are convincing, and what other capabilities would be helpful. At the workshop, we would be especially interested in discussing topics such as: the parallels and differences in scientific methodology between HCI and other fields such as psychology, opportunities for meta-analysis of scientific analysis practice in HCI, to what degree reproducibility is wanted or needed in HCI, and the role of RDoFs, confirmatory analysis, and Human-in-the-Loop formats in current pushes for involving AI into scientific practice.

Declaration on Generative AI

The authors employed the free version of Grammarly for spell checking and grammar and reviewed each suggested change manually.

References

[1] R. G. Hudson, Should we strive to make science bias-free? a philosophical assessment of the reproducibility crisis, Journal for General Philosophy of Science 52 (2021) 389 – 405. URL: https://api.semanticscholar.org/CorpusID:235505853.

[2] J. P. Ioannidis, Why most published research findings are false, PLoS medicine 2 (2005) e124.

[3] M. Baker, Reproducibility crisis, nature 533 (2016) 353–66.

[4] O. S. Collaboration, Estimating the reproducibility of psychological science, Science 349 (2015) aac4716.

[5] C. F. Camerer, A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al., Evaluating replicability of laboratory experiments in economics, Science 351 (2016) 1433–1436.

[6] J. M. Wicherts, C. L. S. Veldkamp, H. E. M. Augusteijn, M. Bakker, R. C. M. van Aert, M. A. L. M. van Assen, Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking, Frontiers in Psychology 7 (2016). URL: https://api.semanticscholar.org/CorpusID:12942653.

[7] L. K. John, G. Loewenstein, D. Prelec, Measuring the prevalence of questionable research practices with incentives for truth telling, Psychological Science 23 (2012) 524 – 532. URL: https://api.semanticscholar.org/CorpusID:8400625.

[8] A. Gelman, E. Loken, The garden of forking paths : Why multiple comparisons can be a problem , even when there is no " fishing expedition " or " p-hacking " and the research hypothesis was posited ahead of time *, 2019. URL: https://api.semanticscholar.org/CorpusID:198164638.

[9] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, M. D. Jennions, The extent and consequences of p-hacking in science, PLoS Biology 13 (2015). URL: https://api.semanticscholar.org/CorpusID:17475145.

[10] N. L. Kerr, Harking: Hypothesizing after the results are known, Personality and Social Psychology Review 2 (1998) 196 – 217. URL: https://api.semanticscholar.org/CorpusID:22724226.

[11] R. M. Heirene, D. A. LaPlante, E. R. Louderback, B. Keen, M. Bakker, A. Serafimovska, S. M. Gainsbury, Preregistration specificity and adherence: A review of preregistered gambling studies and cross-disciplinary comparison, Meta-Psychology (2024). URL: https://api.semanticscholar.org/CorpusID:270937245.

[12] A. Claesen, S. Gomes, F. Tuerlinckx, W. Vanpaemel, Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies, Royal Society Open Science 8 (2021). URL: https://api.semanticscholar.org/CorpusID:239890029.

[13] P. Dragicevic, Fair statistical communication in hci, 2016. URL: https://api.semanticscholar.org/CorpusID:64470036.

[14] A. Cockburn, C. Gutwin, A. Dix, Hark no more: On the preregistration of chi experiments, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–12. URL: https://doi.org/10.1145/3173574.3173715. doi:10.1145/3173574.3173715.

[15] R. M. Kaplan, V. L. Irvin, Likelihood of null effects of large nhlbi clinical trials has increased over time, PloS one 10 (2015) e0132382.

[16] C. D. Chambers, Registered reports: A new publishing initiative at cortex, Cortex 49 (2013) 609–610. URL: https://www.sciencedirect.com/science/article/pii/S0010945212003735. doi:https://doi.org/10.1016/j.cortex.2012.12.016.

[17] Y. Pang, K. Reinecke, R. Just, Apéritif: Scaffolding preregistrations to automatically generate analysis code and methods descriptions, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3491102.3517707. doi:10.1145/3491102.3517707.

[18] X. Pu, L. Zhu, M. Kay, F. Conrad, Designing for preregistration: A user-centered perspective, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–6.

[19] M. Bakker, C. L. S. Veldkamp, M. A. L. M. van Assen, E. A. V. Crompvoets, H. H. Ong, B. A. Nosek, C. K. Soderberg, D. T. Mellor, J. M. Wicherts, Ensuring the quality and specificity of preregistrations, PLoS Biology 18 (2018). URL: https://api.semanticscholar.org/CorpusID:228087995.

[20] G. Ofosu, D. N. Posner, Pre-analysis plans: An early stocktaking, Perspectives on Politics 21 (2021) 174 – 190. URL: https://api.semanticscholar.org/CorpusID:233604973.

[21] O. R. van den Akker, M. Bakker, M. A. L. M. van Assen, C. R. Pennington, L. Verweij, M. M. Elsherif, A. Claesen, S. D. M. Gaillard, S. K. Yeung, J.-L. Frankenberger, K. Krautter, J. P. Cockcroft, K. S. Kreuer, T. R. Evans, F. M. Heppel, S. F. Schoch, M. Korbmacher, Y. Yamada, N. Albayrak-Aydemir, S. Alzahawi, A. Sarafoglou, M. M. Sitnikov, F. Děchtěrenko, S. Wingen, S. Grinschgl, H. Hartmann, S. L. K. Stewart, C. M. F. de Oliveira, S. Ashcroft-Jones, B. J. Baker, J. M. Wicherts, The potential of preregistration in psychology: Assessing preregistration producibility and preregistration-study consistency., Psychological methods (2024). URL: https://api.semanticscholar.org/CorpusID:273293378.

[22] Y. Liu, T. Althoff, J. Heer, Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–14. URL: https://doi.org/10.1145/3313831.3376533. doi:10.1145/3313831.3376533.

[23] P. A. O'Keefe, C. S. Dweck, G. M. Walton, Implicit theories of interest: Finding your passion or developing it?, Psychological Science 29 (2018) 1653–1664. doi:10.1177/0956797618780643.