Apéritif: Scaffolding Preregistrations to Automatically Generate Analysis Code and Methods Descriptions

Yuren Pang
Paul G. Allen School of Computer Science, University of Washington
Seattle, Washington, USA

Katharina Reinecke
Paul G. Allen School of Computer Science, University of Washington
Seattle, Washington, USA

René Just
Paul G. Allen School of Computer Science, University of Washington
Seattle, Washington, USA

ABSTRACT

The HCI community has been advocating preregistration as a practice to improve the credibility of scientific research. However, it remains unclear how HCI researchers preregister studies and what preregistration users perceive as benefits and challenges. By systematically reviewing the past four CHI proceedings and surveying 11 researchers, we found that only 1.11% of papers presented preregistered studies, though both authors and reviewers of preregistered studies perceive it as beneficial. Our formative studies revealed key challenges ranging from a lack of detail about the study design, hindering comprehensibility, to inconsistencies between preregistrations and published papers. To explore ways for addressing these issues, we developed Apéritif, a research prototype that scaffolds the preregistration process and automatically generates analysis code and a methods description. In an evaluation with 17 HCI researchers, we found that Apéritif reduces the effort of preregistering a study, facilitates researchers' workflows, and promotes consistency between research artifacts.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Preregistration, Experiment design, Reproducibility, Data analysis

ACM Reference Format:
Yuren Pang, Katharina Reinecke, and René Just. 2022. Apéritif: Scaffolding Preregistrations to Automatically Generate Analysis Code and Methods Descriptions. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3491102.3517707

1 INTRODUCTION

Preregistration is the practice of documenting a study's objective and analysis plan prior to conducting it and observing its outcomes [62]. When preregistering a study, researchers define their research questions, hypotheses, sampling plans, and analysis plans. To do so, public preregistration repositories, such as the Open Science Framework (OSF) [29] and AsPredicted [48], are available across disciplines. Previous work has shown that preregistration boosts the credibility of quantitative [61], qualitative [34, 49], and exploratory research [21]. A growing number of research communities advocate preregistration, including in medicine for clinical trials [18, 37], psychology [61], economics [66], biology [31], political science [59], and human-computer interaction [14].

While efforts encouraging preregistration are also underway in the HCI community [15, 43, 44, 70], adoption is still in its early stages [14, 15]. It remains unclear whether adoption has increased with growing awareness of preregistration and what challenges to preregistration authors and reviewers perceive. To address this, we conducted a systematic review of CHI proceedings between 2018–2021, finding 32 out of 2,874 papers (1.11%) that included one or more links to a total of 47 preregistrations. The majority of these (91.49%, 43/47) were quantitative studies, of which 74.42% (32/47) were reporting on Null-Hypothesis Statistical Testing (NHST). Given that the vast majority of CHI publications report on at least one study [52], we conclude that preregistration in the HCI research community remains a rare practice. A content analysis of the preregistrations and their corresponding papers revealed two main challenges for the reproducibility of preregistered studies: (1) varying levels of detail and (2) inconsistencies between a preregistration and its corresponding paper. In a survey of 11 researchers, participants pointed out the benefits of preregistration but also expressed a desire for *scaffolding* of preregistration questions, integration of preregistration into the research workflow, and for supporting consistency between preregistrations and papers.

In the learning sciences, scaffolding refers to supporting learners with concrete structure and guidance to ensure successful completion of a task [77, 93]. The term has been adopted in the HCI community to refer to step-by-step instructions, such as providing non-experts with a structured workflow required for designing rigorous scientific experiments [68]. In line with these definitions and in the context of this paper, we refer to providing structure, interactions, and automation for completing a preregistration form as scaffolding the preregistration process.

Guided by our formative studies and existing design requirements for preregistration, we developed a research prototype called Apéritif to explore how the preregistration process could be improved. Apéritif is a Chrome extension for the popular preregistration platform AsPredicted. The system scaffolds the preregistration process of quantitative studies with finer-grained questions and suggestions, offers a high degree of automation for generating analysis code (in Python and R), a methods description for a future publication, and enables version control to allow changes throughout the research process.

Our evaluation of Apéritif with 17 HCI researchers showed that, when compared to using the conventional AsPredicted template,

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517707

Apéritif reduces the time to preregister a study, enhances the preregistration user experience, and increases the likelihood of continued preregistration usage for future research.

In summary, this paper contributes:

(1) An empirical study showing that the number of preregistrations in HCI remains low and that level of detail and consistency with the corresponding paper vary substantially between existing preregistrations (Section 3).
(2) A set of design requirements for preregistrations, compiled from past literature and the findings of our formative study. These include best practices for preregistration templates, integration into the research workflow, and consistency across research artifacts (Section 3.3).
(3) Apéritif, a research prototype for scaffolding the preregistration process and automating the generation of analysis scripts and methods descriptions (Section 4).
(4) An empirical evaluation, based on Apéritif, showing that scaffolding and automation improves the preregistration experience (Section 5).

2 RELATED WORK

Our work is motivated by prior work showing the benefits of preregistration and recent tools that help with the design and statistical analysis of experiments.

2.1 Preregistration Platforms and Practices

Current preregistration platforms, including domain-specific and domain-general registries, support researchers in various disciplines [61]. *Domain-specific* registries exist for clinical trials [64], economics [6], and political science [36]. Furthermore, over 1,100 journals and conferences have implemented the Transparency and Openness Promotion (TOP) guideline, a standard addressing open science practices, which include study and analysis preregistration [61]. In contrast to domain-specific registries, AsPredicted [65] offers a *domain-general* preregistration template, where researchers answer nine questions about their research designs and analyses. The platform subsequently generates a short time-stamped preregistration PDF. The Open Science Foundation (OSF) [28] offers another domain-general registry, where authors preregister in fields such as social and behavioral science, education, business, and life science. OSF contains nine preregistration templates, including a standard template, a more flexible open-ended template, a qualitative template, a secondary data template, a registered report protocol template, a pre-data collection template, the template from AsPredicted, a post-completion template, and a preregistration template in social psychology. Both platforms elicit similar information but ask slightly different questions to guide researchers through the preregistration process (see Table 1 for a list of questions of both templates and supplementary materials for both template interfaces). Notably, both AsPredicted and OSF provide detailed answer examples for each field. Researchers can also use open registration by posting a customized PDF instead of the form on existing registries [62].

2.2 Preregistration Benefits and Drawbacks

Preregistration can help research teams to reach consensus on specific research procedures at an early stage in their work [83]. Designing a research study typically demands that researchers identify abstract concepts, determine measures, assign different conditions, formulate hypotheses, and decide sample sizes, among other study-specific requirements [40, 90]. These tasks consume considerable time. Preregistration facilitates this process by inviting all research members to reach agreement in a timely manner. For example, a team member can preregister an initial study plan, after which the team can make changes and deliberate until a final agreement is reached. As a result, it encourages collaboration and team science that facilitates complete and high-powered study designs [60].

Preregistration also promotes trustworthy and transparent analysis. Researchers have argued that it may mitigate HARKing (Hypothesizing After Results are Known [46]), a sloppy research practice of post hoc "fishing" for a significant effect after data has been collected [81]. This typically occurs when researchers conduct Null Hypothesis Significance Testing (NHST), which relies on a critical p-value to assess whether an effect is significant. In response, preregistration timestamps any variables, hypotheses, and analysis plans before data collection, mitigating "questionable research practices" [39]. Evidence in clinical trials shows that preregistration reduced the number of studies that rejected null hypotheses from 57% to 8% [42]. As a result, it de-emphasizes a commonly used dichotomous threshold (e.g., $p < 0.05$) and reinforces rigorous study designs and method planning [94]. Researchers may be more likely to conduct risky but rigorously planned studies that could produce null results [15]. More studies reporting null results increases the likelihood of more publications with non-significant findings; this weakens the "file-drawer effect" [73], whereby studies that do not reach the critical threshold are more likely to be rejected for publication and kept hidden. Because other researchers cannot benefit from seeing such null results, they may be prone to repeating similar dead-end research studies.

While rapidly gaining attention, preregistration is also under careful scrutiny. A recent study of 27 papers with preregistered badges in *Psychological Science* found that all published studies deviated from the preregistered plans, but only one explained the deviation [13]. Yamada [94] noted that researchers could selectively report data by over-issuing multiple preregistrations or "preregistering" after completing an experiment. Instead of solely emphasizing preregistration, some researchers argue that a reputable scientific report should focus on whether an analysis is appropriate given the nature of the data [50, 63, 69, 76].

Promoting preregistration of qualitative studies also has led to intense debate in the research community. Compared to quantitative methods, qualitative studies emphasize the qualities of a particular technology and how people interact with it, thus generating theories or hypotheses rather than testing hypotheses based on experimental manipulation [4, 80]. For example, qualitative researchers may apply thematic or grounded theory analysis for data collected from ethnographic, interview, focus groups, diary or observational studies [4]. The high level of subjectivity in qualitative studies might challenge the notion of preregistration, which generally promotes replication of objective findings [78]. In fact, the

AsPredicted Template	OSF Standard Preregistration Template
1. Data collection. Have any data been collected for this study already? 2. Hypothesis. What's the main question being asked or hypothesis being tested in this study? 3. Dependent variable. Describe the key dependent variable(s), specifying how they will be measured. 4. Conditions. How many and which conditions will participants be assigned to? 5. Analyses. Specify exactly which analyses you will conduct to examine the main question/hypothesis. 6. Outliers and Exclusions. Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations. 7. Sample Size. How many observations will be collected, or what will determine sample size? 8. Other. Anything else you would like to preregister? 9. Name. Give a title for this AsPredicted preregistration.	1. Registration Metadata. (Title, Description, Contributors, Category, Affiliated institutions, License, Subjects) 2. Study Information (Hypothesis) 3. Design Plan (Study type, Blinding, Is there any additional blinding in this study? Study design, Randomization) 4. Sampling Plan (Existing Data, Explanation of existing data, Data collection procedures, Sample size, Sample size rationale, Stopping rule) 5. Variables (Manipulated variables, Measured variables, Indices) 6. Analysis Plan (Statistical models, Transformations, Inference criteria, Data exclusion, Missing data, Exploratory analysis) 7. Others

Table 1: The questions in the two main preregistration templates on AsPredicted and OSF. Most questions are followed by an open-ended text box and include example content.

call for preregistration and even transparency in research, primarily focusing on quantitative methods, has provoked disagreement among qualitative scholars in other social-science fields [15, 38, 49]. For example, the Qualitative Transparency Deliberations (QTD), a three-year deliberative process by hundred of political scientists, questioned the usefulness of research transparency [38].

In summary, preregistration does not address all questionable research practices [39, 94] and adapting preregistration to qualitative studies is open for discussion in the broader research community [15, 38, 49]. Our work extends this prior literature by investigating how the preregistration process in quantitative studies can be improved to promote sound research practices.

2.3 Preregistration in HCI

The HCI community has increasingly advocated adoption of preregistration practices, but the extent to which it should be adopted invites ongoing conversations. Prior work has addressed the limitations of NHST [11, 15, 23, 45] and shown a continued prevalence of dichotomous inferences in CHI studies [8]. A Special Interest Group (SIG) at CHI 2016 identified HARKing as a problem of statistical practices in HCI and discussed ways to move forward with transparent statistics [43]. In 2017, organizers of a CHI workshop proposed developing detailed guidelines that would add voluntary preregistration to the CHI reviewing process [44]. In 2018, another SIG on Transparent Statistics Guidelines solicited feedback from the HCI community on a first working draft of the guideline, which included preregistrations [86]. Cockburn et al. raised awareness of preregistration and encouraged HCI researchers to preregister their studies [15]; the authors anticipated a substantial growth of preregistered studies and a consequent reduction in the proportion of significant study results. Preregistration has also been treated as a user-centered design problem in an interview with researchers who ranged in seniority and experience [70].

Despite the continued interest, empirical evidence of whether and how HCI authors adopt preregistration remains unexplored. A recent study of TOP Guidelines in HCI studies [7] found that only 2 of 51 HCI journals (4%) specify guidelines for preregistration. A study of 119 CHI PLAY papers [84] showed that over half of their papers employ NHST without specific statistical hypotheses or research questions, a finding that preregistration could redress. Our work is motivated by the benefits of preregistration and the HCI community's continued interest in this topic. We aim to improve preregistration practices to promote its use among HCI researchers.

2.4 Tools for Study Planning and Data Analysis

Apéritif's goal of expediting preregistration, data analysis, and report drafting is also informed by preregistration's relevance for statistical analysis and method planning. Prior to preregistering their finalized plan, researchers use various tools to design rigorous experiments. WexTor [71] is a web-based tool that helps researchers learn how to design an experiment in a step-by-step process. Touchstone [57] is another platform designed to specifically support HCI researchers as they design and launch studies, albeit one with a limited data analysis function. Touchstone2 extends the platform by supporting counterbalancing processes and a priori power analyses [25]. It offers a declarative language to specify each experiment. Similarly, DeclareDesign [9] describes research designs in R code and simulates them to better illustrate their properties. Argus [89] simulates data and visualizes statistical power across different scenarios to faciliate effect size usage and a priori power analysis. NexP [58] assists beginners who lack expertise in statistical analysis and controlled experiment implementation. The Flex-ER platform [56] helps researchers prototype and conduct user studies to evaluate interaction techniques and is designed specifically for immersive visualizations.

After designing studies, researchers run experiments and analyze data to test and examine hypotheses in quantitative studies. Programming languages, such as R, Python and STATA, allow direct, lower-level interactions with a dataset but require knowledge of the language's syntax and its libraries. ExperiScope [33] supports users in analyzing complex data logs for interaction techniques. Statsplorer [87], an educational web tool, supports statistical analysis by interacting with visualizations. Tea, a domain-specific language (DSL), lets users specify their variables, study designs, and hypotheses at a high level [41]. Employing constraint-based reasoning, Tea automatically selects statistical tests for a given study design, with or without collected data.

While some existing tools guide users in designing experiments and analyzing results, none integrate with existing preregistration platforms. Additionally, existing tools and DSLs target different aspects of a research process such as data analysis, visualization, or power analysis. Apéritif aims to scaffold the preregistration process, integrating specialized tools where appropriate, using a unified interface on top of an existing preregistration platform.

3 FORMATIVE STUDY

Our formative study was guided by three research questions: (RQ1) How prevalent is the preregistraton of CHI studies?; (RQ2) Is the information included in CHI preregistrations sufficient to comprehend the research study design and is it consistent with the paper?; and (RQ3) What are the motivations, benefits, and challenges for CHI authors and reviewers of preregistered studies?

To answer these questions, we started with a systematic analysis of preregistrations and their corresponding papers over the past four ACM CHI conference proceedings (2018-2021) and followed up with a survey of 11 authors of preregistered studies. We chose CHI because it is the premier HCI conference and its papers cover diverse methodologies and a variety of subdisciplines.

3.1 Method

3.1.1 Analysis of CHI Conference Proceedings. The HCI community does not use a single, domain-specific registry, unlike clinical trails. To analyze the prevalence of preregistrations (RQ1), we identified CHI papers between 2018–2021 with preregistered studies by querying the ACM Digital Library, using full text search and the following keywords[1]: *preregistration, pre-registration, preregister, pre-register, AsPredicted, Open Science Framework, OSF* and *Github*. The first four keywords cover papers that explicitly address preregistration. For example, a paper might link to the preregistration by stating: "We preregistered our hypotheses and methods prior to the study at *URL*." The latter four keywords cover papers that link to a preregistration or research repository, in the paper or supplementary materials. Our initial query returned 137 candidate CHI papers, from which we excluded 99 false positives by reading the paper and searching the research repository and supplementary materials. Finally, we followed the link and validated whether each linked preregistration was available online.

3.1.2 Analysis of Preregistrations. To analyze whether the content of the preregistrations fulfilled prior recommendations (RQ2), we compiled a list of what should be included in preregistrations from prior literature (shown in Table 2). Based on these recommendations, we extracted the following categories from each preregistration and corresponding paper: (1) a priori description, (2) research questions/anticipated outcomes (hereby RQ), (3) variables, (4) analysis plan, (5) design plan, (6) sampling plan, (7) analysis script, (8) and any additional information. We then labeled each of these eight categories according to three binary variables:

- *Available*: Is the category provided in the preregistration?
- *Comprehensible*: Is the category's content sufficiently comprehensible to support reproducibility?
- *Consistent*: Is the category's content consistent with the description in the paper?

A category was labeled as *available* if its content related to this category appeared anywhere in the preregistration, regardless of where this information was entered. We followed any link in the preregistration (e.g., to supplementary materials of a survey design). None of the preregistration templates explicitly requested analysis code. Therefore, we searched for any analysis code in the research repository (OSF and Github) if the paper provided a link to it.

We labeled a category as *comprehensible* if it contained sufficient information for us to comprehend the plan and potentially reproduce it. The first author initially coded each section in each preregistration based on Table 2. Then, all authors met twice to discuss the preregistrations and resolve any uncertain cases for consistent labeling. Our labeling was guided by the recommendations in Table 2, and each of the eight categories was scored with either a 0 or 1. Our labeling was independent of our perceived correctness of the study design. For example, we did *not* consider a study design to contain insufficient information if the study could have used a different or more sophisticated statistical model that might have been better aligned with the research questions. We also did *not* rate the preregistration based on presentation quality.

We labeled a category as *consistent* between a preregistration and corresponding paper if the two artifacts generally expressed the same intent, study design, and analysis methods, even if the wording differed. Note that we included both undisclosed and disclosed deviations from the preregistered protocol—the latter being explicitly mentioned in the paper (e..g, "beyond our preregistered method, we analyzed ...") [82]. Prior work found that disclosed deviations from a preregistered objective can boost confidence in claims whereas undisclosed deviations may reduce a study's credibility [62, 82].

One author of this paper reviewed all preregistrations and papers and discussed potential inconsistencies with the other authors in research meetings. If an inconsistency was found, we recorded whether the paper reported the deviation from the preregistration. We labeled analysis scripts, a priori description, and additional information as NA: rerunning the analysis scripts was beyond the scope of this project, and evaluating the latter two would have required domain knowledge and would be highly subjective.

3.1.3 Survey with Preregistration Users. We additionally surveyed HCI researchers who had authored and/or reviewed papers with preregistered studies to gain a deeper understanding of the benefits and challenges of preregistration (RQ3). We recruited participants via the *Transparent Statistics in HCI* group Slack workspace [35].

[1] The ACM Digital Library dataset is provided for non-commercial research purposes, courtesy of the Association for Computing Machinery, Inc. [2021].

Design Requirements	Recommendations
Preregistration as a template	A preregistration should include the following categories: **A priori Description**: this should describe the goals, intended outcomes and methods as specifically as possible. Experience has shown that the utility of preregistration increases with the specificity of the information recorded. [15, 62, 83] **Research Questions/Anticipated Outcomes**: For quantitative studies, especially studies that use NHST, preregistration should specify expected relationships between two or more variables. Qualitative studies should clearly state the fact if no formal hypotheses are present. [11, 15, 62, 83, 91] **Variables**: Preregistration should clearly define dependent and independent variables and how researchers will measure them. Conditions are levels of the independent variable that are manipulated by the researcher in order to assess the effect on a dependent variable. [11, 15, 62, 83, 91] **Design Plan**: For quantitative studies, especially studies that use NHST, preregistration should include how researchers assign their manipulated conditions (within- or between-subjects). All preregistrations should include study procedures, including the participants' tasks and measures. [11, 15, 62, 83, 91] **Analysis Plan**: For quantitative studies, especially studies that use NHST, preregistration should include assumptions and statistical models. For qualitative studies, preregistration should include the data analysis approach (e.g., thematic analysis, content analysis, and grounded theory) and the data analysis process (e.g., who will be involved in the analysis, and what evidentiary criteria will be used to assess the qualitative research question). [11, 15, 34, 62, 83, 91] **Sampling Plan**: Preregistration should include sample size, sample size rationale, and exclusion and inclusion criteria. [10, 11, 15, 34, 62, 83, 91] **Additional Information**: Informational that is not covered by template questions. [15, 62] **Analysis Script**: Data analysis files, such as Python or R scripts, should be uploaded. [15]
Preregistration as a practice	**Research Process**: Preregistration should be integrated into researchers' existing workflow. [15, 70] **Consistency Across Research Artifacts**: Preregistration should enable easy preregistration-to-paper checking. [15, 70]

Table 2: Design requirements and recommendations for preregistrations based on past literature.

The group ran Special Interest Groups (SIGs) on transparent statistics at CHI 2016 and 2018 [43, 86], and a workshop at CHI 2017 [44]. Our survey asked open-ended questions about the *motivation* for, and the *challenges* of, reviewing preregistered studies and completing preregistrations. We preregistered our survey[2] and included the survey questions as supplementary materials. 11 participants completed the survey, including 3 who had both authored and reviewed papers with preregistered studies. The participants had previously preregistered studies using AsPredicted (3/11) and OSF (8/11).

3.2 Results

3.2.1 Prevalence of Preregistrations in CHI (RQ1). Of 2,874 papers published in the CHI proceedings from 2018–2021, 38 papers (1.32%) included at least one link to a preregistration. Out of those, 6 papers contained links to non-existent pages or empty preregistrations. Overall, we found 32 papers (1.11%) with available preregistrations, authored by 108 unique authors, 96 of whom authored exactly one of 18 papers and 12 authored at least two of the remaining 14. The 32 papers linked a total of 47 preregistrations, which spanned quantitative (43/47) and qualitative studies (4/47). Among the 43 quantitative studies, 32 used NHST, 8 used interval estimation, and 3 used Bayesian statistics. Among the 4 qualitative studies, 2 used interviews and 2 used surveys. Among the 32 NHST studies, 2 reported non-significant findings.

Figure 1 shows an increase in the number of papers using prereigstration 2018–2020 and a similar number for 2020 and 2021. Of the 32 papers, 16 used AsPredicted, 12 used OSF, and 4 used open preregistration. The first two provide a fixed template, while the latter allows authors to customize their preregistration. (None of the open preregistrations were timestamped, and all papers that linked multiple preregistrations consistently used the same registry.)

3.2.2 Comprehensibility of Preregistrations and Consistency with Papers (RQ2). Figure 2 shows that most of the 47 preregistrations included information about RQs (43), variables (44), design plan (40), analysis plan (43), and sampling plan (44). This makes sense given

[2] https://osf.io/uk75c/?view_only=1aa3003d1e42411a99f2391606566e65. Note that we initially preregistered to survey authors of CHI papers, similar to the approach in [85], but decided to recruit preregistration users from the active Slack workspace due to the low number of preregistered CHI studies.

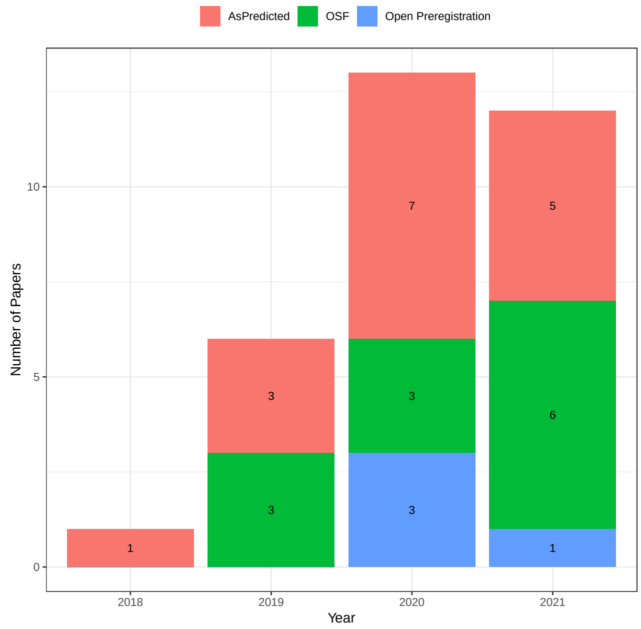


Figure 1: The number of CHI papers with at least one preregistration, broken down by year and registry. There were a total of 32 papers with a total of 47 preregistrations.

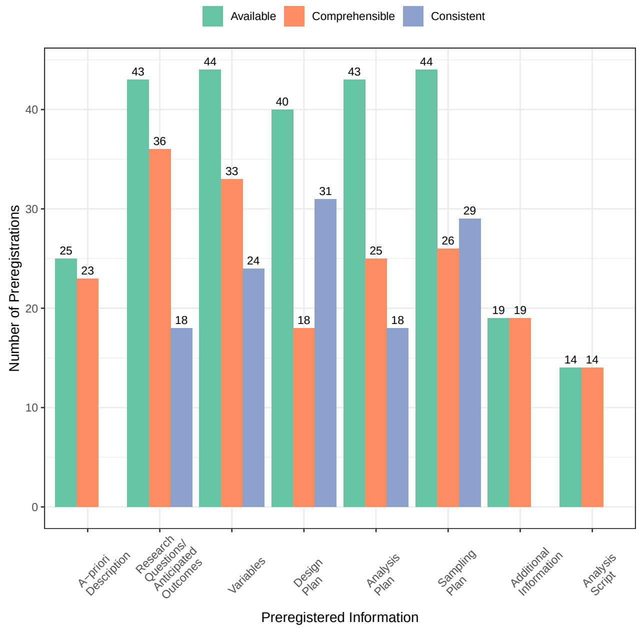


Figure 2: The number of preregistrations (out of 47) with information that was comprehensible and consistent with the paper, broken down by category of information. (Note that we did not label consistency for a-priori description, additional information, and analysis scripts.)

that both the AsPredicted and OSF templates specifically ask about these five categories. Only 25 preregistrations contained an a priori description, presumably because the AsPredicted template does not explicitly ask about it. Several preregistrations also reported on additional information (19) and analysis scripts (14).

From analyzing the *comprehensibility* of each preregistration (PR), we found that not all preregistered information was sufficiently detailed for us to understand and potentially reproduce the research study. Specifically, comprehensibility varies across categories: RQs (36/43), variables (33/44), design plan (18/40), analysis plan (25/43), and sampling plan (26/44). For example, PR20 introduced the goal of the study and explained the related work in detail but did not describe any research questions. In a study that used NHST, PR19 preregistered the hypothesis section ("This study is examining the usability of [Interaction A]") but did not mention null or alternative hypotheses that other researchers could test. In terms of variables, P3 preregistered a list of high-level concepts, such as *user perception*, but did not explain how to concretely measure them, or what their variable type is (e.g., categorical or continuous). When specifying the analysis plan, PR22 broadly stated that "we will test if there is a significant difference in [Measure A] between [Condition A] and [Condition B]" without specifying any statistical methods to be used in the study. For the sampling plan, PR31 intended to recruit 40–60 participants but did not mention the rationale, inclusion or exclusion criteria, and the sampling procedure.

When analyzing *consistency*, we observed that 45/47 (95.74%) preregistrations (PR) were inconsistent with the corresponding 31/32 (96.87%) papers (P). In other words, only one paper precisely adhered to the preregistered plan. For example, P17 reports results using interval estimation, guided by five hypotheses, none of which appears in the preregistration (PR21). P31 reported that the researchers conducted 31-40 trials per participants but preregistered 30 trials per participant in PR47. P35, a paper that used Bayesian statistics, described a non-preregistered analysis with Bayesian linear regression but claimed that the method was preregistered. In terms of sampling plan, P24 planned to recruit 120 participants in PR31, but the actual sample size was 50. In short, we found that a number of preregistrations were inconsistent with the papers in terms of RQs (18/43), variables (24/44), design plans (31/40), analysis plans (18/43), and sampling plans (29/44). Only 3 of the 31 papers with inconsistencies reported on all of them. These papers explicitly addressed the inconsistencies and offered a justification for a deviation from the plan. For example, P6 mentioned that "We originally pre-registered a 2 × 2 repeated measure ANOVA" and went on to justify the performed 2 × 3 repeated measure ANOVA with 3 levels of a condition in a dedicated paragraph. Note that our results are consistent with prior work that analyzed 27 preregistered studies published in *Psychological Science* [13], finding similar deviations. Overall, our results underline that inconsistencies are frequent—likely due to the fact that preregistrations are a separate artifact from the final paper, with little to no support to keep the contents of the two in sync.

3.2.3 Motivations, Benefits, and Challenges for CHI Authors and Reviewers of Preregistered Studies (RQ3). Our survey results show that both authors and reviewers recognized the benefits of using preregistration when stating their *motivation* for preregistration.

Participants (P refers to participants hereinafter) had various motivations for preregistration. Four participants (P3-6) indicated that preregistration helped them make a snapshot of research decisions, and four researchers (P1-3, P7) wanted to openly communicate research goals and design. Three participants (P3, P6-7) said preregistration helped them define research specifics in the presence of too many degrees of freedom. As P7 said:

> "My research starts with exploratory studies of user needs followed by empirically testing if [an interaction] could be effective. Sometimes, I felt that my studies had too much flexibility, and I was letting it slide based on the shared belief iterative design processes do require some flexibility. However, preregistration would help me have a structured plan prior to data collection to limit the excessive flexibility." [P7]

Two researchers (P2-3) said preregistration avoids publication bias in case the result is null. P1 also acknowledged that "it might look better to reviewers to show that we were rigorous in planning out the study." Our survey generally confirmed Pu et al.'s finding [70] that preregistration delimits flexibility and increases transparency.

All participants who had previously reviewed preregistrations stated that they use them to obtain more information and check for any deviations from the paper. For example, P10 stated the reason was to understand "some important details that were unclear in the paper." P11's motivation was to "investigate how sincere and meticulous the authors were with method planning and justification" and whether "they had followed their initial plan and gave proper justifications for why they had to stray from the initial plans if they did in their study." In one paper he reviewed, P9 indicated that the sample size and exclusion criteria differed and he felt less confident in the scientific rigor of the work. P10 found that "one study has changed the analysis description, but the rationale for the change was not sufficiently explained." These findings suggest that reviewers benefit from additional information provided in preregistrations and that consistency is an important concern for them.

One of the reviewers' major challenge was the different levels of detail in the preregistrations they reviewed. P9 said "some preregistration can have lots of information repeated multiple times and go into too much details about the motivation, which the preregistration isn't really for." This made them tedious and time consuming to review. P10 thought that "It could be unclear where to find the information I want." They re-emphasized the importance of consistency, expressing that "I wish I could be directed to the related parts of the preregistration."

Participants who preregistered studies reported similar challenges. It was generally perceived as difficult to know how much detail to provide (P1-3, 5-8). For example, P3 noted that "it is still a bit unclear what has to be included or which template is the right one." P7 thought that "there was so much freedom with the preregistration template," and added "more than what was required to give the readers a better understanding." P2's lab resolved the issue by having a separate standard to the analysis plan.

Another challenge is additional time and efforts for authors. P11 expressed that "definitely time" is a concern for him that might have made him choose not to preregister otherwise. He explained "If I have enough time to put in extra effort to the study planning, I would do so but sometimes it's not realistically possible." Three other participants (P1, P3, P7) also considered time and effort to be a challenge, in particular when it is not required.

Additionally, two researchers (P1, P8) stressed that the current preregistration is hard to evolve as the study unfolds. P1 said, "I am the kind of person who makes changes to studies frequently, so I wanted to thread the line between being specific, but making sure I had a little wiggle room if some particular feature of my design didn't work." P8 commented that "preregistration definitely takes extra time and effort, but it helped me to design my study but didn't do much when we changed [our plan]. So I didn't attach the URL in my submission." These findings indicate that authors perceived preregistrations as potentially inflexible and potentially disadvantageous if the final study and analysis plan deviated.

3.3 Design Requirements

Based on the results of our formative study, we conclude that an effective interface to preregistration should:

D1 **Scaffold** the preregistration template to elicit necessary and specific information. This addresses the issue of current templates posing a challenge for anticipating an adequate level of detail for the provided information.

D2 **Integrate** preregistration into the research process and connect it with other research artifacts. This addresses the issue of inconsistent information in preregistrations and papers. Preregistration authors could also gain a quick overview of where inconsistencies arise.

D3 **Track** the evolution of a research plan. This addresses the desire for more flexibility, in particular making (justified) changes after a study has been preregistered. Since flexibility should not come at the expense of consistency, version control for preregistrations could support evolving, yet consistent, research artifacts.

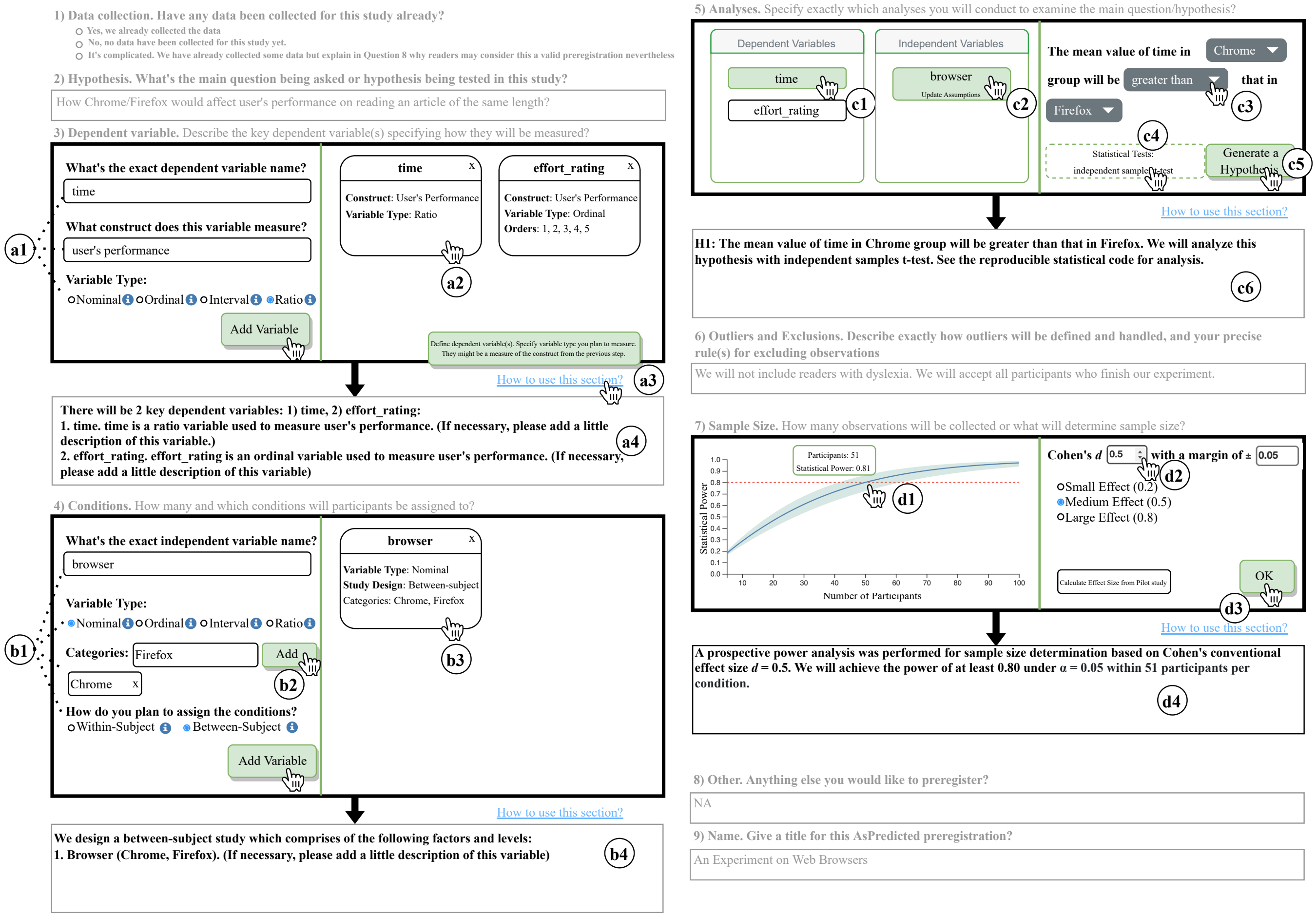
D4 **Reduce** the time and cognitive effort needed to complete a preregistration. The overall goal is to lower the barriers to using preregistration, avoid errors, and speed up the process.

In addition, our findings revealed that most preregistrations in CHI used the AsPredicted platform, a vast majority of them preregistered quantitative experiments, and most of them used NHST, which continues to be the most prevalent statistical analysis paradigm in HCI [8, 24]. We therefore decided that supporting the preregistration of quantitative studies with NHST statistical analyses is a reasonable first step to studying possible remedies to the challenges identified in our formative study.

4 APÉRITIF

Based on our findings and the design requirements, we developed Apéritif, a research prototype that is built on top of the AsPredicted preregistration platform. Apéritif is designed to scaffold the preregistration process and integrate it into the research workflow. In particular, Apéritif is designed to meet the design requirements *D1-4* and allows us to study how scaffolding is perceived by researchers. Apéritif is a Chrome extension and publicly available as open-source to enable others to build on it.[3]

[3] https://github.com/rrrrrrockpang/aperitif



1) Data collection. Have any data been collected for this study already?

○ Yes, we already collected the data
○ No, no data have been collected for this study yet.
○ It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid preregistration nevertheless

2) Hypothesis. What's the main question being asked or hypothesis being tested in this study?

How Chrome/Firefox would affect user's performance on reading an article of the same length?

3) Dependent variable. Describe the key dependent variable(s) specifying how they will be measured?

What's the exact dependent variable name? time

What construct does this variable measure? user's performance

Variable Type: ○Nominal ○Ordinal ○Interval ●Ratio

time — Construct: User's Performance; Variable Type: Ratio

effort_rating — Construct: User's Performance; Variable Type: Ordinal; Orders: 1, 2, 3, 4, 5

Define dependent variable(s). Specify variable type you plan to measure. They might be a measure of the construct from the previous step.

There will be 2 key dependent variables: 1) time, 2) effort_rating:
1. time. time is a ratio variable used to measure user's performance. (If necessary, please add a little description of this variable.)
2. effort_rating. effort_rating is an ordinal variable used to measure user's performance. (If necessary, please add a little description of this variable)

4) Conditions. How many and which conditions will participants be assigned to?

What's the exact independent variable name? browser

Variable Type: ●Nominal ○Ordinal ○Interval ○Ratio

Categories: Firefox; Chrome

How do you plan to assign the conditions? ○Within-Subject ●Between-Subject

browser — Variable Type: Nominal; Study Design: Between-subject; Categories: Chrome, Firefox

We design a between-subject study which comprises of the following factors and levels:
1. Browser (Chrome, Firefox). (If necessary, please add a little description of this variable)

5) Analyses. Specify exactly which analyses you will conduct to examine the main question/hypothesis?

Dependent Variables: time, effort_rating

Independent Variables: browser

The mean value of time in Chrome group will be greater than that in Firefox

Statistical Tests: independent samples t-test

H1: The mean value of time in Chrome group will be greater than that in Firefox. We will analyze this hypothesis with independent samples t-test. See the reproducible statistical code for analysis.

6) Outliers and Exclusions. Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations

We will not include readers with dyslexia. We will accept all participants who finish our experiment.

7) Sample Size. How many observations will be collected or what will determine sample size?

Participants: 51
Statistical Power: 0.81

Cohen's *d* 0.5 with a margin of ± 0.05

○Small Effect (0.2)
●Medium Effect (0.5)
○Large Effect (0.8)

A prospective power analysis was performed for sample size determination based on Cohen's conventional effect size *d* = 0.5. We will achieve the power of at least 0.80 under α = 0.05 within 51 participants per condition.

8) Other. Anything else you would like to preregister?

NA

9) Name. Give a title for this AsPredicted preregistration?

An Experiment on Web Browsers

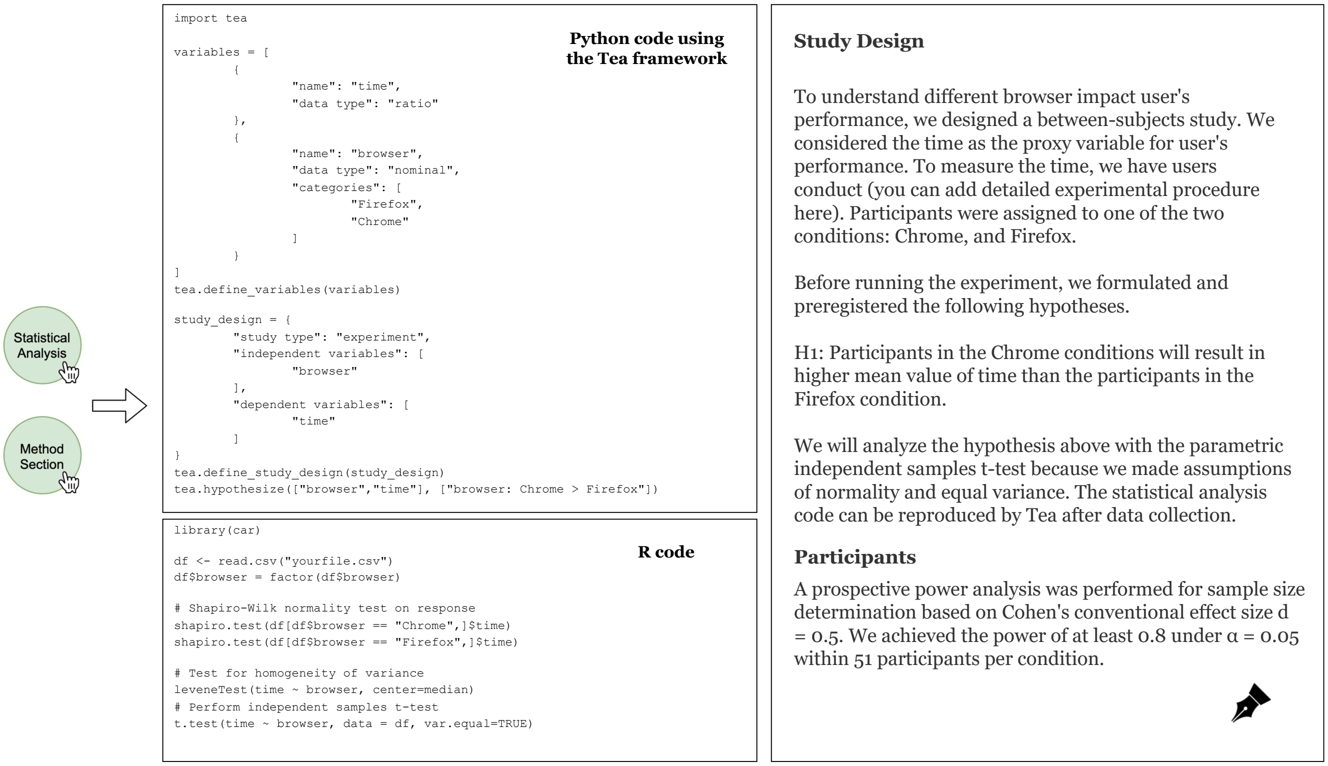
Figure 3: Apéritif's user interface. With Apéritif, a researcher finish a preregistration form by completing the Apéritif toolboxes (abcd) and subsequently editing the text boxes (a4, b4, c6, d4). Question 1-9 in grey are the same set of questions on AsPredicted. Bolded boxes are Apéritif toolboxes (abcd), whereas the grey boxes are original textboxes on AsPredicted. The bolded texts (a4, b4, c6, d4) in the grey textboxes are generated by Apéritif.

4.1 Overview of Apéritif

Apéritif augments the AsPredicted preregistration template, which includes nine questions (Table 1), with specific questions, interactive tools, and visualizations. In the original template, each question is followed by an open-ended textbox, with the exception of the first question (Data collection). Apéritif introduces four *toolboxes* (shown with a bold border in Figure 3), each connecting a question prompt above to a *textbox* below. These four *toolboxes* guide users' answers to Question 3 (dependent variable), Question 4 (conditions), Question 5 (analyses), and Question 7 (sample size). Rather than providing only an open-ended textbox, Apéritif asks specific subquestions and constrains the input to scaffold what information can be entered. The toolboxes for Questions 3–4 elicit information about variables and study design on the left (**a1** and **b1**) and display the captured information as blocks (**a2** and **b3**) on the right. Users can add, edit, and remove these blocks as necessary. The toolbox for Question 5 establishes a link between the dependent and independent variables provided in the previous two toolboxes to the specified analyses. Guided by the text suggestion on the right (**c3–c4**), users can specify the relationship between any variables on the left. In Question 7, users can hover over the visualization on the left, which is derived from the selected effect size on the right. Apéritif does not augment the remaining questions in the AsPredicted template.

For each question, users complete the corresponding toolbox, and Apéritif automatically populates the preregistration textbox (**a4**, **b4**, **c6**, and **d4**) with the answer text. Users can edit the text to provide more details, and Apéritif explicitly reminds them to expand on the variable that needs further description.

As users complete the toolboxes, Apéritif offers real-time updates to the generated analysis code and methods section, which users can view and edit by clicking the respective buttons shown in Figure 4. Apéritif currently does not update the preregistration based on changes made to the analysis code and methods section. After completing the nine questions with Apéritif, users can choose



```python
import tea

variables = [
        {
                "name": "time",
                "data type": "ratio"
        },
        {
                "name": "browser",
                "data type": "nominal",
                "categories": [
                        "Firefox",
                        "Chrome"
                ]
        }
]
tea.define_variables(variables)

study_design = {
        "study type": "experiment",
        "independent variables": [
                "browser"
        ],
        "dependent variables": [
                "time"
        ]
}
tea.define_study_design(study_design)
tea.hypothesize(["browser","time"], ["browser: Chrome > Firefox"])
```

```r
library(car)

df <- read.csv("yourfile.csv")
df$browser = factor(df$browser)

# Shapiro-Wilk normality test on response
shapiro.test(df[df$browser == "Chrome",]$time)
shapiro.test(df[df$browser == "Firefox",]$time)

# Test for homogeneity of variance
leveneTest(time ~ browser, center=median)
# Perform independent samples t-test
t.test(time ~ browser, data = df, var.equal=TRUE)
```

Study Design

To understand different browser impact user's performance, we designed a between-subjects study. We considered the time as the proxy variable for user's performance. To measure the time, we have users conduct (you can add detailed experimental procedure here). Participants were assigned to one of the two conditions: Chrome, and Firefox.

Before running the experiment, we formulated and preregistered the following hypotheses.

H1: Participants in the Chrome conditions will result in higher mean value of time than the participants in the Firefox condition.

We will analyze the hypothesis above with the parametric independent samples t-test because we made assumptions of normality and equal variance. The statistical analysis code can be reproduced by Tea after data collection.

Participants

A prospective power analysis was performed for sample size determination based on Cohen's conventional effect size d = 0.5. We achieved the power of at least 0.8 under α = 0.05 within 51 participants per condition.

Figure 4: The statistical analysis and methods description generated by Apéritif.

to connect Apéritif with their Github account, create a repository, and commit the three research artifacts (preregistration, analysis code, and methods section). By using version control for these artifacts, authors and reviewers can see any changes that have been made after the initial preregistration was created.

4.2 Scaffolding the Preregistration Process

Apéritif's questions prompt users to disambiguate and operationalize research questions, develop concrete measurements, and weigh trade-offs among different study designs. For example, it asks users to specify a variable by including its name, any construct it measures, and its type (**a1** in Figure 3). Apéritif supports nominal (e.g., interaction techniques), ordinal (e.g., Likert-scale), interval (e.g., test scores), and ratio (e.g., time) data types [22]. If a variable is categorical (nominal or ordinal), Apéritif interactively shows a question asking users to specify the categories (e.g., Chrome/Firefox or Strongly Agree/.../Strongly Disagree, see **b2**). Additionally, it allows users to assign conditions to participants (e.g., between- and within-subject designs). Users can add multiple independent variables, but currently use only one independent variable per hypothesis (see section 4.3 for details). Apéritif captures users' alternative hypotheses, predicting an effect or relationship between two variables [40, 41]). After all variables are specified, users formulate hypotheses by clicking on the dependent and independent variables (**c1**–**c2**) and specifying the relationship between variables (**c3**).

Apéritif also lets users state assumptions based on domain knowledge (e.g., for an independent variable in **c2**). Often, users' assumptions are specific to variables and properties, such as independence, normality, and homoscedasticity (or equal variances) [41]. Note that all assumptions are selected in Figure 3. In our formative study, only two preregistrations considered assumptions when planning statistical analysis. Apéritif makes assumptions optional in case the knowledge required to express assumptions is lacking at this stage. A lack of assumptions results in a more conservative test selection, which Apéritif indicates to a user.

Users can perform an a priori power analysis (**d**), which allows them to calculate the minimum sample size required to achieve a sufficient level of statistical power ($1 - \beta$) based on the probability of Type I errors (α) and a minimal effect size for the target population. Apéritif asks users to input the anticipated effect size and visualizes statistical power as a function of the sample size. Apéritif defaults to $\alpha = 0.05, 1 - \beta = 0.8$ [16]. This design is inspired by Touchstone2 [25], which was shown to lift barriers to estimating and interpreting standard effects. We did not integrate the tool itself to keep Apéritif's interface unified, but rather relied on the `statsmodels` Python library [3] for the backend. The current Apéritif prototype implements the power analysis for parametric tests (i.e., t-test and ANOVA), both for within and between subjects designs. Users can specify the expected effect size (e.g., derived from a pilot study) in terms of mean differences or Cohen's d. A user can hover over the visualization (**d1**) to explore the sample size and corresponding statistical power for a given effect size (**d2**).

4.3 Integrating into the Research Process

Apéritif aims to seamlessly integrate preregistration into the investigative process by providing reusable statistical analysis scripts

Statistical Tests
Comparison Tests *(DV=Interval/Ratio, Ordinal; IV=Nominal)*
• Independent samples t-test
• Paired-samples t-test
• Wilcoxon rank-sum test (Mann-Whitney U test)
• Wilcoxon signed-rank test
• One-way ANOVA
• One-way repeated measures ANOVA
• Kruskal-Wallis test
• Friedman test
Correlation Tests *(DV=Interval/Ratio, Ordinal; IV=Interval/Ratio, Ordinal)*
• Pearson's r
• Kendall's τ
• Spearman's ρ

Table 3: Statistical tests, along with the corresponding type of dependent variable (DV) and independent variable (IV), for which Apéritif can generate analysis code.

and text suggestions for a methods section. A key challenge to generating statistical analysis scripts is to determine a valid statistical analysis for the specific hypotheses in question. Apéritif uses Tea [41], a constraint-based framework and domain-specific language (DSL) that selects statistical tests based on study specifications and hypotheses. Apéritif translates high-level preregistration information to lower-level Tea input and obtains an appropriate statistical test using its API. In an evaluation of 12 statistical tutorials, Tea generally agreed with expert recommendations and was more conservative in the presence of non-normal data, minimizing the risk of false positive findings [41]. Apéritif presents Tea code as part of the generated analysis code (Figure 4). With the recommended tests, Apéritif also generates equivalent R code to support a wider population of researchers. The generated R code includes tests of assumptions (e.g., for normality and homoscedasticity), a procedure built into the Tea framework. After collecting data, users can rerun the analysis and examine the result using the preregistered plan, which is encapsulated in the analysis code.

We chose Tea for its ability to recommend statistical tests in the absence of an input dataset. However, due to limitations in Tea, a user can specify multiple hypotheses but only one independent variable per hypothesis in the initial version of Apéritif. Table 3 shows the statistical tests, as well as the independent and dependent variable combination, for which Apéritif can generate analysis code by querying Tea. Replacing Tea or improving it, e.g., with a more sophisticated constraint system could extend model coverage and address interaction effects. Such improvements are likely necessary to support a wide range of preregistrations.

Apéritif also translates all preregistered information into a methods description that can serve as both an a priori description and a methods section in a paper. It does so by inserting the preregistered information from the toolbox into a template, which is based on reporting guidelines in HCI [92] and psychology [91]. Apéritif synchronously updates the statistical analysis script and methods section as users complete each preregistration question.

4.4 Tracking the Evolution of a Research Plan

Apéritif leverages Git, a version control system. After completing preregistration, a user can initialize a Github repository using Apéritif, which uses the Github API. Current preregistration sites, including AsPredicted and OSF, prevent users from editing the preregistration once submitted. However, the concept of a preregistration that captures all aspects of an experiment is ideal in theory, but drafting a preregistration may require a number of iterations. When a revised preregistration is submitted on AsPredicted, the earlier version is permanently deleted. The OSF repository has built-in version control for all files, but it retains only copies of a file added to OSF instead of concrete change history records. We chose Git for its tracking functionality and scalability, as well as its support through the Github REST API.

4.5 Implementation Details

The Apéritif Chrome browser extension and web application are implemented in HTML, CSS, JavaScript, and Python using the Flask framework [32]. Apéritif uses Heroku [1] for hosting and MongoDB [2] for data storage. It can be extended to support the OSF platform by reusing the existing backend and overlaying Apéritif toolboxes on corresponding template sections. Likewise, components in Apéritif, such as the automated test selection and power analysis visualization, can be exchanged for other tools. The scaffolding, integrating, and tracking functions can also be extended to a web application if the HCI community decides to host its own domain-specific registry.

5 PREREGISTERED USER EVALUATION

We conducted a within-subjects user study to answer the following three questions:

(1) To what extent does Apéritif reduce the time and cognitive effort needed to complete a preregistration?
(2) What is the value, if any, of Apéritif's artifact generation over simply completing a preregistration template?
(3) How can Apéritif be improved?

We used Apéritif to preregister this study and reused the automatically generated analysis code and methods description as part of our research process. We slightly edited the description, changes that later informed improvements to Apéritif, and documented these changes. Our preregistration is available on AsPredicted[4], and our code, data, and documentation are available on OSF[5].

5.1 Apparatus

Our study was conducted remotely over Zoom. To facilitate remote participation, we did not ask participants to install Apéritif on their personal computer. Instead, we created a controlled set up—a temporary website that hosts the original AsPredicted template and the AsPredicted template augmented by Apéritif. Each participant interacted with both templates through the same website in our within-subjects experiment. For the evaluation, we opted for a self-contained system[5], as opposed to interfacing with external tools. Specifically, our system implemented the relevant test selection and power analysis functionality directly in JavaScript.

[4] https://aspredicted.org/blind.php?x=/CS4_JDR
[5] https://osf.io/tgacn/?view_only=cd81b7c90092458a95c25c49ec469f0f

5.2 Procedure

Participants were given a 10-minute introduction to the concept of preregistration as well as common preregistered information, including variables, statistical analyses, study designs, hypotheses, and sampling plans. We did not introduce Apéritif per se so we could evaluate how well novice users could interact with it. The within-subjects experiment included two conditions and research study scenarios: In one, participants were asked to preregister a prespecified research study with the original AsPredicted interface. In the other one, they were given a different research study of similar complexity and asked to register it using Apéritif. To ensure the same level of difficulty of the two conditions, we chose two similar study scenarios, which were adapted from an online course that teaches how to design, run, and analyze experiments in HCI [74]:

(1) Your research team is interested in learning about how iPhone vs Samsung devices affect a user's text entry input time. You are asked to preregister your study and analyze the data.
(2) Your research team is interested in learning about how Chrome vs Firefox browsers affect a user's speed when reading a long article. You are asked to preregister your study and analyze the data.

We used a counterbalanced, randomized design to avoid any systematic bias due to the order of the two conditions or the differences between the two scenarios. Each participant designed their own experiments based on these scenarios and moved forward if they decided the information would be sufficient.

After preregistering these study scenarios, participants were directed to write statistical analysis code to test their hypotheses in the first condition and to edit Apéritif's analysis code in the second. We provided synthetic data files for these analyses, and the participants analyzed the data with any software they felt comfortable with, such as Jupyter Notebook, PyCharm or RStudio. (We asked participants to set up all software on their own computer at the beginning of the experiment to avoid conflating our time measurement with the time to choose, open, and set up additional software.) To quantitatively evaluate participants' performance, we measured the time it took to create the preregistration and write the analysis code.

At the end, we administered a five-question exit questionnaire to probe the usefulness and usability of Apéritif (see Figure 5). Additionally, we conducted a semi-structured interview to further collect feedback about the utility of the interactions and the overall workflow when using Apéritif.

We did not ask participants to draft a report that described their method analysis to reduce the time of the study; a pilot study with three participants had previously shown that participants spent more than one hour preregistering the synthetic study, coding the analysis script and drafting a method section.

5.3 Participants

We hypothesized that Apéritif significantly improves researchers' efficiency, yielding a large effect size. A prospective power analysis was performed for sample size determination with Cohen's d = 0.8. According to the a priori power analysis, 14 participants were required to achieve 0.80 power at $\alpha = 0.05$. We used purposive sampling to recruit HCI researchers by identifying and contacting authors who had conducted experiments in prior publications. We included participants with varying experience with preregistration to evaluate whether Apéritif helped both existing and new users of preregistrations. Participants were required to have experience designing experiments and analyzing data with Python or R.

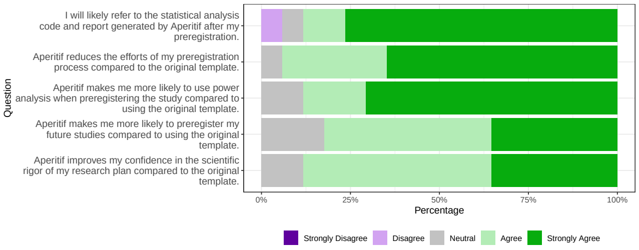


Figure 5: Likert-scale responses to the usefulness of Apéritif.

A total of 17 participants (9 identified as female, 8 as male) took part in our study. Among them, 15 were PhD students and 2 were undergraduate students. Five participants had used preregistration on OSF or AsPredicted; eight participants had heard of preregistration but had not used it; and four participants had not heard of preregistration. None of them were familiar with Apéritif, nor the content of this study. Each participant was compensated with a $15 Amazon gift card upon finishing the user study.

5.4 Results

5.4.1 Time. To quantify participants' performance, we measured the time it took them to complete the preregistration and write analysis code. We conservatively preregistered the non-parametric Wilcoxon signed rank test but instead conducted the Shapiro-Wilk normality test (Apéritif: $W = 0.97, p = 0.81$ and AsPredicted: $W = 0.93, p = 0.23$) and Levene's test for homogeneity of variance ($F(1) = 2.72, p = 0.11$). The test of assumptions, which was also generated by Apéritif, informed us to use a paired samples t-test. There was a statistically significant effect of condition ($t(16) = 5.03, p < 0.001$) such that participants completed tasks significantly faster with Apéritif (M = 11.25 minutes, SD = 3.28) than with AsPredicted alone (M = 17.67 minutes, SD = 5.34).

5.4.2 Questionnaire on Apéritif. Figure 5 shows an overview of participants' answers to our questionnaire. Almost all participants indicated that Apéritif reduced the perceived effort of completing a preregistration (64.71% strongly agreed, 29.41% agreed) and that Apéritif made them more likely to preregister their future studies compared with the original template (35.29% strongly agreed, 47.06% agreed). As a corollary of Apéritif's precise question scaffolding, we also saw that participants were more likely to use power analysis to determine sample size (70.59% strongly agreed, 17.65% agreed). Participants indicated that they would refer to the statistical analysis code and report generated by Apéritif after their preregistration (76.47% strongly agreed, 11.76% agreed). Overall, Apéritif improved participants' confidence in the scientific rigor of the research plan compared to the original template (35.29% strongly agreed, 52.94% agreed).

5.4.3 Post-Study Interview. In our interviews, all participants indicated that Apéritif's fine-grained questions gave them a good idea of what exact information to include in the template. Participants found that "by interacting with the visual components the preregistration seems less like a laborious form-filling exercise" (P6). The interaction and automatic research artifacts generation makes the preregistration "almost like a small game" (P9) that encourages users to "just want to continue to see what's next" (P10). Participants acknowledged that the generation of analysis code was beneficial. P4 stated that when he preregistered with the original interface, he spent "so much time googling what tests I should use, though I took a research method class last year". P5 echoed that "I needed to find my R script where I did my last experiment". Further, participants found that Apéritif saved them time on determining the analysis (P1-P5, P17) and implementing the hypothesis tests (P1-P8, P11, P14-P17). P6 mentions that "the final analysis might not be the same [as the preregistered one] in real-world research projects, but hey at least I have something to start with."

Although Apéritif's generated methods description is concise, participants commented that it provided "what I needed to start my own writing" (P14). P14 also commented that "if I were to preregister my own study, I would read so many other published tutorials [on the analysis methods] and spend hours before writing down my own answer". Participants indicated that they can just use the generated methods description as the input for the free-text preregistration summary (P1-3, P6-8, P14).

Participants also noted that Apéritif provides incentives for preregistration in the future. For example, P2 said, "I heard about preregistration, but I never think it's worth the effort. Now I can see some [analysis] code and a method section. That's pretty neat." P6 also mentioned that "now preregistration is advertised as a way to avoid HARKing. If you don't preregister for your experiments, people might wonder whether you did p-fishing. I think Apéritif gave me an alternative incentive to preregister."

Although Apéritif adds explanations for terms, such as construct, data type and within-subject/between-subject design, 9/17 participants went online to search for definitions of these. P14 said that "I ran into the terminology [construct] in the past, but I just didn't recall what it means. I needed to think about it for a second." To address these issues, we refined and detailed the explanations for the used terminology. Understanding the power analysis was a hurdle for 14/17 of the participants. Three participants had not heard of power analysis, and 11 participants had not used power analysis. In response, we added more text explanation to Apéritif.

We also found that several participants raised questions about how Apéritif may be able to handle qualitative studies. P6, a researcher who had preregistered before, mentioned that "this tool works really well with [a] quantitative study, but I'm not sure how I can use this for some of my interview study design." P8, who had also preregistered previously, cautioned that "[Apéritif] looks really cool to preview statistical scripts and the method section, but it seems that preregistration has to be written in a pretty rigid way. It may hamper the open mind that researchers need for qualitative work such as open coding and grounded theory." While the current version of Apéritif does not address this concern, future improvements should focus on similar scaffolding approaches specifically designed for qualitative and explorative research studies.

6 DISCUSSION

Our work shows that the adoption rate of preregistration in HCI remains low, despite researchers having anticipated a significant increase in preregistered studies [15]. The majority of preregistrations reported on studies that used NHST, and only two reported a non-significant finding. In contrast to the prediction [15], studies that report null results are still rare—or, in other words, the file drawer does not seem to have opened up among CHI researchers.

We additionally found that preregistration is still practiced by only a small number of researchers. However, even among this small group, preregistrations vary substantially in their level of details. Our survey results suggest that the varying levels of detail are a result of a lack of structure in existing preregistration templates, such as those provided by OSF or AsPredicted. Despite both templates providing example answers, researchers commonly stated that they do not know what level of detail to provide, nor is it easy for them to anticipate what information others may need.

Our study also revealed that authors of preregistrations commonly change study or analysis details after the preregistration has been "locked in." In other words, there are often inconsistencies between the description of a study in a preregistration and the corresponding paper. The finding reveals a trade-off between asking researchers to register their studies a priori and the reality, in which most research studies evolve and experience changes in one way or another. Changes to minor details, such as adjustments of participant numbers or specific analyses, should be expected; researchers have frequently been found to adapt their analysis procedure [30, 54, 75]. Our work confirms these findings by showing that the current preregistration process does not match the reality of the research process, in which researchers often make changes or add additional information while a study is underway.

Our work shows that scaffolding the preregistration process, as we exemplified in our research prototype Apéritif, could address both the issue that preregistrations commonly contain insufficient details and that they are often inconsistent with the description in the corresponding paper. Because preregistration should support planning a study rather than serving as a strict accountability tool [20], scaffolding focuses on adding specific information that is needed for understanding and potentially reproducing a study. Additionally, version control enables users to preserve the information of any version and to generate an up-to-date analysis script and methods section.

Of course, a key question is whether scaffolding has the potential to foster a greater adoption of preregistrations in the HCI community. In evaluating Apéritif, we found that the scaffold that the tool provides serves as an incentive for both thinking through details and for learning about rigorously designing studies. This added benefit may convince additional researchers to preregister their study, as suggested by our participants. It was encouraging to see that Apéritif helps novice researcher to consider important study design aspects, such as the proper definition of variables and determination of sample size, that might otherwise be overlooked.

While an added benefit is certainly needed to increase the adoption of preregistrations, a lack of time is usually another important factor that prevents researchers from preregistering their study [69, 83]. Scaffolding requires taking time up front to think through and answer specific questions, but reduces the time to write analysis code and a methods section by partially automating both. As the results of our evaluation showed, participants with Apéritif spent significant less time to preregister, write the analysis code, and analyze the data than participants using the conventional AsPredicted preregistration template. Participants expressed positive feedback on Apéritif's approach to scaffolding, which suggested they may be more likely to preregister future studies.

Interestingly, despite its detailed questions, we found that Apéritif does not add to researchers' cognitive load when completing a preregistration compared to the original template—in line with prior work that appropriate scaffolding with questions and tools facilitates the learning experience and helps with decision-making. Expert-curated scaffolding has been used to create scientific theories [67] and teach domain expertise [88] with crowds, design video-based reflection exercises to support student learning [72], and support developers to weigh trade-offs for coding solution [53]. Similarly, instead of asking researchers to simply "figure it out," Apéritif provides scaffolding for preregistration process.

6.1 Ways for Increasing the Use of Preregistration in the HCI Community

While scaffolding lowers barriers and provides incentives that may encourage more researchers to preregister (quantitative) studies, additional measures are likely needed to substantially increase adoption rate. For example, a recent study of the 51 journals in which CHI authors most frequently publish indicates that journals do not set or specify clear guidelines on preregistration [7]. In response, HCI conferences and journals could provide more specific expectations for preregistration and even create a submission track for registered reports, a two-phase preregistration submission process such as that used in psychology and political science. With registered reports, authors submit their research questions and methodology before observing the outcomes of the research, obtaining an in-principle acceptance for the paper that adheres to the preregistered plan [12, 62]. In computer science, the *Mining Software Repository* conference, a subcommunity in software engineering, established a registered report submission track in 2020 [17]. Once the report is accepted, the full study is guaranteed for publication in *Empirical Software Engineering* journal. CHI and other HCI venues could start similar tracks. Given the diversity of methodologies in the HCI community, it would be essential to encourage preregistration of all kinds of studies, including quantitative, qualitative, and exploratory.

Conferences could also award preregistered studies with *Preregistered Badges* [27] which have been adopted by 79 journals to promote preregistration. This will extend the current ACM badging system [26] that includes badges for *Artifacts Evaluated, Artifacts Available* and *Results Validated.* The *Preregistered Badge* would both acknowledge open science practices and incentivize more authors to preregister. Scaffolding can be useful approach for guiding new users through the process. When introducing such measures, it is likely that reviewers would have to learn how to review preregistrations. Our recommendations for what preregistrations should include, which we compiled based on a thorough literature review (see Table 2), could provide guidance and could even be turned into a checklist for authors and reviewers.

7 LIMITATIONS AND FUTURE WORK

The initial prototype of Apéritif allowed us to study the potential of scaffolding preregistration, but more sophisticated systems are necessary to support a wide range of preregistrations. We chose NHST as a starting point because the majority of studies in our CHI paper analysis used it, and because it generally remains widely used in HCI [8, 24]. Apéritif currently supports statistical analyses and sample size recommendations based on a limited set of tests in Table 3 under the NHST paradigm, though the HCI community has raised awareness and made efforts to incorporate other methods such as Bayesian analysis [45] and interval estimation [23]. Extending tools like Apéritif to support preregistering other forms of user interface evaluation, such as demonstration, usage, technical benchmarks and heuristics [51], under NHST [19, 40] and non-NHST frameworks [23, 45], will be an important next step.

Apéritif is based on the AsPredicted template which is tailored to quantitative work, but we are also aware of templates for qualitative studies [5, 34]. Our user study surfaced challenges to adapting Apéritif to qualitative studies with one participant suggesting that it might "hamper the open mind that researchers need for qualitative work." While qualitative researchers have raised concerns over the push for overarching transparency [38], future work needs to investigate how useful the current preregistration templates are for qualitative studies and explore how to support planning for qualitative studies with scaffolding.

Apéritif's approach to scaffolding leverages version control to track the evolution of a study design and analysis as inspired by our formative study. After submitting this paper, OSF pushed an update in December 2021 that enabled support for tracking the history of preregistrations and justifying changes, which suggests that version control has been a long-standing need. However, our user study did not thoroughly evaluate the benefits of this feature, how well it integrates with the research process for authors, or how it may be interpreted by reviewers. Future longitudinal studies should qualitatively assess how researchers use this feature as a research project evolves and what consequences may arise from it. For example, it may be necessary to integrate visualizations of a preregistration's history for optimal support [47], similar to analytic decision graphs [54]. The multiverse analysis concept [24, 55, 79] can also capture alternative analyses and decision making through a large set of reasonable scenarios and could provide inspiration for visualizing version control in preregistrations.

Importantly, future work should study deviations between papers and preregistrations in greater detail. For example, adding justifications to every change may be a great benefit—or a great burden—for researchers. It is also feasible that reviewers or readers might inadvertently penalize papers that exhibited many deviations from the preregistration. Future studies may inform detailed guidelines for communicating these changes.

8 CONCLUSIONS

In this work, we presented empirical evidence that preregistration remains an uncommon practice in HCI. In our content analysis of CHI proceedings, we found that current preregistrations suffer from insufficient details and inconsistencies between preregistrations and the corresponding papers. We treated these problems from a user-centered perspective by surveying 11 preregistration users. Based on this formative study and literature review, we identified the needs to scaffold the preregistration process, integrate it into researcher's workflow, and maintain artifact consistency. In response, we developed a research prototype Apéritif that builds on top of existing preregistration templates and generates analysis code and methods descriptions, based on specific questions. Our evaluation shows that this approach reduces the time of preregistering studies and enables artifact consistency as well as tracking of evolving preregistrations. We look forward to our work sparking new conversations on how to preregister in the HCI community and further the creation of tools that support preregistration practices.

ACKNOWLEDGMENTS

We thank Eunice Jun, Sherry Wu, Leilani Battle, Tal August, Qisheng Li, Enhao Zhang, Ruotong Wang, Mary Peng, Sandy Kaplan, and the anonymous reviewers for their valuable feedback and suggestions.

REFERENCES

[1] 2021. *Heroku.* Retrieved 2021-09-08 from https://www.heroku.com/

[2] 2021. *MongoDB.* Retrieved 2021-09-08 from https://www.mongodb.com/

[3] 2021. *Statsmodels.* Retrieved 2022-12-30 from https://www.statsmodels.org

[4] Anne Adams, Peter Lunt, and Paul Cairns. 2008. A qualititative approach to HCI research. (2008).

[5] David Thomas Mellor Alexandra Hartman, Florian Kern. 2018. *Preregistration for Qualitative Research Template.* Retrieved Accessed: 2021-12-30 from https://osf.io/j7ghv/

[6] American Economic Association. 2012. *American Economic Association's registry for randomized controlled trials.* Retrieved Accessed: 2021-09-08 from https://www.socialscienceregistry.org

[7] Nick Ballou, Vivek R. Warriar, and Sebastian Deterding. 2021. Are You Open? A Content Analysis of Transparency and Openness Guidelines in HCI Journals. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 176, 10 pages. https://doi.org/10.1145/3411764.3445584

[8] Lonni Besançon and Pierre Dragicevic. 2019. The continued prevalence of dichotomous inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–11.

[9] Graeme Blair, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. Declaring and diagnosing research designs. *American Political Science Review* 113, 3 (2019), 838–859.

[10] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems.* 981–992.

[11] Paul Cairns. 2007. HCI... not as it should be: inferential statistics in HCI research. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21.* 1–7.

[12] Christopher D Chambers, Eva Feredoes, Suresh Daniel Muthukumaraswamy, and Peter Etchells. 2014. Instead of" playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience* 1, 1 (2014), 4–17.

[13] Aline Claesen, Sara Lucia Brazuna Tavares Gomes, et al. 2019. Preregistration: Comparing dream to reality. (2019).

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

[15] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. *HARK No More: On the Preregistration of CHI Experiments.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173715

[16] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences.* Academic press.

[17] The 2020 Online Mining Software Repositories Conference. [n.d.]. *Registered Reports (MSR 2020).* Retrieved Accessed: 2021-09-09 from https://2020.msrconf.org/track/msr-2020-Registered-Reports?

[18] US Congress. [n.d.]. Food and Drug Administration Amendments Act of 2007 (FDAAA).

[19] Rik Crutzen and Gjalt-Jorn Y Peters. 2017. Targeting Next Generations to Change the Common Practice of Underpowered Research. *Frontiers in psychology* 8 (2017), 1184.

[20] Alexander DeHaven. 2017. Preregistration: A plan, not a prison. *Retrieved October* 29 (2017), 2019.

[21] Ulrich Dirnagl. 2020. Preregistration of exploratory research: Learning from the golden age of discovery. *PLoS biology* 18, 3 (2020), e3000690.

[22] Alan Dix. 2020. Statistics for HCI: Making Sense of Quantitative Data. *Synthesis Lectures on Human-Centered Informatics* 13, 2 (2020), 1–181.

[23] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *MModern Statistical Methods for HCI.* Springer, 291–330.

[24] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–15.

[25] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2019. Touchstone2: An Interactive Environment for Exploring Trade-Offs in HCI Experiment Design *(CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300447

[26] Association for Computing Machinery (ACM). [n.d.]. *Artifact Review and Badging.* Retrieved Accessed: 2021-09-08 from https://www.acm.org/publications/policies/artifact-review-and-badging-current

[27] Center for Open Science. 2011. *Open Science Badges enhance openness, a core value of scientific practice.* Retrieved Accessed: 2021-09-08 from https://www.cos.io/initiatives/badges?_ga=2.74055258.93961764.1630280699-1403603742.1629748234

[28] Center for Open Science. 2011. *Open Science Framework.* Retrieved Accessed: 2021-09-08 from https://osf.io

[29] Erin D Foster and Ariel Deardorff. 2017. Open science framework (OSF). *Journal of the Medical Library Association: JMLA* 105, 2 (2017), 203.

[30] Andrew Gelman and Eric Loken. 2013. 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. *Department of Statistics, Columbia University* 348 (2013).

[31] Paul Glasziou, Douglas G Altman, Patrick Bossuyt, Isabelle Boutron, Mike Clarke, Steven Julious, Susan Michie, David Moher, and Elizabeth Wager. 2014. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet* 383, 9913 (2014), 267–276.

[32] Miguel Grinberg. 2018. *Flask web development: developing web applications with python.* " O'Reilly Media, Inc.".

[33] François Guimbretière, Morgan Dixon, and Ken Hinckley. 2007. ExperiScope: an analysis tool for interaction data. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* 1333–1342.

[34] Tamarinde L Haven, Timothy M Errington, Kristian Skrede Gleditsch, Leonie van Grootel, Alan M Jacobs, Florian G Kern, Rafael Piñeiro, Fernando Rosenblatt, and Lidwine B Mokkink. 2020. Preregistering qualitative research: a Delphi study. *International Journal of Qualitative Methods* 19 (2020), 1609406920976417.

[35] Transparent Statistics in Human–Computer Interaction Working Group. 2019. Transparent Statistics Guidelines. https://doi.org/10.5281/zenodo.1186169 (Available at https://transparentstats.github.io/guidelines).

[36] Berkeley Institute of Governmental Studies at the University of California. 2020. *Evidence in Governance and Politics.* Retrieved Accessed: 2021-09-08 from https://egap.org/

[37] John PA Ioannidis. 2005. Why Most Published Research Findings Are False. *PLoS medicine* 2, 8 (2005), e124.

[38] Alan M Jacobs, Tim Büthe, Ana Arjona, Leonardo R Arriola, Eva Bellin, Andrew Bennett, Lisa Björkman, Erik Bleich, Zachary Elkins, Tasha Fairfield, et al. 2021. The Qualitative transparency deliberations: Insights and implications. *Perspectives on Politics* 19, 1 (2021), 171–208.

[39] Leslie K John, George Loewenstein, and Drazen Prelec. 2012. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological science* 23, 5 (2012), 524–532.

[40] Eunice Jun, Melissa Birchfield, Nicole De Moura, Jeffrey Heer, and René Just. 2022. Hypothesis Formalization: Empirical Findings, Software Limitations, and Design Implications. *ACM Trans. Comput.-Hum. Interact.* 29, 1, Article 6 (jan 2022), 28 pages. https://doi.org/10.1145/3476980

[41] Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke. 2019. Tea: A high-level language and runtime system for automating statistical analysis. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology.* 591–603.

[42] Robert M Kaplan and Veronica L Irvin. 2015. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PloS one* 10, 8 (2015), e0132382.

[43] Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2016. Special interest group on transparent statistics in hci. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* 1081–1084.

[44] Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, and Chat Wacharamanotham. 2017. Moving transparent statistics forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* 534–541.

[45] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing*

Systems. 4521–4532.

[46] Norbert L Kerr. 1998. HARKing: Hypothesizing After the Results are Known. *Personality and social psychology review* 2, 3 (1998), 196–217.

[47] Mary Beth Kery, Bonnie E John, Patrick O'Flaherty, Amber Horvath, and Brad A Myers. 2019. Towards Effective Foraging by Data Scientists to Find Past Analysis Choices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[48] Goran Knežević, Ljiljana B Lazarevic, and Michael Bosnjak. [n.d.]. As Predicted PreRegistration. ([n. d.]).

[49] Tamarinde L. Haven and Dr Leonie Van Grootel. 2019. Preregistering qualitative research. *Accountability in Research* 26, 3 (2019), 229–244.

[50] David D Laitin and Rob Reich. 2017. Trust, Transparency, and Replication in Political Science. *PS: Political Science & Politics* 50, 1 (2017), 172–175.

[51] David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation strategies for HCI toolkit research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–17.

[52] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[53] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A. Myers. 2019. Unakite: Scaffolding Developers' Decision-Making Using the Web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 67–80. https://doi.org/10.1145/3332165.3347908

[54] Yang Liu, Tim Althoff, and Jeffrey Heer. 2020. 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*. 1–14.

[55] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. 2020. Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1753–1763.

[56] María Jesús Lobo, Christophe Hurter, and Pourang Irani. 2020. Flex-ER: A Platform to Evaluate Interaction Techniques for Immersive Visualizations. *Proceedings of the ACM on Human-Computer Interaction* 4, ISS (2020), 1–20.

[57] Wendy E Mackay, Caroline Appert, Michel Beaudouin-Lafon, Olivier Chapuis, Yangzhou Du, Jean-Daniel Fekete, and Yves Guiard. 2007. Touchstone: exploratory design of experiments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1425–1434.

[58] Xiaojun Meng, Pin Sym Foong, Simon Tangi Perrault, and Shengdong Zhao. 2017. NexP: A Beginner Friendly Toolkit for Designing and Conducting Controlled Experiments. In *INTERACT*.

[59] James E Monogan. 2015. Research Preregistration in Political Science: The Case, Counterarguments, and a Response to Critiques. *PS: Political Science & Politics* 48, 3 (2015), 425–429.

[60] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour* 1, 1 (2017), 1–9.

[61] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. 2015. Promoting an open research culture. *Science* 348, 6242 (2015), 1422–1425.

[62] Brian A Nosek, Emorie D Beck, Lorne Campbell, Jessica K Flake, Tom E Hardwicke, David T Mellor, Anna E van't Veer, and Simine Vazire. 2019. Preregistration is hard, and worthwhile. *Trends in cognitive sciences* 23, 10 (2019), 815–818.

[63] Klaus Oberauer and Stephan Lewandowsky. 2019. Addressing the theory crisis in psychology. *Psychonomic bulletin & review* 26, 5 (2019), 1596–1618.

[64] U.S. National Library of Medicine. [n.d.]. *ClinicalTrials.gov*. Retrieved September 8, 2021 from https://clinicaltrials.gov

[65] Wharton School of the University of Pennsylvania. 2017. *AsPredicted*. Retrieved Accessed: 2021-09-08 from https://aspredicted.org

[66] Benjamin A Olken. 2015. Promises and Perils of Pre-analysis Plans. *Journal of Economic Perspectives* 29, 3 (2015), 61–80.

[67] Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2017. *Gut Instinct: Creating Scientific Theories with Online Learners*. Association for Computing Machinery, New York, NY, USA, 6825–6836. https://doi.org/10.1145/3025453.3025769

[68] Vineet Pandey, Krzysztof Z Gajos, and Anoopum S Gupta. 2020. From novices to co-pilots: Fixing the limits on scientific knowledge production by accessing or building expertise. In *Proceedings of the 7th International Conference on ICT for Sustainability*. 294–304.

[69] Michel Tuan Pham and Travis Tae Oh. 2021. Preregistration Is Neither Sufficient nor Necessary for Good Science. *Journal of Consumer Psychology* 31, 1 (2021), 163–176.

[70] Xiaoying Pu, Licheng Zhu, Matthew Kay, and Frederick Conrad. 2019. Designing for preregistration: A user-centered perspective. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[71] Ulf-Dietrich Reips and Christoph Neuhaus. 2002. WEXTOR: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers* 34, 2 (2002), 234–240.

[72] Wendy Roldan, Ziyue Li, Xin Gao, Sarah Kay Strickler, Allison Marie Hishikawa, Jon E. Froehlich, and Jason Yip. 2021. *Pedagogical Strategies for Reflection in Project-Based HCI Education with End Users*. Association for Computing Machinery, New York, NY, USA, 1846–1860. https://doi.org/10.1145/3461778.3462113

[73] Robert Rosenthal. 1979. The file drawer problem and tolerance for null results. *Psychological bulletin* 86, 3 (1979), 638.

[74] Jacob O. Wobbrock Scott Klemmer. [n.d.]. *Designing, Running, and Analyzing Experiments*. Retrieved Accessed: 2021-09-08 from https://www.coursera.org/learn/designexperiments

[75] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22, 11 (2011), 1359–1366. https://doi.org/10.1177/0956797611417632 arXiv:https://doi.org/10.1177/0956797611417632 PMID: 22006061.

[76] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2021. Pre-registration is a Game Changer. But, Like Random Assignment, it is Neither Necessary Nor Sufficient for Credible Science. *Journal of Consumer Psychology* 31, 1 (2021), 177–180.

[77] Elliot Soloway, Mark Guzdial, and Kenneth E. Hay. 1994. Learner-Centered Design: The Challenge for HCI in the 21st Century. *Interactions* 1, 2 (April 1994), 36–48. https://doi.org/10.1145/174809.174813

[78] Lisa Spitzer and Stefanie Mueller. 2021. Registered Report Protocol: Survey on attitudes and experiences regarding preregistration in psychological research. *PloS one* 16, 7 (2021), e0253950.

[79] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11, 5 (2016), 702–712.

[80] Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage.

[81] Denes Szucs and John Ioannidis. 2017. When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience* 11 (2017), 390.

[82] Allison A Toth, George C Banks, David Mellor, Ernest H O'Boyle, Ashleigh Dickson, Daniel J Davis, Alex DeHaven, Jaime Bochantin, and Jared Borns. 2021. Study Preregistration: An Evaluation of a Method for Transparent Reporting. *Journal of Business and Psychology* 36, 4 (2021), 553–571.

[83] Anna Elisabeth van't Veer and Roger Giner-Sorolla. 2016. Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology* 67 (2016), 2–12.

[84] Jan B. Vornhagen, April Tyack, and Elisa D. Mekler. 2020. *Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency*. Association for Computing Machinery, New York, NY, USA, 4–18. https://doi.org/10.1145/3410404.3414229

[85] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[86] Chat Wacharamanotham, Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2018. Special Interest Group on Transparent Statistics Guidelines. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3170427.3185374

[87] Chat Wacharamanotham, Krishna Subramanian, Sarah Theres Völkel, and Jan Borchers. 2015. Statsplorer: Guiding novices in statistical analysis. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2693–2702.

[88] Nai-Ching Wang, David Hicks, and Kurt Luther. 2018. Exploring Trade-Offs Between Learning and Productivity in Crowdsourced History. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 178 (Nov. 2018), 24 pages. https://doi.org/10.1145/3274447

[89] Xiaoyi Wang, Alexander Eiselmayer, Wendy E Mackay, Kasper Hornbæk, and Chat Wacharamanotham. 2020. Argus: Interactive a priori Power Analysis. *IEEE Transactions on Visualization and Computer Graphics* (2020).

[90] Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology* 7 (2016), 1832.

[91] Leland Wilkinson. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American psychologist* 54, 8 (1999), 594.

[92] Jacob O Wobbrock. 2015. Catchy Titles Are Good: But Avoid Being Cute. (2015).

[93] David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry* 17, 2 (1976), 89–100.

[94] Yuki Yamada. 2018. How to Crack Pre-registration: Toward Transparent and Open Science. *Frontiers in psychology* 9 (2018), 1831.