HARK No More: On the Preregistration of CHI Experiments

Andy Cockburn
University of Canterbury
Christchurch, New Zealand
andy@cosc.canterbury.ac.nz

Carl Gutwin
University of Saskatchewan
Saskatoon, Canada
gutwin@cs.usask.ca

Alan Dix
University of Birmingham
Birmingham, UK
alan@hcibook.com

ABSTRACT
Experimental preregistration is required for publication in many scientific disciplines and venues. When experimental intentions are preregistered, reviewers and readers can be confident that experimental evidence in support of reported hypotheses is not the result of HARKing, which stands for Hypothesising After the Results are Known. We review the motivation and outcomes of experimental preregistration across a variety of disciplines, as well as previous work commenting on the role of evaluation in HCI research. We then discuss how experimental preregistration could be adapted to the distinctive characteristics of Human-Computer Interaction empirical research, to the betterment of the discipline.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
Experimental preregistration; HARKing, p-fishing; NHST controversy; file drawer effect; replication.

INTRODUCTION
Researchers in HCI use a wide variety of evaluation methods, including qualitative and quantitative techniques, in field and laboratory settings, and taking objective and subjective measurements or observations. For those researchers who use quantitative methods and statistical techniques to assess their data, a primary way in which research is carried out is null-hypothesis significance testing (NHST). When using NHST, an assumption of no difference is rejected when the probability (p) of encountering data as extreme as that observed falls below a threshold value (the α level, normally .05). The intention is to draw reliable and repeatable inferences about how a population of users would perform with the evaluated interface(s), based on the performance of the sample.

Null-hypothesis significance testing has been a key component of the scientific method for many decades. However, criticisms of NHST have been raised for nearly sixty years [37]. While the range of NHST criticisms is broad, this paper primarily focuses on issues arising from *publication bias* – the tendency for papers that reject the null hypothesis to be accepted at a much higher rate than those that do not [11, 36].

Publication bias creates incentives for researchers to ensure that their NHST studies reject the null hypothesis, which can motivate a variety of imperfect scientific practices, including various forms of 'HARKing' [28]. HARKing stands for Hypothesising After the Results are Known, and it has various other names including 'p-fishing', 'p-hacking', 'outcome switching', and 'experimenter degrees of freedom'. HARKing in NHST studies is troublesome because the malleability to post-hoc 'fish' for significant effects inflates the risks of falsely identifying an effect and labelling it as significant. A survey of over 2000 psychology researchers indicates that HARKing is disturbingly prevalent [22]. Publication bias and HARKing are not limited to NHST studies, but they are predominantly raised with respect to NHST due to its widespread use.

Publication bias also causes a 'file drawer' effect [35, 14], in which studies that do not reach the threshold α level are rejected for publication (or are not submitted). Their findings are therefore hidden from the research community. Similar studies may be repeated multiple times, producing findings that also go unpublished. However, a single study of the same effect that observes a significant outcome is much more likely published and could become part of a discipline's 'knowledge'. Yet at $\alpha = .05$, we should expect one study in twenty to falsely reject the null hypothesis. In other words, publication bias and the file drawer effect can combine to propagate the dissemination of studies falsely claiming a difference, while suppressing those studies that correctly find null results.

One further outcome of publication bias is that it can influence the types of research conducted, potentially deterring researchers from asking important but risky questions in favour of less important work that is more likely to yield statistical significance. The shape and thrust of entire disciplines can be influenced by the undesirable implications of publication bias.

These issues are not new. They have been comprehensively reported across many disciplines, including medicine [21], psychology [22], political science [20], biology [13], and general science [14]. They have also been identified and discussed within HCI (e.g., [12]). However, one point of difference between HCI and many other disciplines, is that

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

 ISBN 978-1-4503-5620-6/18/04…$15.00
DOI: https://doi.org/10.1145/3173574.3173715

other disciplines have introduced means for preregistration of experiments, which alleviates many of the problems.

Through preregistration, researchers record an ultimately public statement of their study's objectives, hypotheses, and methods before conducting their experiment. Preregistration offers many substantial advantages for researchers and the scientific community. Diligent and honest scientists who's experiments reject the null hypothesis are relieved of many forms of potential criticism – they can demonstrate, through reference to their registration, that their experimental analysis is free from HARKing and selective reporting. Similarly, the discipline as a whole gains assurance in its work because reviewers and readers gain increased trust in the veracity of papers' claims. Furthermore, if the study registration process is coupled with mechanisms for storing and retrieving subsequent results files and analysis scripts, then substantial new opportunities for study, replication, and meta-analysis are enabled.

Regardless of the potential benefits of preregistration, experimental research in HCI has distinct features that could influence the practicality and value of preregistration. In particular, user interface design and evaluation is conducted in a highly iterative manner, often with exploratory objectives, and if poorly implemented preregistration could impede valid and desirable HCI practices. This paper therefore presents a review of considerations associated with experimental preregistration in HCI. Background on related issues within and outside HCI are presented first. We then examine the challenges, potential benefits and costs associated with the use of preregistration within HCI. We finish by examining ways in which preregistration could improve support for HCI replication and meta-analysis.

BACKGROUND

To establish foundations for the upcoming discussion of experimental preregistration, this section provides a brief and non-technical summary of inferential statistics and NHST before reviewing the wide-ranging criticisms of NHST from disciplines other than HCI. We then review literature that has raised issues relating to the conduct of experiments and data analysis within HCI. While the majority of the review concerns NHST, this emphasis stems from its prevalence in HCI research; many of the issues raised are not limited to NHST.

Brief Summary of Inferential Statistics and NHST

One of the main reasons for conducting statistical analyses in HCI is to draw inferences about how a population of users would interact with a particular form of user interface, compared to the existing state of the art. Such analyses ask whether the population of potential users would be faster, more accurate, or prefer the new interface (or other measures) if it were broadly disseminated. Of course, measuring the entire population is impossible, so instead samples are measured and statistics are used to make inferences about the population. Inferential statistical analysis allows us to gain confidence that findings are not due to chance, indicating that the inferences drawn are rigorous and repeatable. Ideally, the inferences are both internally and externally valid, where internal validity means that the findings are due to the experimental manipulations conducted, rather than other extraneous factors. Internal validity also means that a replication of the study is likely to reproduce the same findings, with the replication of findings being a cornerstone of science. External validity means that experimental results are generalisable beyond tested settings – for example, if externally valid, a lab study showing 15% performance improvements with a new mouse would also mean 15% improvements when using the same mouse in practice.

When using NHST, a *null hypothesis* of no effect (i.e., no difference in mean values or no association between conditions) is tested through a family of statistical techniques. Sample data is collected from a set of conditions, and the statistical tests determine a p value, which represents the likelihood that data at least as extreme as the sample would be observed if the null hypothesis were true. The tests are intentionally conservative, with the null hypothesis only rejected when the probability of observing this data is sufficiently low, suggesting that the assumption of no difference is unlikely to be correct. To avoid malleable argumentation over what constitutes 'sufficient' rarity, the threshold for rejecting the null hypothesis, called the α level, is established beforehand, and is typically 5% (or .05).

NHST therefore results in a dichotomous outcome for each test: the null hypothesis is either rejected or not rejected. If rejected, researchers infer that a difference is 'real', accepting an alternative hypothesis, such as 'the interface is faster'. If not rejected, meaningful interpretation is difficult – perhaps there is no difference between the conditions, or perhaps the experiment had insufficient sensitivity to expose the difference. Both dichotomous outcomes may be made in error. A 'Type I error' occurs if the null hypothesis is incorrectly rejected – falsely inferring that a difference exists. A 'Type II error' occurs when an experiment falsely fails to reject a null hypothesis. This paper deals primarily with Type I errors.

Brief Review of Criticism of NHST Outside HCI

Problems of statistical misuse have long been reported in the general scientific literature. For example, in 1949 Lewis and Burke [29] analysed papers published in the Journal of Experimental Psychology between 1944 and 1946. Nine of the 14 papers that used the chi-square test applied it in a manner that was 'clearly unwarranted', and another one provided insufficient details to determine validity.

More broadly, in a 1959 paper, Sterling [37] raised three important assumptions regarding publication decisions and NHST: $A1$ – results are more likely to be published if they reject the null hypothesis; $A2$ – replication is unlikely when a previous publication rejects the null hypothesis; $A3$ – a great many more studies are conducted than are published [37, p. 33]. Sterling supported his assumption $A1$ by analysing 362 papers published in major psychology journals between 1955 and 1956. Of the 294 papers that used NHST, 286 (97.3%) rejected the null hypothesis, while only 8 (2.7%) failed to do so. Furthermore, the total number of papers reporting a replication was zero. Although these problems led Sterling (nearly sixty years ago) to observe that 'such a field consists in substantial part of false conclusions', many disciplines including CHI appear to maintain similar publication biases. Although arguably an exaggeration, a recent comment by Forstmeier et al. [13, p. 1]

is likely to resonate with CHI authors – 'you can publish if you found a significant effect.'

Some of the implications of Sterling's assumptions were later named the *file drawer effect*: 'journals are filled with the 5% of studies that show Type I errors, while the drawers back at the lab are filled with the 95% that show nonsignificant (e.g., $p > .05$) results.' [35, p. 638]. Quantitative data demonstrating the prevalence of the file drawer effect has become available through the introduction of experimental registries (which are described later). For example, Franco's analysis of 221 time-sharing studies in the social sciences [14] showed several data points supporting the existence of the file drawer effect:

- studies with null results were unwritten (31 of 49, 63%) more frequently than those showing mixed (10 of 86, 11.6%) or strong effects (4 of 93, 4.5%);
- a further 7 of 49 papers (14.2%) papers showing null results were unpublished after being written up;
- studies with null results are much less likely to be published (10 of 49, 20.4%) than studies with mixed (40 of 86, 46.5%) or strong rejections of the null (56 of 93, 60.2%).

Authors across disciplines are aware of the influence that rejecting the null hypothesis has on the likelihood that their work will be accepted for publication, and this creates strong motivation for assuring that statistically significant findings are obtained. Consciously or subconsciously, scientists tend to steer their studies to successfully reject the null hypothesis; yet any such 'steerage' is opposed to the key objectives of NHST. While many scientists might agree that *other* scientists are susceptible to inappropriate experimental behaviour, evidence suggests that it is troublingly widespread. In a survey of over 2000 psychology researchers, John *et al.* [22] examined the prevalence of questionable experimental practices. Their survey questions serve as a useful classification of ten different forms of HARKing, as follows. The percentage values show the respondent's self-admission rates for the following questionable practices [22, Table 1]:

1. failing to report all dependent measures, which opens the door for selective reporting of favourable findings – 63.4%;
2. deciding to collect additional data after checking if the effects were significant – 55.9%;
3. failing to report all of the study's conditions – 27.7%;
4. stopping data collection early once the significant effect is found – 15.6%;
5. rounding off a p value (e.g., reporting $p = .05$ when the actual value is $p = .054$) – 22.0%;
6. selectively reporting studies that worked – 45.8%;
7. excluding data after looking at the impact of doing so – 38.2%;
8. reporting an unexpected finding as having been predicted – 27.0%;
9. reporting a lack of effect of demographic variables (e.g., gender) without checking – 3.0%;
10. falsifying data – 0.6%.

One recent proposal to ease concerns regarding excessive occurrence of Type I errors is to redefine the accepted norms for determining the dichotomous outcome of NHST tests. Instead of using $\alpha = .05$ as the default threshold for declaration of significance, the proposal is to use the much more conservative level of $\alpha = .005$ [23]. This recommendation is based on statistical analysis that equates classical hypothesis tests (p values) with evidence thresholds in Bayesian tests. While this proposal may seem shockingly stringent to HCI researchers who frequently examine effects at the margins of significance when $\alpha = .05$, this proposal recently gained substantial support from 75 senior co-authors across psychology, medicine, statistics, economics, political science, earth sciences, biology, and other disciplines [3]. This collection of authors additionally recommend that experiments yielding $0.005 < p < 0.05$ be labelled as 'suggestive'.

In summary, when commenting on the use of statistics in scientific writing, Cohen appealed for 'informed judgement from the investigator' [8, p. 1304], noting that the 'prevailing yes-no decision at the magic .05 level from a single research is a far cry from the use of informed judgement' [8, p. 1311].

Concerns About Experimental Work Within HCI

Like other disciplines, researchers in Human-Computer Interaction have raised concerns about experimental work and the importance and validity of findings generated. The following subsections highlight some of these issues, broadly categorised across concerns regarding experimental design, the types of data reported, the possibility of using Bayesian approaches for data analysis, and the need for replication.

Experimental design: do the right study and analyse correctly

In 2002, Lieberman wrote a self-proclaimed 'rant' on the 'Tyranny of Evaluation' [30] in which he argued that there existed an over-reliance on 'evaluation standards' in CHI (ostensibly NHST) that was stifling innovation in the field – 'papers that present innovative user interfaces but lack airtight evaluations are being rejected... what tends to get accepted is carefully crafted studies of uninteresting questions.' Conversations at CHI and its PC meetings suggests that similar sentiments persist, sixteen years after Lierberman's 'rant'. His final comments strongly reflect those of Cohen's above – 'We need more judgment in evaluation of user interfaces, not just more calculation.'

Zhai [40] responded to Lieberman's essay, arguing that although imperfect, evaluation remains the best method available to HCI researchers for drawing sound conclusions about interactive systems. Without it, he argues, HCI would turn into a 'faith-based enterprise' in which we "simply accept the inventors' and designers' claims." Zhai's essay also provides recommendations for the conduct of experimental studies that avoid poor practices. In particular, he notes that good studies have well founded motivation and hypotheses (far beyond a 'mindless A-B comparison'), with cautious generalistion of results, and they reveal the role of relevant secondary factors.

Greenberg and Buxton [16] further commented on the practice and role of evaluation in HCI, titling their paper 'Usability Evaluation Considered Harmful (Some of the Time)'. Their

topic is broader than that of Lieberman and Zhai in that it concerns all forms of evaluation – as taught in education, practiced in industry, and deployed by researchers. Given the wide range of objectives, it is perhaps unsurprising that their main message is a general call that 'the choice of evaluation methodology – if any – must arise from and be appropriate for the actual problem or research question under consideration'.

Cairns [6] surveyed the use of inferential statistics in HCI research, with two main observations – approximately half of the sampled papers used statistics, and all but one of these papers had some form of problem in their use. The survey included a total of 80 papers published in the 2005/6 Proceedings of the British HCI Conference (51 papers), and in the 2006 volumes of the Human-Computer Interaction Journal (10) and ACM Transactions on CHI (19), with 41 making some use of inferential statistics. Statistical problems were categorised into four types, and a summary table showed the number of papers containing at least one problem of each type – reporting (25), checking assumptions (16), over-testing (15), and inappropriate tests (12). Cairns' survey finishes with useful recommendations for those using inferential statistics in HCI, including the need to improve statistical reporting by following standards such as APA recommendations [1], the creation of data repositories associated with papers for improved checking, and the need for better statistical education.

Favour estimation, information, and Bayesian approaches

Dragicevic [12] provides a broad-ranging and insightful review of problems with current statistical reporting in HCI, together with a clear and useful set of recommendations for reporting empirical results in an 'accurate and transparent way without using any tests or p values.' These recommendations and sentiments align with those of several outside HCI [8, 9, 21]. Dragicevic's paper culminates in a set of 34 tips for conducting experiments, including the forceful 'Tip 25: Ban dichotomous interpretations', reflecting the decision of the editorial board of Basic and Applied Social Psychology to ban NHST from their journal [38]. None of these tips explicitly advocate experimental preregistration (our focus), although Tip 5 'Plan all analyses using pilot data' approaches the idea, noting that planned analyses are 'more convincing than post-hoc analyses because they leave less room for self-deception and prevent questionable practices such as "cherry picking"' [12, p. 23]. Related observations on 'cherry picking' were made at a CHI 2016 workshop [26]. However, without preregistration, it is still possible for experimenters to consciously or unconsciously engage in HARKing, and to reframe the experimental intention to match the outcome.

Other researchers, including those in HCI, propose replacing NHST with Bayesian statistical methods. One of the key motivators for doing so concerns a common misapprehension, yet desired interpretation, of the p value in NHST. Researchers wish to understand the probability that the null hypothesis is true, given the data observed ($P(H_0|D)$), and p is often misunderstood to actually represent this value. However, as stated above, the p value actually represents the probability of observing data at least as extreme as the sample, given that the null hypothesis is true: $P(D|H_0)$. In contrast to NHST, Bayesian statistics (in certain contexts) enable the desired computation of $P(H_0|D)$. A full discussion of Bayesian methods is beyond the scope of this paper, and interested readers are directed to Kay [27] and Masson [32] for brief introductions to Bayesian methods, and to Howard et al. [18] for a review of the relative strengths of NHST, Meta-Analysis and Bayesian analysis.

Replicate

Experimental replication and refutation is another important aspect of the scientific method. However, as discussed above, various forms of publication bias (e.g., 'you can publish if you found a significant effect' [13]) create a disincentive for conducting replications – if significant findings are replicated, then authors' work is likely to be rejected for lack of novelty because they 'merely' replicate previous findings; and if the results are not replicated (producing null results instead) then the paper is likely to be rejected for its lack of significance, with the study being confined to the burgeoning 'file drawer'.

Many researchers have argued that more needs to be done to encourage replication, including access to null findings. Hornbæk et al. [17] recently analysed 891 papers, finding that only 3% attempted some form of replication, and that many of these were accidental replications, where the study was not initially planned as such. Recommendations for encouraging replications are presented, including new practices by publication venues, but none of the recommendations include experimental preregistration. Banovic [2] also called for increased replication in CHI, with an emphasis on understanding and use of mobile technologies. He notes the importance of increasing accessibility to the findings of replicated studies, but opportunities for linking this with experimental preregistration are not noted. A series of CHI workshops on 'RepliCHI' provided a good forum for promoting the important issues related to replication of experimental work on HCI (e.g., [39]).

Finally, HCI researchers have proposed the introduction of experimental platforms, such as Touchstone [31], to assist in the design, conduct, analysis, and replication of experiments. The utilities provided by these platforms contribute to the increasing scientific maturity of HCI, but they address different issues to those of experimental preregistration.

EXPERIMENTAL PREREGISTRATION (OUTSIDE HCI)

Correct scientific inferences regarding the outcome of NHST are critically important in clinical trials – a Type I error could result in millions of people taking a new medication, exposing themselves to risks of potential side-effects, due to falsely identified medical benefits. In response to decades of concerns similar to those above, various authorities have instituted registries in which researchers preregister their intention, methods, and hypotheses for upcoming experiments. By doing so, risks of HARKing are substantially reduced, and the file drawer problem is eased (assuming the outcome of all experiments are entered into the registry). In 1997, the US Food and Drug Administration Modernization Act (FDAMA) established the registry ClinicalTrials.gov, although with some resistance from pharmaceutical companies, which established their own registries [24]. Within its first ten years over 96,000 experiments were registered on ClinicalTrials.gov, with adoption assisted by the decision of the International Committee of

Medical Journal Editors to make preregistration a requirement for publication in their journals [10].

Long-term analysis of studies suggests that registries have been successful. In particular, Kaplan and Irvin [25] examined all large studies (direct costs >$500,000 per year) funded by the National Heart, Lung, and Blood Institute (NHLIBI) between 1970 and 2012, giving 55 studies in total. They then examined whether or not those studies showed a significant effect (rejecting the null hypothesis). The majority (17 of 30, 57%) of studies published prior to 2000 showed a significant benefit of the intervention, while only 2 of 25 (8%) published after 2000 did so. The authors attributed this dramatic change in the rate of null findings to the introduction of mandatory preregistration on ClinicalTrials.gov in 2000.

The successes of clinical trial registries has led to other disciplines introducing their own, including the American Economic Association (https://www.socialscienceregistry.org/) and the political science 'dataverse' [34], hosted by Harvard at https://dataverse.harvard.edu/dataverse/registration. Many journals are also offering or developing support for preregistration, with several summarised at https://tinyurl.com/yaxt6ott. The Open Science Framework (OSF) also supports preregistration, ranging from simple and brief descriptions through to complete specification of all experimental aspects: http://osf.io, which includes a preregistration template at https://osf.io/t6m9v. Another general registry is available at: https://aspredicted.org/

EXPERIMENTAL PREREGISTRATION WITHIN HCI

Despite HCI's heavy reliance on evaluation, preregistration has received little to no published attention in HCI. One viable explanation for this is that HCI experimental research has distinct characteristics that reduce the appeal of preregistration. This section considers first the distinctive aspects of HCI research, and second, the benefits of preregistration for HCI.

Distinctive characteristics of HCI experimental research

In clinical trials the dichotomous outcome of NHST is appealing – for instance, either there is significant evidence for the intended benefit of a new drug, or the trial is inconclusive. Similarly, when testing a new psychological theory, evidence is either compelling (i.e., significant) or not. In such experiments the conditions under study are exact and unchanging.

However, research in HCI almost always includes iteration, during which researchers explore design alternatives. They typically use parallel *elaborative design* to identify variant approaches to an interactive solution ('getting the right design') as well as using iterative *reductive design* to hone pursued variants ('getting the design right') [16]. Evaluations will typically be conducted throughout these processes, making quantitative and qualitative observations.

The need for iteration and exploration in HCI stems from a combination of factors. First, the discipline is relative young when compared to human factors or experimental psychology, reducing the opportunity for a comprehensive empirical foundation. Second, technology evolves rapidly, and lessons derived with one technology may not transfer to another – for example, over 30 years HCI has examined interaction with keyboards and command-based interfaces, WIMP interfaces, touchscreens, very small and very large displays, speech and eyes-free input, virtual and augmented reality, and so on; and all of these technologies introduce new factors for empirical analysis. Third, interface technologies are continually deployed in (or are used to establish) new contexts of use, such as hypertext, the world-wide-web, and social networking. The result is that HCI research has been, and is likely to remain, continually on the edge of uncharted territory. Consequently, HCI experimental work often lacks strict hypotheses, with researchers iteratively gaining new information about the technique, the task, its context, or the users.

The introduction of preregistration could therefore raise challenges for researchers in understanding at what point they should preregister their experiments – prior to initial formative studies, before initial pilot studies (where methods are likely to change), or before final studies (where methods are largely finalised, but still may change)? However, closely related problems already exist for HCI researchers (at least at some institutions[1]) due to the requirements of ethical review boards. Membership of these boards is often predominantly from the classical sciences, who expect detailed study descriptions. Yet the requirement for detail demands that HCI researchers either submit new ethics approval applications for each iteration (which could become over-burdensome) or that they write vague applications that encapsulate a broad set of potential iterative adaptations (risking refusal from the review board).

An HCI experimental registry could accommodate iterative design and exploratory empiricism by supporting researchers in explicitly stating that their objectives are exploratory (as further described later). Doing so could offer advantages, not least because researchers engaged in exploratory studies would be empowered to defend against any claim during review that the outcomes of their experiment actually represent a failed NHST experiment. However, it is important to distinguish between two variants of HCI empirical exploration.

Appropriate Exploration. Intentionally exploratory studies are a cornerstone of HCI. For such studies, changing interface designs, variables, methods, and procedures mid-study is entirely appropriate. When an exploratory study is intended, researchers could explicitly register their exploratory intention, including any initial hypotheses. After the study, researchers would upload to the registry at least a brief description of the outcome of their studies, including a description and explanation of modifications to systems, hypotheses, and methods. Importantly, such studies should not be used as the basis for NHST dichotomous testing, although they could be used as pilot studies for purposes such as statistical power analysis.

Inappropriate Explorative NHST HARKing. Conducting a series of exploratory studies, subjecting the data to *p*-fishing, and then presenting the work as if a significant outcome was the only hypothesis is inappropriate HARKing. Such posthoc selective reporting largely invalidates the use of NHST as evidentiary criteria due to inflation of Type I errors.

[1] Saul Greenberg, personal communication.

Benefits of preregistration within HCI

This section reviews the potential benefits of preregistration for HCI, together with corollaries of inhibited weak science.

Promote theory and ease NHST for barrage analysis

NHST is best used when testing *risky* and *specific* hypotheses that are founded on a *generalisable theory* of interaction. By *specific*, we mean that the set of predictions is small – only a few statistical tests will be made (reducing the likelihood of Type I errors). In contrast to specificity, many experiments in HCI include a swathe of data points all of which are subject to NHST, often without associated correction for the multitude of comparisons conducted (such as Bonferroni). In such barrage analyses, the likelihood of Type I errors are extremely high. Note, that we are not criticising experiments that sample a wide range of different aspects of interaction (rather, we applaud them). Instead, we are arguing that NHST is the wrong tool for analysing a swathe of different measures – instead, when analysing extensive datasets researchers might consider the *estimation* approach advocated by Dragicevic [12].

The term *risky* means that the experiment involves a genuine attempt to refute the hypothesis. In HCI this might involve seeking out conditions in which an interface fails to provide anticipated benefits, or preferably it might empirically validate a model that exposes how an interface's performance transitions from merits to detriments in comparison with some baseline interface when a secondary parameter is manipulated. As Buxton famously notes 'Everything is best for something and worst for something else' https://www.billbuxton.com. In contrast to risky experiments, some experiments in HCI can be considered to be 'existence proofs' [16], which seek to demonstrate the existence of a condition in which a novel interface outperforms its competitors; however, the scientific value of existence proof experiments is limited unless the condition in question represents an important domain of interaction. Greenberg and Buxton characterised existence proof empiricism as being 'at best, weak science' [16, p. 113].

Finally, the knowledge derived from an experiment (or set of experiments) is strongest when support for a hypothesis contributes to evidence for some theory of interaction that generalises beyond the specific experimental context (see [19] for a discussion of Interaction Science, which promotes theory-oriented research within HCI). Experimental preregistration encourages the explicit recording of hypotheses and *predicted* experimental outcomes. Forming and recording anticipatory outcomes is likely to encourage researchers to contemplate and state the rationale for their predictions, which is part of theory formation. Conversely, preregistration will appropriately discourage the use of NHST for barrage analyses. For example, preregistering an intention to concurrently analyse dozens (say) of data points should be a red-flag to the researcher due to the likelihood of Type I errors if corrections for the multitude of tests are not planned; and if appropriate corrections for the multitude of tests are planned, then Type II errors are near certain, again hopefully representing a red-flag to the researcher. As Cohen stated, "you're not likely to solve your multiple tests problem with the Bonferroni maneuver... In short, the results of this humongous study are a muddle" [8, p. 1304].

Unlike barrage analyses, for which NHST is almost certainly inappropriate, preregistration of an existence proof experiment should help researchers to convince their reviewers that the experiment was well founded. Researchers would state in advance why superior performance in the selected condition(s) is important, thereby relieving authors from accusations that results were 'cherry picked' from a broader study.

Promote exploratory research and ease NHST compulsion

Exploratory studies typically examine a range of issues concerning interaction with a novel technology. To reiterate, such studies are an essential part of HCI literature, often representing initial empirical analysis of a breakthrough technology.

Given the lack of specific hypotheses and the wide range of potential dependent measures, NHST is largely inappropriate for exploratory analyses. Like barrage analyses described above, Type I errors are likely when many effects are concurrently analysed – "The greater the number and the lesser the selection of tested relationships ..., the less likely the research findings are to be true" [21].

Unfortunately, however, the heavy reliance on NHST within HCI means that researchers may feel compelled to use NHST to analyse exploratory outcomes, even when they are sensibly reluctant to do so. This compulsion could be derived from their fear and expectation that reviewers will criticise a lack of significance testing because it fails to satisfy the 'you can publish if your results are significant' criteria [13, 30]. Furthermore, researchers' compulsion to use NHST is exacerbated by their *a-priori* knowledge that reviewers may suspect that the reported 'exploratory' study is actually a post-hoc reframing of an NHST experiment with null findings.

Experimental preregistration overcomes these problems. When an experiment is intended to be exploratory, researchers would preregister their exploratory intent, and they could also state their reasons for that intent (possibly also stating why NHST is inappropriate). Due to preregistration, reviewers can no longer suspect that the stated exploratory intent is actually a post-hoc reframing of a failed NHST experiment, which should ease authors' compulsion to use NHST.

In summary, reviewers should not expect exploratory studies to employ NHST, and researchers should not feel compelled to do so. Over time, the introduction of preregistration could reduce the standing of NHST within the field as the almost defacto standard method for data analysis, helping to achieve objectives expressed by Dragicevic [12].

Promote trust and ease scepticism

Many authors will have encountered reviews that make unfounded criticisms of their work, such as stating that a 'Strawman' condition was included in the study merely to assure that significant findings were produced, or casting doubt that the experiment was originally intended to investigate the stated hypotheses. Sometimes such criticisms are warranted and correct; but sometimes they are not. Similarly, many reviewers will have read papers in which they suspect that the stated experimental outcomes differ from those that the authors originally intended. Sometimes reviewers will state their doubts

and adjust their review score accordingly, reducing the likelihood that the paper will be published. In other cases, reviewers will give authors the benefit of the doubt.

In essence, these problems in authorship and review stem from suspicion that authors may have engaged in different forms of HARKing. Experimental preregistration largely prevents authors from doing so, removing the exaggerated role that trust, faith and scepticism currently play in our review processes.

Promote risky research and ease NHST-induced increments

A perennial criticism of HCI research is that much of the work is incremental, potentially stifling space for more risky and creative research into novel interactions, new application areas, and interesting new interaction domains. However, as noted in Meyer's provocatively titled article 'Incremental research vs. paradigm-shift mania' [33], scientific advancements are almost invariably incremental. In HCI, the word 'incremental' often seems to be applied as a criticism for research that is shaped more by the NHST method than by the intended improvement in interactive technology.

Studies using NHST are well represented in the accepted CHI literature, and it is unsurprising if graduate students, young faculty, and others are drawn towards the methods that are achieving acceptance. Consequently, the high proportion of published papers that use NHST is likely to compel researchers to follow suit; and the more that do so, the more reviewers will expect to find the ubiquitous p, and criticise its absence. This has similarities to Goodhart's law, which was originally associated with economic measures [15] and recently applied to citation metrics [4] – when a measure (or method in our case) becomes a target, it ceases to be a good measure because people start to game it. The past successes of NHST at CHI may be contributing to a situation where research projects are selected based on their suitability for NHST. This concern is echoed by several previous comments on the state of HCI research [12, 30, 16].

There is therefore a real risk that a proportion of CHI publications are shaped by the method (NHST), rather than a more sensible situation where interesting research questions are selected first, and an appropriate method for its analysis is selected second. Interesting research questions that are not readily amenable to NHST may be less attractive to researchers (particularly young ones), potentially stifling the field.

Experimental preregistration should ease the compulsion to study safe increments with NHST, not least due to an expected increase in studies finding null outcomes (the next heading).

Promote access to null outcomes and open the file drawer

Based on the experiences of other disciplines, following the introduction of experimental preregistration we should anticipate a substantial reduction in the proportion of HCI studies rejecting the null hypothesis – Kaplan and Irvin's [25] study of clinical trials showed a reduction from 57% to 8%. A corollary to the reduced proportion of studies rejecting the null hypothesis, is that an increased proportion will find a null outcome, with several likely consequences. First, with more studies reporting null outcomes, we might expect reviewers and program committees to focus more on the nature of the contribution (for example, is the proposed interaction or method important, interesting, or intriguing, and are the results indicative of a successful method) and less on the attainment of a critical value of p. Second, with a higher proportion of studies reporting null outcomes, it is likely that more studies reporting null outcomes will be published, which should serve to weaken the file drawer effect. Third, over time reviewers' expectations for the rejection of the null hypothesis are likely to shift, easing expectations for both the *use* of NHST and for the rejection of the null. Any perceived enslavement to NHST methodology should be eased over time, allowing researchers to examine outcomes using the most appropriate methods, with reduced fear that reviewers will argue for rejection based on failing to attain statistical significance.

Promote replication, public data, and meta-analyses

The benefits of preregistration are arguably small where replications are common [7], but the lack of replication in HCI is a recognised problem [17, 39]. Many CHI researchers have sensibly called for more replication, but the impetus for researchers to do so is currently weak – a confirmatory replication is likely to be criticised as lacking novelty, and a failure to replicate is likely to be criticised for not satisfying the 'you can publish if your results are significant' criteria.

However, preregistration should shift the goalposts. Evidence from other disciplines suggests that the proportion of studies rejecting the null effect would decrease, which should lead to a lowering of reviewers' expectation for statistically significant findings. Also, the existence of a pool of studies on the registry showing null outcomes should increase the publication relevance and value replications, increasing possibilities for meta-analyses that are currently frustratingly difficult or impossible. We elaborate on these possibilities in the Discussion.

In short, experimental preregistration offers extensive opportunities to mature HCI research, including changes in the value systems and culture of academia, removing some of the tensions between career success and scientific reliability.

AN HCI REGISTRY: FEATURES AND USE

The attainment of preregistration benefits relies in part on the design of the registry and the features it supports. Some of the issues associated with their design and use in Economics are discussed by Coffman and Niederle [7], and Dickersin & Rennie [24] describe the evolution of registries and their use for clinical trials. This section briefly reviews issues in the design and use of experimental registries for HCI.

Desirable Features

Experimental preregistration is not intended to be a bureaucratic exercise. It should not substantially increase the workload for HCI researchers, however it is likely to shift the timing of some of that work. For example, rather than writing the Introduction to a study and its method *after* conducting the experiment, preregistration requires that these activities are conducted *beforehand*. Many researchers already advocate writing the motivation, method, and hypotheses before experimentation [12], and many institutional review boards will require the same, so the additional burden of preregistration need not be substantial.

The following paragraphs discuss desirable features of HCI experimental registries. Each of the features is prefixed by the word 'Require' (for data that must be provided for each registered experiment) or 'Support' (for data or features that researchers are likely to find valuable, but are not essential).

Require at least an a-priori narrative description

Early in the development of clinical trial registries, a working group proposed that registries should only require a small amount of information in order to reduce investigator burden [5]. While this may have helped adoption during the inception of the registries, experience has shown that the utility of registries increases with the specificity of the information recorded within them. Vague, incomplete, and imprecise experimental registration yields little value to the research community. It is therefore desirable that those registering their experiment should be as specific as possible regarding their intended outcomes and methods.

Research papers in HCI often begin with a description of the problem, its context, previous attempts to resolve the problem, and a description of a new approach to resolving it. This narrative establishes a foundation for assessing experimental outcomes, and it should be relatively straight forward for HCI researchers to enter this narrative into an experimental registry.

Require primary intended outcomes/hypotheses

In addition to the narrative, researchers should be required to identify primary anticipated outcomes (or hypotheses). Key dependent measures for these outcomes should be identified, together with evidentiary criteria used to assess the success/failure of the experiment (e.g., the expected analysis methods and tests whether NHST, Bayesian or other; NHST α level; criteria in an estimation-based approach). If the experiment is exploratory, without formal hypotheses, the experimenters should clearly state this fact.

Support full method, design, and analysis scripts

Beyond the narrative and primary outcomes, researchers should also be encouraged to add extensive details of the experimental method, design, and procedure. When data will be analysed using NHST, the more specificity the better, including details of methods for outlier identification, removal and treatment. Ideally, data analysis scripts should be uploaded (e.g., R scripts) together with templates showing the structure of experimental data files.

Institutional review boards (IRBs) for human ethics clearance often require much of the information associated with experimental intention and method, so the additional burden for researchers in uploading this information to a registry need not be substantial. Typically, the effortful components of experimental design are in conceiving the method and piloting alternatives; formally articulating the method for ethical review (or for uploading to a registry) should constitute only a relatively small additional burden.

When institutional review board application is required for the experiment, the registry should include information stating the date at which the ethical clearance application was submitted. Once clearance is granted, the details of the authorisation (the approval code and date) should be entered into the registry.

One of the reasons for including a dated link to IRB application is to deter post-hoc '*pre*'registration, whereby researchers first conduct the experiment, then formulate HARKed hypotheses, and finally register the experiment and associated results. While it may seem unlikely that researchers would do so, evidence from clinical trials suggests that related activities were not uncommon – Dickersin and Rennie [24] report that contrary to registry requirements, 52% of trials were registered *after* participant enrollment. While linking to IRB applications does not prohibit such practices, it escalates the risk of punitive measures for proceeding with an experiment without IRB clearance. A substantial difference between dates of IRB approval and experimental registration would warrant explanation on the registry.

Support a limited privacy window, with access key

Researchers may be concerned by the public availability of their experimental plans before publication. These concerns are likely to be particularly strong if the requirements for implementing the technology and method are low. The registry should therefore allow a limited time-period during which registry entries can be held private.

There are tradeoffs between researchers' potential desire for an extended privacy period and potential problems arising from a long-term 'closed file drawer'. For example, researchers intending to submit their work to CHI might originally request a privacy window that opens after the acceptance notification date; but if their work is rejected, they may want to extend the privacy period until after its acceptance to another venue. However, if authors are permitted to continually extend their privacy period (possibly forever, in order to bury an unsuccessful experiment), then the potential benefits of the registry's open file drawer are impaired – recall that a key motivation for preregistration is to promote the public record of null experimental outcomes, effectively opening the file drawer. We suggest a maximum privacy period of one year.

When creating a new entry on the registry researchers would be given an access key that permits review of the registered experiment during the privacy window. Authors submitting their preregistered work for review would include the experiment's URL and key in their submission, allowing reviewers to confirm the registration and the experiment's predicted outcomes. Separate keys would be required for researchers and reviewers to enable double-blind review procedures.

Support links to related studies

Publication venues such as CHI require authors to identify their closest related work for each submission. This requirement arose from concerns that excessively similar papers by the same authors were being concurrently submitted for consideration, contributing to an increased reviewer load and potential 'double dipping' in acceptance likelihood.

Related problems could hypothetically occur in an experimental registry. For example, a researcher might preregister an experiment predicting an experimental outcome of X, and concurrently preregister an experiment predicting *not* X. Asking authors to explicitly identify their closest related study should deter this type of activity by elevating the reputational risk

associated with failing to identify related work – the explicit request for a statement of their closest related studies reduces the viability of a researcher resorting to an argument that they were unaware of the importance of stating the connection.

On a more positive note, the explicit identification of related prior studies should facilitate subsequent researchers in conducting reviews and meta analyses.

Require results

As stated several times, one key reason for using experimental preregistration is to reduce the file drawer effect, increasing public access to studies with null findings. It is therefore important that researchers upload their experimental outcomes to the registry, regardless of whether predicted outcomes occur or not. One possibility for encouraging researchers to ensure that outcomes are eventually uploaded would be to have a ratio-score associated with each researcher on the registry – the score would show the ratio of registered experiments with outcomes to the total count of registered experiments (only counting those with expired privacy periods). However this could be easily circumvented by a researcher uploading dummy or nonsense results, so alternatively the registry might instead simply summarise each researcher's data, allowing users to draw their own conclusions based on the evidence of outcomes uploaded by the researcher.

Support explicit HARKing

Experiments play a pivotal role in theory formation, especially when the results are surprising or oppose hypotheses. The registry should support researchers in explaining why they believe their experiments produced the observed outcomes. If the experiment is preregistered as exploratory, the researchers' explanation could speculate about underlying causal factors, contributing to theory formation. If the experiment was preregistered for NHST, there are four broad categories of outcomes:

1. significant positive results, with the hypothesis supported;
2. negative results, with non-existent or unimportant effects;
3. inconclusive, e.g., insufficient power for confidence;
4. the experiment was not performed or abandoned.

With the existence of the registry, the value of outcomes (1), (2) and (3) are elevated: (1) readers know the supported hypothesis was not HARKed; (2) the file drawer is open on the negative outcome, reducing the likelihood of accidental replication of a null effect; (3) subsequent researchers have data from which to conduct a power analysis for estimating sample size for a replication. Additionally, a large set of similar experiments, each showing inconclusive results, might expose a reliable effect through meta-analysis (e.g., multiple student projects studying the same effect, each with small samples, all of which inconclusively expose the same trend).

Finally, if the experiment was abandoned (4), researchers should be able to explain why this occurred (e.g., the PhD student involved abandoned their studies).

DISCUSSION

Infrastructure: What can ACM do?

Several disciplines within Computer Science and Information Systems make extensive use of experimental methods, with HCI being among the largest. As the 'principal curator of publication data for the field'[2], ACM's Publications Board is committed to 'maintaining ACM as a brand of quality', with a goal of 'aggressively developing the highest-quality content'. In pursuing these goals, the publication board could enact a decision, like that by the International Committee of Medical Journal Editors [24], to *require* preregistration for experimental work published in ACM venues. Such a requirement might seem draconian, especially for experiments that are substantially exploratory in their objectives. However, at a minimum, preregistration merely requires that a narrative description of the experiment, its primary anticipated outcomes, and broad results be expressed on the registry.

ACM's Publication Board might also consider proposing that ACM host its own experimental registry. The ACM Digital Library is an outstanding resource for computer scientists, and it includes a wide range of facilities for storing, search, and retrieving supplementary material in association with archival papers. Modifying the digital library to additionally host an experimental registry would be a substantial undertaking, but one which would serve the goal of 'aggressively developing the highest-quality content'.

Use: What can researchers do?

Regardless of action from ACM's Publications Board, HCI researchers can begin registering their experimental intentions on registries such as the Harvard Dataverse https://dataverse.harvard.edu/, As Predicted https://aspredicted.org/, and the Open Science Framework http://osf.io. Although limited in its functionality, As Predicted may be particularly appropriate for work that is intended for CHI because it permits the production of an anonymous registration that meets CHI's requirements for anonymous submissions. It also allows the registration to remain private until the author explicitly makes it public. However, As Predicted currently lacks the ability for researchers to upload results, and therefore it offers little value in opening access to null findings.

Hopefully, as researchers begin to include references to their preregistered intentions within their CHI submissions and publications, expectations will change, and studies without preregistration will become abnormal.

Expectations: What can journals and conferences do?

Within clinical trials, it was mandatory action by funding authorities and publication venues that triggered widespread adoption of preregistration [10]. In HCI, decisions by journal editorial boards or by conference steering committees could similarly prompt rapid change in the adoption of preregistration. Even without forceful measures in *requiring* preregistration, publication venues could easily stress that preregistration is a desirable activity that should positively influence the submission's consideration during its review.

[2] https://www.acm.org/publications/publications-board-committees

Potential drawbacks of preregistration

Would we regret opening the file drawer?

One argument against preregistration is that many studies are deservedly consigned to the hidden file drawer, and that opening the drawer (through preregistration and uploading of results) would increase the amount of noise in publicly accessible science.

Finding a null outcome from an experiment means one of two things – either the experiment was insensitive and incorrectly failed to reject the null hypothesis (a Type II error), or the tested effect does not exist (e.g., there really is no difference between the conditions). Assuming an HCI experiment is well founded, it is unlikely that two different conditions truly have identical population means (to the *n*th decimal place) – as Cohen put it, 'the null hypothesis is always false' [8]. However, it is also possible that an experiment's foundation is poor, and the null hypothesis true – for example (facetiously), a Fitts' Law examination of the effect of sock colour on target selection performance should return a null outcome[3].

The discipline of HCI would gain few benefits from opening the file drawer on unfounded experiments that show null outcomes. However, the costs of exposing these experiments are unlikely to be high because search tools are increasingly capable in supporting selective identification of target material.

Conversely, for studies in which null results represent a Type II error, there are substantial potential benefits in opening the file drawer. For example, subtle effects in interaction, which require large samples to reliably expose their impact, might consistently show marginal effects that exceed $p = .05$, leading to repeated rejection for failing to satisfy NHST criteria. If the file drawer were open on such outcomes, meta-analysis would allow reliable inference from the multitude of studies.

Which studies should be preregistered?

HCI researchers often engage in multiple preliminary or pilot studies in which conditions and methods are progressively tuned to create sensitive experiments that reliably reveal significant effects. There are therefore legitimate questions about where the boundaries lie between pilot studies that are intended to refine conditions and full studies that are intended to test formal hypotheses through NHST. As described earlier, even when experimental objectives are exploratory, researchers may find the benefits of preregistration appealing and choose to do so. However, we believe that researchers should feel compelled to preregister their experiment whenever it is *possible* that the results of a study will be submitted for consideration for publication with NHST analysis.

How do I publish when my results are null?

Evidence from other disciplines suggests that preregistration has a dramatic effect in reducing the proportion of experiments that reject the null hypothesis [25, 14, 21]. Assuming the same reduction occurred in HCI, we could anticipate some combination of the following outcomes – a substantial drop in the number of published papers, an increase in the proportion of studies reporting null outcomes, and a reduction in the proportion of papers employing NHST. The latter two outcomes are both arguably positive for the discipline, and the former is determined by publication venues.

Furthermore, preregistration does not constrain authors in their choice of posthoc methods. Any analysis that might be used in an unregistered experiment could also be used in a preregistered one, but the language used to describe the analysis is likely to change (eliminating any opportunity for hidden HARKing). As suggested on aspredicted.org, authors of preregistered experiments could explain their results using phrases such as 'Contrary to expectations...' or 'In addition to the preregistered analysis, we also ran...'.

Loss of freedom

Experimental preregistration intentionally removes researchers' freedom to hide the fact that they have altered their hypotheses in light of their results – they can HARK no more. This is important not least because it reduces the incidence of Type I errors. While this loss of freedom may seem unappealing to some, we reiterate that the loss does not imply an impediment to researchers' ability to conduct exploratory research. Rather, preregistration should aid and enhance research that is exploratory for at least two reasons: 1) researchers who preregister their exploratory intent remove the validity of any reviewer criticism that the stated exploratory objectives were due to the failure of NHST outcomes; 2) as the proportion of studies rejecting the null hypothesis is likely to decrease with the introduction of preregistration, the use of NHST is likely to lose some of its appeal, particularly for exploratory studies, hopefully easing reviewer expectation for assessment of NHST outcomes.

CONCLUSION

Many scientific disciplines have concluded that experimental preregistration is an essential part of their discipline. It overcomes or reduces many substantial problems that are demonstrable in their literature, including HARKing, publication bias, and the file drawer effect. When registries are appropriately equipped, experimental preregistration also facilitates replication and meta-analyses.

The nature of HCI empirical research raises some challenges for preregistration, particularly its strong and appropriate reliance on iterative exploratory design and evaluation. However, preregistration can offer advantages even for exploratory work. Regardless of exploratory studies, much of the research knowledge within HCI is derived from formal experiments that make use of NHST, and for these studies our discipline is every bit as susceptible to problems of absent preregistration as any other; the potential benefits are equivalent too. Ideally preregistration would promote empirical publications being judged on the quality of the research, including the potential interest of the hypothesis, not simply whether they give the 'right' results. HCI research can sharpen its methods and rigour by introducing experimental preregistration.

ACKNOWLEDGMENTS

Thanks to Saul Greenberg, Philip Quinn, Shumin Zhai and reviewers for their helpful comments on drafts of this paper.

[3] We have not yet conducted this experiment, but anticipate one day reviewing it!

REFERENCES

1. APA. 2010. *Publication Manual of the American Psychological Association* (6th ed.). American Psychological Association.
2. Nikola Banovic. 2016. To Replicate or Not to Replicate? *GetMobile: Mobile Comp. and Comm.* 19, 4 (March 2016), 23–27. DOI: http://dx.doi.org/10.1145/2904337.2904346
3. Daniel Benjamin, James Berger, Magnus Johannesson, Brian Nosek, E. Wagenmakers, Richard Berk, Kenneth Bollen, Bjorn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher Chambers, Merlise Clyde, Thomas Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy Field, Malcom Forster, Edward George, Tarun Ramadorai, Richard Gonzalez, Steven Goodman, Edwin Green, Donald Green, Anthony Greenwald, Jarrod Hadfield, Larry Hedges, Leonhard Held, Teck Hau Ho, Herbert Hoijtink, James Jones, Daniel Hruschka, Kosuke Imai, Guido Imbens, John Ioannidis, Minjeong Jeon, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott Maxwell, Michael McCarthy, Don Moore, Stephen Morgan, Marcus Munafo, Shinichi Nakagawa, Brendan Nyhan, Timothy Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix Schonbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Zandt, Simine Vazire, Duncan Watts, Christopher Winship, Robert Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen Johnson. 2017. Redefine Statistical Significance. *PsyArXiv* (July 22 2017). DOI: http://dx.doi.org/10.17605/OSF.IO/MKY9J
4. Mario Biagioli. 2016. Watch out for cheats in citation game. *Nature* 535, 7611 (Jul 14 2016), 201. DOI: http://dx.doi.org/10.1038/535201a
5. J. P. Boissel. 1993. International Collaborative Group on Clinical Trial Registries: Position paper and consensus recommendations on clinical trial registries. *Clinical Trials and Meta-Analysis* 28, 4-5 (1993), 255–266.
6. Paul Cairns. 2007. HCI... Not As It Should Be: Inferential Statistics in HCI Research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1 (BCS-HCI '07)*. British Computer Society, Swinton, UK, UK, 195–201. http://dl.acm.org/citation.cfm?id=1531294.1531321
7. Lucas C. Coffman and Muriel Niederle. 2015. Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible. *Journal of Economic Perspectives* 29, 3 (September 2015), 81–98. DOI: http://dx.doi.org/10.1257/jep.29.3.81
8. Jacob Cohen. 1990. Things I have learned (so far). *American Psychologist* 45, 12 (1990), 1304 – 1312.
9. G. Cumming. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Routledge.
10. Rennie D. 2004. Trial registration: A great idea switches from ignored to irresistible. *JAMA* 292, 11 (2004), 1359–1362. DOI: http://dx.doi.org/10.1001/jama.292.11.1359
11. K. Dickersin, S. Chan, T. C. Chalmers, H. S. Sacks, and H. Smith Jr. 1987. Publication bias and clinical trials. *Controlled Clinical Trials* 8, 4 (1987), 343–353.
12. Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer International Publishing, Cham, 291–330. DOI: http://dx.doi.org/10.1007/978-3-319-26633-6_13
13. Wolfgang Forstmeier, Eric-Jan Wagenmakers, and Timothy H. Parker. 2016. Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews* (2016), n/a–n/a. DOI: http://dx.doi.org/10.1111/brv.12315
14. Annie Franco, Neil Malhotra, and Gabor Simonovits. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science* 345, 6203 (2014), 1502–1505. DOI:http://dx.doi.org/10.1126/science.1255484
15. C. A. E. Goodhart. 1984. *Problems of Monetary Management: The UK Experience*. Macmillan Education UK, London, 91–121. DOI: http://dx.doi.org/10.1007/978-1-349-17295-5_4
16. Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, USA, 111–120. DOI: http://dx.doi.org/10.1145/1357054.1357074
17. Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is Once Enough?: On the Extent and Content of Replications in Human-computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3523–3532. DOI: http://dx.doi.org/10.1145/2556288.2557004
18. George S. Howard, Scott E. Maxwell, and Kevin J. Fleming. 2000. The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods* 5, 3 (2000), 315 – 332. http://dx.doi.org/10.1037/1082-989X.5.3.315
19. Andrew Howes, Benjamin R. Cowan, Christian P. Janssen, Anna L. Cox, Paul Cairns, Anthony J. Hornof, Stephen J. Payne, and Peter Pirolli. 2014. Interaction Science SIG: Overcoming Challenges. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, New York, NY, USA, 1127–1130. DOI: http://dx.doi.org/10.1145/2559206.2559208

20. Macartan Humphreys, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration. *Political Analysis* 21, 1 (2013), 1. DOI:http://dx.doi.org/10.1093/pan/mps021

21. John P. A. Ioannidis. 2005. Why Most Published Research Findings Are False. *PLOS Medicine* 2, 8 (08 2005). DOI: http://dx.doi.org/10.1371/journal.pmed.0020124

22. 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. DOI: http://dx.doi.org/10.1177/0956797611430953 PMID: 22508865.

23. Valen E. Johnson. 2013. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* 110, 48 (2013), 19313–19317. DOI: http://dx.doi.org/10.1073/pnas.1313476110

24. Dickersin K and Rennie D. 2012. The evolution of trial registries and their use to assess the clinical trial enterprise. *JAMA* 307, 17 (2012), 1861–1864. DOI: http://dx.doi.org/10.1001/jama.2012.4230

25. 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 (08 2015), 1–12. DOI: http://dx.doi.org/10.1371/journal.pone.0132382

26. Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2016a. Special Interest Group on Transparent Statistics in HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 1081–1084. DOI: http://dx.doi.org/10.1145/2851581.2886442

27. Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016b. 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 (CHI '16)*. ACM, New York, NY, USA, 4521–4532. DOI: http://dx.doi.org/10.1145/2858036.2858465

28. Norbert L. Kerr. 1998. HARKing: Hypothesizing After the Results are Known. *Personality & Social Psychology Review (Lawrence Erlbaum Associates)* 2, 3 (1998), 196.

29. Don Lewis and C. J. Burke. 1949. The use and misuse of the chi-square test. *Psychological Bulletin* 46, 6 (1949), 433 – 489. https://www.ncbi.nlm.nih.gov/pubmed/15392587

30. H Lieberman. 2002. The Tyranny of Evaluation. http://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html. (2002). Last accessed: June 21, 2017.

31. 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 (CHI '07)*. ACM, New York, NY, USA, 1425–1434. DOI: http://dx.doi.org/10.1145/1240624.1240840

32. Michael E. J. Masson. 2011. A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods* 43, 3 (2011), 679–690. DOI: http://dx.doi.org/10.3758/s13428-010-0049-5

33. Bertrand Meyer. 2012. Incremental Research vs. Paradigm-shift Mania. *Commun. ACM* 55, 9 (Sept. 2012), 8–9. DOI:http://dx.doi.org/10.1145/2330667.2330670

34. James E. Monogan, III. 2013. A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections. *Political Analysis* 21, 1 (2013), 21. DOI:http://dx.doi.org/10.1093/pan/mps022

35. Robert Rosenthal. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin* 86, 3 (1979), 638 – 641. http://dx.doi.org/10.1037/0033-2909.86.3.638

36. J. D. Scargle. 1999. Publication Bias (The "File-Drawer Problem") in Scientific Inference. *ArXiv Physics e-prints* (Sept. 1999). http://adsabs.harvard.edu/abs/1999physics...9033S

37. Theodore D. Sterling. 1959. Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance — or Vice Versa. *J. Amer. Statist. Assoc.* 54, 285 (1959), 30–34. DOI: http://dx.doi.org/10.1080/01621459.1959.10501497

38. David Trafimow and Michael Marks. 2015. Editorial. *Basic and Applied Social Psychology* 37, 1 (2015), 1–2. DOI:http://dx.doi.org/10.1080/01973533.2015.1012991

39. Max L. L. Wilson, Paul Resnick, David Coyle, and Ed H. Chi. 2013. RepliCHI: The Workshop. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. ACM, New York, NY, USA, 3159–3162. DOI: http://dx.doi.org/10.1145/2468356.2479636

40. S Zhai. 2002. Evaluation is the worst form of HCI research except all those other forms that have been tried. www.shuminzhai.com/papers/EvaluationDemocracy.htm. (2002). Last accessed: June 21, 2017.