Threats of a replication crisis in empirical computer science

Andy Cockburn, Pierre Dragicevic, Lonni Besançon, Carl Gutwin

Threats of a Replication Crisis in Empirical Computer Science

Andy Cockburn[1] Pierre Dragicevic[2] Lonni Besançon[3] Carl Gutwin[4]

[1]University of Canterbury, New Zealand
[2]Inria, Université Paris-Saclay, France
[3]Linköping University, Sweden
[4]University of Saskatchewan, Canada

This is the authors' own version. The final version is available at https://doi.org/10.1145/3360311

Key insights:

- Many areas of computer science research (e.g., performance analysis, software engineering, artificial intelligence, and human-computer interaction) validate research claims by using statistical significance as the standard of evidence.
- A loss of confidence in statistically significant findings is plaguing other empirical disciplines, yet there has been relatively little debate of this issue and its associated 'replication crisis' in computer science.
- We review factors that have contributed to the crisis in other disciplines, with a focus on problems stemming from an over-reliance on – and misuse of – null hypothesis significance testing.
- Our analysis of papers published in a cross section of computer science journals suggests that a large proportion of computer science research faces the same threats to replication as those encountered in other areas.
- Computer science research can be greatly improved by following the steps taken by other disciplines, such as using more sophisticated evidentiary criteria, and showing greater openness and transparency through experimental preregistration and data/artifact repositories.

> *"If we do not live up to the traditional standards of science, there will come a time when no one takes us seriously" Peter J. Denning, 1980. [13]*

Almost forty years ago, Denning argued that computer science research could be strengthened by increased adoption of the scientific experimental method. Through the intervening decades, Denning's call has been answered. Few computer science graduate students would now complete their studies without some introduction to experimental hypothesis testing, and computer science research papers routinely use p-values to formally assess the evidential strength of experiments.

Our analysis of the ten most-downloaded articles from each of the 41 ACM "Transactions" journals (discussed below) showed that statistical significance was used as an evidentiary criterion in 61 articles (15%) across 21 different journals (51%), and in varied domains: from the evaluation of classification algorithms, to comparing the performance of cloud computing platforms, to assessing a new video-delivery technique in terms of quality of experience.

While computer science research has increased its use of experimental methods, the scientific community's faith in these methods has been eroded in several areas, leading to a 'replication crisis' [32, 27] in which experimental results cannot be reproduced and published findings are mistrusted. Consequently, many disciplines have taken steps to understand and try to address these problems. In particular, misuse of statistical significance as the standard of evidence for experimental success has been identified as a key contributor in the replication crisis. But there has been relatively little debate within computer science about this problem or how to address it. If computer science fails to adapt while others move on to new standards then Denning's concern will return – other disciplines will stop taking us seriously.

Some terminology

- *Publication bias:* papers supporting their hypotheses are accepted for publication at a much higher rate than those that do not.
- *File drawer effect:* null findings tend to be unpublished and therefore hidden from the scientific community.
- *p-hacking:* manipulation of experimental and analysis methods to produce statistically significant results. Used as a collective term in this paper for a variety of undesirable research practices.
- *p-fishing:* seeking statistically significant effects beyond the original hypothesis.
- *HARKing:* Hypothesising After the Results are Known; post-hoc reframing of experimental intentions to present a p-fished outcome as having been predicted from the start.

Beyond issues of statistical significance, computer science research raises some distinct challenges and opportunities for experimental replication. Computer science research often relies on complex artefacts such as source code and data sets, and with appropriate packaging, replication of some computer experiments can be substantially automated. The replicability problems associated with access to research artefacts have been broadly discussed in computer systems research (e.g., [25, 9]), and the ACM now awards badges to recognise work that is *repeatable* (the original team of researchers can reliably produce the same result using the same experimental setup), *replicable* (a different team can produce the same result using the original setup), and *reproducible* (a different team can produce the same result using a different experimental setup) [5]. However, these definitions are primarily directed at experiments that analyse the results of computations (such as new computer algorithms, systems, or methods), and uptake of the badges has been slow in fields involving experiments with human participants. Furthermore, the main issues contributing to the replication crisis in other experimental disciplines do not stem from access to artefacts; rather, they largely stem from a misuse of evidentiary criteria used to determine whether an experiment was successful or not.

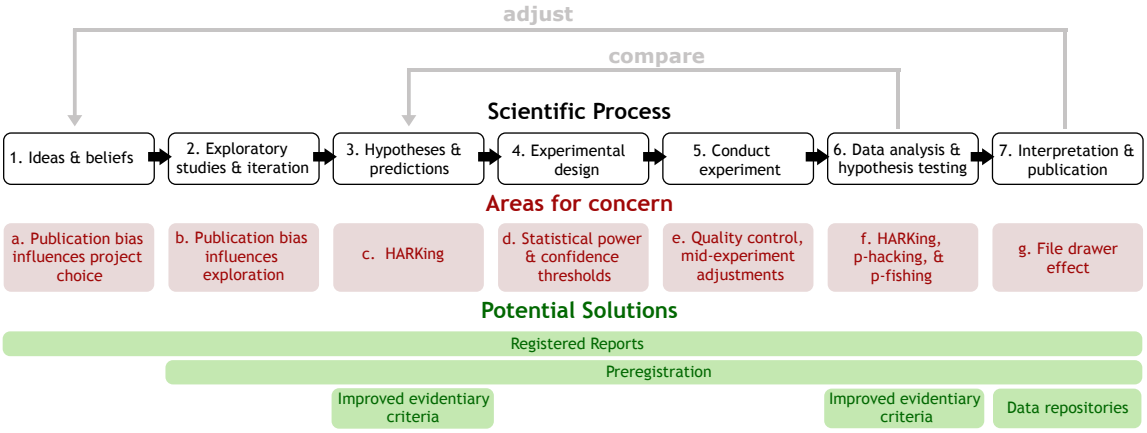

Figure 1. Stages of a typical experimental process (top, adapted from [18]), prevalent concerns at each stage (middle), and potential solutions (bottom).

The following section reviews the extent and causes of the replication crisis in other areas of science, with a focus on issues relating to the use of null hypothesis significance (NHST) as an evidentiary criterion. We then report on our analysis of a cross section of computer science publications to identify how common NHST is in our discipline. Two sections then review potential solutions, dealing first with alternative ways to analyse data and present evidence for hypothesised effects, and second arguing for improved openness and transparency in experimental research.

The Replication Crisis in Other Areas of Science

In assessing the scale of the crisis in their discipline, cancer researchers attempted to reproduce the findings of landmark research papers, finding that they could not do so in 47 of 53 cases [3], and psychology researchers similarly failed to replicate 39 out of 100 studies [31]. Results of a recent Nature survey of more than 1,500 researchers found that 90% agree that there is a crisis, that more than 70% had tried and failed to reproduce another scientist's experiments, and that more than half had failed to replicate their own findings [2].

Experimental process

A scientist's typical process for experimental work is summarised along the top row of Figure 1, with areas of concern and potential solutions below. In this process, initial ideas and beliefs (item 1) are refined through formative explorations (2), leading to the development of specific hypotheses and associated predictions (3). An experiment is designed and conducted (4, 5) to test the hypotheses, and the resultant data is analysed and compared with the predictions (6). Finally, results are interpreted (7), possibly leading to adjustment of ideas and beliefs.

A critical part of this process concerns the evidentiary criteria used for determining whether

experimental results (at 6) conform with hypotheses (at 3). Null hypothesis significance testing (NHST) is one of the main methods for providing this evidence. When using NHST, a p-value is calculated that represents the probability of encountering data at least as extreme as the observed data if a null hypothesis of no effect were true. If that probability is lower than a threshold value (the α level, normally .05, representing the Type I error rate of false positives) then the null hypothesis is deemed untenable and the resultant finding is labelled 'statistically significant'. When the p-value exceeds the α level, results interpretation is not straightforward – perhaps there is no effect, or perhaps the experiment lacked sufficient power to expose a real effect (a Type II error or false negative, where β represents the probability of this type of error).

Publication bias

In theory, rejection of the null hypothesis should elevate confidence that observed effects are real and repeatable. But concerns about the dichotomous interpretation of NHST as 'significant' or not have been raised for almost 60 years. Many of these concerns stem from a troublesome *publication bias* in which papers that reject the null hypothesis are accepted for publication at a much higher rate than those that do not. Demonstrating this effect, Sterling [41] analysed 362 papers published in major psychology journals between 1955 and 1956, noting that 97.3% of papers that used NHST rejected the null hypothesis.

The high publication rates for papers that reject the null hypothesis contributes to a *file drawer effect* [35] in which papers that fail to reject the null go unpublished because they are not written up, written up but not submitted, or submitted and rejected [16]. Publication bias and the file drawer effect combine to propagate the dissemination and maintenance of false knowledge: through the file drawer effect, correct findings of no effect are unpublished and hidden from view; and through publication bias, a single incorrect chance finding (a 1:20 chance at $\alpha = .05$, if the null hypothesis is true) can be published and become part of a discipline's *wrong* knowledge.

Ideally, scientists are objective and dispassionate throughout their investigations, but knowledge of the publication bias strongly opposes these ideals. Publication success shapes careers, so researchers need their experiments to succeed (rejecting the null in order to get published), creating many areas of concern (middle row of Figure 1), as follows.

Publication bias negatively influences project selection

There are risks that the direction of entire disciplines can be negatively affected by publication bias (Figure 1a and g). Consider a young faculty member or graduate student who has a choice between two research projects: one that is mundane, but likely to satisfy a perceived publication criterion of $p < .05$; the other is exciting but risky in that results cannot be anticipated and may end up in a file drawer. Publication bias is likely to draw researchers towards safer topics in which outcomes are more certain, potentially stifling researchers' interest in risky questions.

Publication bias also disincentivises replication, which is a critical element of scientific validation. Researchers' low motivation to conduct replications is easy to understand – a successful

replication is likely to be rejected because it merely confirms what is already 'known', while a failure to replicate is likely to be rejected for failing to satisfy the $p < .05$ publication criterion.

Publication bias disincentivises exploratory research

Exploratory studies and iteration play an important role in the scientific process (Figure 1b). This is particularly true in areas of computer science, such as human-computer interaction, where there may be a range of alternative solutions to a problem. Initial testing can quickly establish viability and provide directions for iterative refinement. Insights from explorations can be valuable for the research community, but if reviewers have been trained to expect standards of statistical evidence that only apply to confirmatory studies (such as the ubiquitous p-value) then publishing insights from exploratory studies and exploratory data analyses may be difficult. In addition, scientists' foreknowledge that exploratory studies may suffer from these problems can deter them from carrying out the exploratory step.

Publication bias encourages HARKing

Publication bias encourages researchers to explore hypotheses that are different to those that they originally set out to test (Figure 1c and f). This practice is called 'HARKing' [23], which stands for Hypothesising After the Results are Known, also known as 'outcome switching'.

Figure 2. HARKing (Hypothesising After the Results are Known) is an instance of the Texas sharpshooter fallacy. Illustration by Dirk-Jan Hoek, CC-BY.

Diligent researchers will typically record a wide set of experimental data beyond that required to test their intended hypotheses – this is good practice, as doing so may help interpret and explain experimental observations. However, publication bias creates strong incentives for scientists to ensure that their experiments produce statistically significant results. Consciously or subconsciously,

they may steer their studies to ensure that experimental data satisfies $p < .05$. If the researcher's initial hypothesis fails (concerning task time, say) but some other data satisfies $p < .05$ (error rate, for example), then authors may be tempted to reframe the study around the data that will increase the paper's chance of acceptance, presenting the paper as having predicted that outcome from the start. This reporting practice, which is an instance of the so-called "Texas sharpshooter fallacy" (see Figure 2), essentially invalidates the NHST procedure due to inflated Type I error rates. For example, if a researcher collects 15 dependent variables and only reports statistically significant ones, and if we assume that in reality the experimental manipulation has no effect on any of the variables, then the probability of a Type I error is 54% instead of the advertised 5% [19].

While many scientists might agree that *other* scientists are susceptible to questionable reporting practices such as HARKing, evidence suggests that they are troublesomely widespread [20, 21]. For example, over 63% of respondents to a survey of 2000 psychology researchers admitted failing to report all dependent measures, which is often associated with the selective reporting of favourable findings [20].

Even without any intention to misrepresent data, scientists are susceptible to cognitive biases that may promote misrepresentations: for example, *apophenia* is the tendency to see patterns in data where none exists, and it has been raised as a particular concern for big-data analyses [6]; *confirmation bias* is the tendency to favour evidence that aligns with prior beliefs or hypotheses [30]; and *hindsight bias* is the tendency to see an outcome as having been predictable from the start [36], which may falsely assuage researchers' concerns when reframing their study around a hypothesis that differs from the original.

Publication bias encourages mid-experiment adjustments

In addition to the modification of hypotheses, other aspects of an experiment may be modified during its execution (Figure 1e), and the modifications may go unreported in the final paper. For example, the number of samples in the study may be increased mid-experiment in response to a failure to obtain statistical significance (56% of psychologists self-admitted to this questionable practice [20]). This, again, inflates Type I error rates, which impairs the validity of NHST.

Publication bias encourages questionable data analysis practices

Dichotomous interpretation of NHST can also lead to problems in analysis: once experimental data has been collected, researchers may be tempted to explore a variety of post-hoc data analyses to make their findings look stronger or to reach statistical significance (Figure 1f). For example, they might consciously or unconsciously manipulate various techniques such as excluding certain data points (e.g., removing outliers, excluding participants, or narrowing the set of conditions under test), applying various transformations to the data, or applying statistical tests only to particular data subsets. While such analyses can be entirely appropriate if planned and reported in full, engaging in a data 'fishing' exercise to satisfy $p < .05$ is not, especially if the results are then selectively reported. Flexible data analysis and selective reporting can dramatically increase Type

I error rates, and these are major culprits in the replication crisis [38].

Is Computer Science Research at Risk? (Spoiler: Yes)

Given that much of computer science research either does not involve experiments, or involves deterministic or large-sample computational experiments that are reproducible as long as data and code are made accessible, one could argue that the field is largely immune to replication issues that have plagued other empirical disciplines. To find out whether this argument is tenable, we analysed the ten most downloaded articles for each of the 41 ACM journals beginning with the name 'Transactions on'. We inspected all 410 articles to determine whether or not they used $p < \alpha$ (with α normally 0.05) as a criterion for establishing evidence of a difference between conditions. The presence of p-values is an indication of statistical uncertainty, and therefore of the use of non-deterministic small-sample experiments (for example involving human subjects). Furthermore, as we have previously discussed, the use of a dichotomous interpretation of p-values as 'significant' or 'not significant' is thought to promote publication bias and questionable data analysis practices, both of which heavily contributed to the replication crisis in other disciplines.

A total of 61 of the 410 computer science articles (15%) included at least one dichotomous interpretation of a p-value[1]. All but two of the papers that used dichotomous interpretations (97%) identified at least one finding as satisfying the $p < .05$ criterion, suggesting that publication bias (long observed in other disciplines [41]) is likely to also exist in empirical computer science. Furthermore, 21 different journals (51%) included at least one article using a dichotomous interpretation of p within the set of 10 papers inspected. The count of articles across journals is summarised in Figure 3, with fields such as applied perception, education, software engineering, information systems, bioinformatics, performance modelling, and security all showing positive counts.

Our survey showed four main ways in which experimental techniques are used in computer science research, spanning work in graphics, software engineering, artificial intelligence, and performance analysis, as well as the expected use in human-computer interaction. First, empirical methods are used to assess the quality of an artifact produced by a technique, using humans as judges (e.g., the photorealism of an image or the quality of streaming video). Second, empirical methods are used to evaluate classification or prediction algorithms on real-world data (e.g., a power scheduler for electric vehicles, using real data from smart meters). Third, they are used to carry out performance analysis of hardware or software, using actual data from running systems (e.g., a comparison of real cloud computing platforms). Fourth, they are used to assess human performance with interfaces or interaction techniques (e.g., which of two menu designs is faster).

Given the high proportion of computer science journals that accept papers using dichotomous interpretations of p, it seems unreasonable to believe that computer science research is immune to the problems that have contributed to a replication crisis in other disciplines. The two following sections review proposals from other disciplines on how to ease the replication crisis, focusing

[1]Data for this analysis is available at osf.io/hkqyt/, including a quote extracted from each counted paper, showing its use of a dichotomous interpretation.

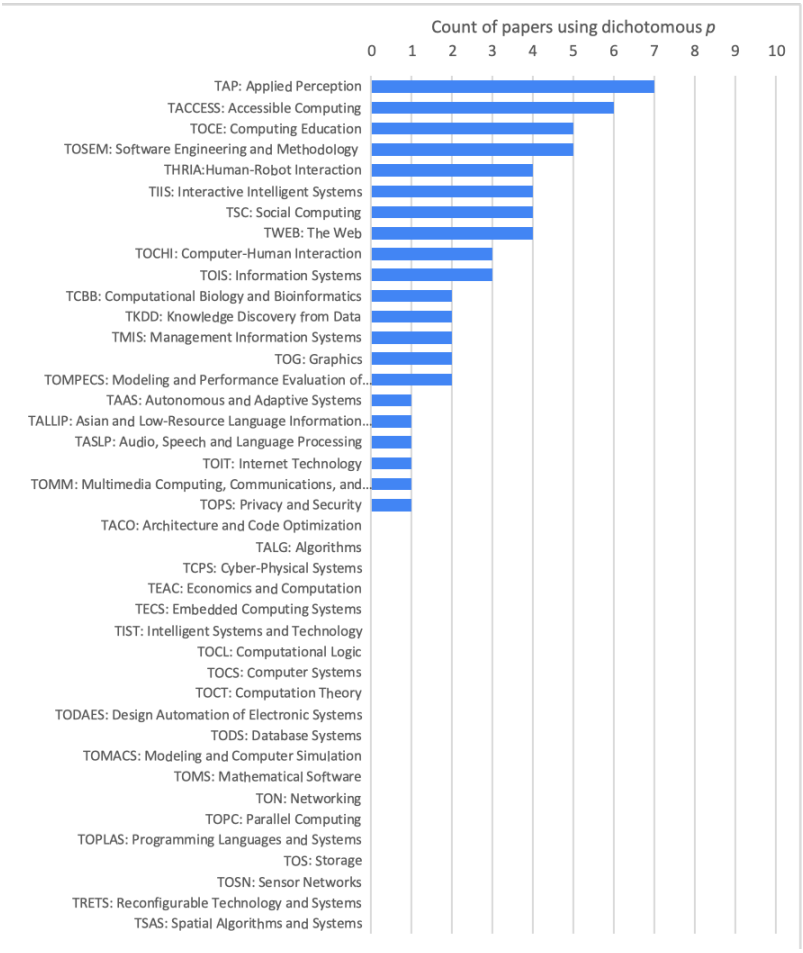

Figure 3. Count of articles from among the '10 most downloaded' (24th May, 2019) that use dichotomous interpretations of p from among ACM journals titled 'Transactions on...'

first on changes to the way in which experimental data is analysed, and second on proposals for improving openness and transparency.

Proposals for Easing the Crisis: Better Data Analysis

Redefine statistical significance

Many researchers attribute some of the replication crisis to the dominant use of NHST. Among the noted problems with NHST is the ease with which experiments can produce false-positive findings, even without scientists contributing to the problem through questionable research practices. To address this problem, a group of 75 senior scientists from diverse fields (including computer

science) proposed that the accepted norm for determining 'significance' in NHST tests be reduced from $\alpha = .05$ to $\alpha = .005$ [4]. Their proposal was based on two analyses – the relationship between Bayes factors and p-values, and the influence of statistical power on false positive rates – both of which indicated disturbingly high false positive rates at $\alpha = .05$. The authors also recommended that the word 'suggestive' be used to describe results in the range $.005 <= p < .05$.

Despite the impressive list of authors, this proposal attracted heavy criticism (see [33] for a review). Some have argued that the reasoning behind the .005 threshold is flawed, and that adopting it could actually make the replication crisis worse (by causing a drop in the statistical power of studies without reducing incentives for p-hacking, and by diverting resources away from replications). Another argument is that the threshold value remains arbitrary, and that focusing instead on effect sizes and their interval estimates (confidence intervals or credible intervals) can better characterise results. There is also a pragmatic problem that until publication venues firmly announce their standards, authors will be free to choose terminology ('statistically significant' at $p < .05$ or 'statistically significant' at $p < .005$) and reviewers/readers may differ in their expectations. Furthermore, the proposal does nothing to discourage or prevent problems associated with inappropriate modification of experimental methods and objectives after they begin.

Abandon statistical significance

Many researchers argue that the replication crisis does not stem from the choice of the .05 cutoff, but from the general idea of using an arbitrary cutoff to classify results, in a dichotomous manner, as statistically significant or not. Some of these researchers have called for reporting exact p-values and abandoning the use of statistical significance thresholds [1]. Recently, a comment published in Nature with more than 800 signatories called for abandoning binary statistical significance [28]. Cumming [12] argued for the banning of p-values altogether, and recommended the use of estimation statistics where strength of evidence is assessed in a non-dichotomous manner, by examining confidence intervals. Similar recommendations have been made in computer science [14]. The editorial board of the Basic and Applied Social Psychology journal went further by announcing that it would not publish papers containing any statistics that could be used to derive dichotomous interpretations, including p-values and confidence intervals [42]. Overall there is no consensus on what should replace NHST, but many methodologists are in favour of banning dichotomous statistical significance language.

Despite the forceful language opposing NHST (e.g., "very few defences of NHST have been attempted" [12, p11]), some researchers believe NHST and the notion of dichotomous hypothesis testing still have their place [4]. Others have suggested that the calls to abandon NHST are a red herring in the replicability crisis [37], not least due to the lack of evidence that doing so will aid replicability.

Adopt Bayesian statistics

Several researchers propose replacing NHST with Bayesian statistical methods. One of the key motivators for doing so concerns a common misunderstanding 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 represent this value. However, the p-value actually represents the probability of observing data at least as extreme as the sample if the null hypothesis were true: $P(D|H_0)$. In contrast to NHST, Bayesian statistics can enable the desired computation of $P(H_0|D)$.

Bayesian statistics are perfectly suited for doing estimation statistics, and have several advantages over confidence intervals [26, 22]. Nevertheless, they can also be used to carry out dichotomous tests, possibly leading to the same issues as NHST. Furthermore, Bayesian analysis is not immune to the problems of p-hacking – researchers can still 'b-hack' to manipulate experimental evidence [39, 37]. In particular, the choice of priors adds an important additional experimenter degree of freedom in Bayesian analysis [39].

Help the reader form their own conclusion

Given the contention over the relative merits of different statistical methods and thresholds, researchers have proposed that when reporting results, authors should focus on assisting the reader in reaching their own conclusions by describing the data and the evidence as clearly as possible. This can be achieved through the use of carefully-crafted charts that focus on effect sizes and their interval estimates, and the use of cautionary language in the author's interpretations and conclusions [11, 14].

While improved explanation and characterisation of underlying experimental data is naturally desirable, authors are likely to encounter problems if relying only on the persuasiveness of their data. First, the impact of using more cautious language on the persuasiveness of arguments when compared to categorical arguments is still uncertain [15]. Second, many reviewers of empirical papers are familiar and comfortable with NHST procedures and its associated styles of results reporting, and they may criticise its absence; in particular, reviewers may suspect that the absence of reported dichotomous outcomes is a consequence of their failure to attain $p < .05$. Both of these concerns suggest that a paper's acceptance prospects could be harmed if lacking simple and clear statements of results outcome, such as those provided by NHST, despite the simplistic and often misleading nature of such dichotomous statements.

Quantify p-hacking in published work

None of the above proposals address problems connected with researchers consciously or subconsciously revising experimental methods, objectives, and analyses after their study has begun. Statistical analysis methods exist that allow researchers to assess whether a set of already published studies are likely to have involved such practices. A common method is based on the p-curve, which is the distribution of statistically significant p-values in a set of studies [40]. Studies of true effects should produce a right-skewed p-curve, with many more low statistically significant

p-values (e.g., .01s) than high values (e.g., .04s); but a set of p-hacked studies are likely to show a left-skewed p-curve, indicative of selecting variables that tipped analyses into statistical significance.

While use of p-curves appears promising, it has several limitations. First, it requires a set of study results to establish a meaningful curve, and its use as a diagnostic tool for evidence of p-hacking in any single article is discouraged. Second, its usefulness for testing the veracity of any particular finding in a field depends on the availability of a series of related or replicated studies; but replications in computer science are rare. Third, statisticians have questioned the effectiveness of p-curves for detecting questionable research practices, demonstrating through simulations that p-curve methods cannot reliably distinguish between p-hacking of null effects and studies of true effects that suffer experimental omissions such as unknown confounds [7].

Openness, Preregistration and Registered Reports

While the debate continues over the merits of different methods for data analysis, there is a wide agreement on the need for improved openness and transparency in empirical science. This includes making materials, resources, and datasets available for future researchers who might wish to replicate the work.

Making materials and data available after a study's completion is a substantial improvement, because it greatly facilitates peer scrutiny and replication. However, it does not prevent questionable research practices, since the history of a data analysis (including possible p-hacking) is not visible in the final analysis scripts. And if others fail to replicate a study's findings, the original authors can easily explain away the inconsistencies by questioning the methodology of the new study or by claiming that an honest Type I error occurred.

Overcoming these limitations requires a clear statement of materials, methods, and hypotheses *before* the experiment is conducted, as provided by experimental preregistration and registered reports, discussed next.

Experimental preregistration

In response to concerns about questionable research practices, various authorities instituted registries in which researchers preregister their intentions, hypotheses and methods (including sample sizes and precise plans for the data analyses) for upcoming experiments. Risks of p-hacking or outcome switching are dramatically reduced when a precise statement of method predates the experimental conduct. Furthermore, if the registry subsequently stores experimental data, then the file drawer is effectively opened on experimental outcomes that might otherwise have been hidden due to failure to attain statistical significance.

Although many think preregistration is only a recent idea, and therefore one that needs to be refined and tested before it can be fully adopted, it has in fact been in place for a long time in medical research. In 1997, the US Food and Drug Administration Modernization Act (FDAMA) established the registry `ClinicalTrials.gov`, and over 96,000 experiments were registered in its

first ten years, assisted by the decision of the International Committee of Medical Journal Editors to make preregistration a requirement for publication in their journals [34]. Results suggest that preregistration has had a substantial effect on scientific outcomes – for example, an analysis of studies funded by the National Heart, Lung, and Blood Institute between 1970 and 2012 showed that the rate at which studies showed statistically significant findings plummeted from 57% before the introduction of mandatory preregistration (in 2000) to only 8% after [21]. The success of ClinicalTrials.gov and the spread of the replication crisis to other disciplines has prompted many disciplines to introduce their own registries, including the American Economic Association (https://www.socialscienceregistry.org/) and the political science 'dataverse' [29]. The Open Science Framework (OSF) also supports preregistration, ranging from simple and brief descriptions through to complete experimental specification (http://osf.io). Although originally focused on replications of psychological studies, it is now used in a range of disciplines, including by computer scientists.

Registered reports

While experimental preregistration should enhance confidence in published findings, it does not prevent reviewers from using statistical significance as a criterion for paper acceptance. Therefore, it does not solve the problem of publication bias and does not help prevent the file drawer effect. As a result, the scientific record can remain biased towards positive findings, and since achieving statistical significance is harder if *p*-hacking is not an option, researchers may be even more motivated to focus on unsurprising but safe hypotheses where the null is likely to be rejected. However, we do not want to simply take null results as equivalent to statistical significance, because null results are trivially easy to obtain; instead, the focus should be on the quality of the question being asked in the research.

Registered reports are a way to provide this focus. With registered reports, papers are submitted for review *prior* to conducting the experiment. Registered reports include the study motivation, related work, hypotheses, and detailed method; everything that might be expected in a traditional paper *except* for the results and their interpretation. Submissions are therefore considered based on the study's motivations (is this an interesting research question?) and method (is the way of answering the question sound and valid?). If accepted, a registered report is published *regardless* of the final results.

A recent analysis of 127 registered reports in the bio-medical and psychological sciences showed that 61% of studies did not support their hypothesis, compared to the estimated 5-20% of null findings in the traditional literature [10]. As of February 2019, the Center for Open Science (https://cos.io/rr/) lists 136 journals that accept registered reports and 27 journals that have accepted them as part of a special issue. No computer science journal is currently listed.

Recommendations for Computer Science

The use of NHST in relatively small-sample empirical studies is an important part of many areas of computer science, creating risks for our own reproducibility crisis [24, 14, 8]. The following recommendations suggest activities and developments that computer scientists can work on to protect the credibility of the discipline's empirical research.

Promote preregistration

The ACM has the opportunity and perhaps the obligation to lead and support changes that improve empirical computer science – its stated purpose includes 'promotion of the highest standards' and the ACM Publications Board has the goal of 'aggressively developing the highest-quality content'. These goals would be supported by propagating to journal editors and conference chairs an expectation that empirical studies should be preregistered, preferably using transdisciplinary registries such as the Open Science Framework (**http://osf.io**). Authors of papers describing empirical studies could be asked or required to include a standardised statement at the end of their papers' abstract providing a link to the preregistration, or explicitly stating that the study was not preregistered (in other disciplines, preregistration is mandatory). Reviewers would also need to be educated on the value of preregistration and the potential implications of its absence.

It is worth noting that experimental preregistration has potential benefits to authors even if they do not intend to test formal hypotheses. If the registry entry is accessible at the time of paper submission (perhaps through a key that is disclosed to reviewers), then an author who preregisters an exploratory experiment is protected against reviewer criticism that the stated exploratory intent is due to HARKing following a failure to reject the null hypothesis [8].

Another important point regarding preregistration is that it does not constrain authors from reporting unexpected findings. 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 in the published paper must make the post-hoc discovery clear, such as 'Contrary to expectations...' or 'In addition to the preregistered analysis, we also ran...'

Publish registered reports

The editorial boards of ACM journals that feature empirical studies could adapt their reviewing process to support the submission of registered reports and push for this publication format. This is perhaps the most promising of all interventions aimed at easing the replication crisis – it encourages researchers to address interesting questions, it eliminates the need to produce statistically significant results (and thus addresses the file drawer problem), and it encourages reviewers to focus on the work's importance and potential validity [10]. In addition, it eliminates hindsight bias among reviewers, i.e., the sentiment that they could have predicted the outcomes of a study, and that the findings are therefore unsurprising.

The prospect of permitting the submission of registered reports to large-scale venues is daunting

(e.g., ACM CHI 2019 conference on Human-Computer Interaction received approximately 3000 submissions to its papers track). However, the two-round submission and review process adopted by conferences within the Proceedings of the ACM (PACM) series could be adapted to embrace the submission of registered reports at round 1. We encourage conference chairs to experiment with registered report submissions.

Encourage data and materials openness

The ACM Digital Library supports access to resources that could aid replication through links to auxiliary materials. However, more could be done to encourage or require authors to make data and resources available. Currently, authors decide whether or not to upload resources. Instead, uploading data could be compulsory for publication, with exceptions made only following special permission from an editor or program chair. While such requirements may seem draconian given the permissive nature of current practice in computer science, the requirement is common in other disciplines and outlets, such as Nature's 'Scientific Data' (www.nature.com/sdata/). A first step in this direction would be to follow *transparency and openness guidelines* (https://cos.io/our-services/top-guidelines/), which encourage authors to state in their submission whether or not they made their data, scripts, and pre-registered analysis available online, and to provide links to them where available.

Promote clear reporting of results

While the debate over standards for data analysis and reporting continues, certain best-practice guidelines are emerging. First, authors should focus on two issues: first, conveying effect sizes (this includes simple effect sizes such as differences between means [11]), and second, helping readers to understand the uncertainty around those effect sizes by reporting interval estimates [14, 26] or posterior distributions [22]. A range of recommendations already exist for improving reporting clarity and transparency, and must be followed more widely. For example, most effect sizes only capture central tendencies and thus provide an incomplete picture. Therefore, it can help to also convey population variability through well-known practices such as reporting standard deviations (and their interval estimates) and/or plotting data distributions. When reporting the outcomes of statistical tests the name of the test and its associated key data (such as degrees of freedom) should be reported. And, if describing the outcomes of a NHST test, the exact p-value should be reported. Since the probability of a successful replication depends on the order of magnitude of p [17], we suggest avoiding excessive precision (one or two significant digits are enough), and using scientific notation (e.g., $p = 2 \times 10^{-5}$) instead of inequalities (e.g., $p < .001$) when reporting very small p-values.

Encourage replications

The introduction of preregistration and registered reports in other disciplines caused a rapid decrease in the proportion of studies finding statistically significant effects. Assuming the same was

to occur in computer science, how would this influence accepted publications? It is likely that many more empirical studies would be published with statistically non-significant findings or with no statistical analysis (such as exploratory studies that rely on qualitative methods). It is also likely that this would encourage researchers to consider conducting experimental replications, regardless of previous outcomes. Replications of studies with statistically significant results help reduce Type I error rates, and replications of studies with null outcomes reduce Type II error rates and can test the boundaries of hypotheses. If better data repositories were available, computer science students around the world could contribute to the robustness of findings by uploading to registries the outcomes of replications conducted as part of their courses on experimental methods. Better data repositories with richer datasets would also facilitate meta-analyses, which elevate confidence in findings beyond that possible from a single study.

Educate reviewers (and authors)

Many major publication venues in computer science are under stress due to a deluge of submissions that creates challenges in obtaining expert reviews. Authors can become frustrated when reviewers focus on equivocal results of a well founded and potentially important study—but reviewers can also become frustrated when authors fail to provide definitive findings on which to establish a clear contribution. In the spirit of registered reports, our recommendation is to educate reviewers (and authors) on the research value of studying interesting and important effects, largely irrespective of the results generated. If reviewers focused on questions and method rather than traditional evidentiary criteria such as $p < .05$, then researchers would be better motivated to identify interesting research questions, including potentially risky ones. One potential objection to risky studies is their typically low statistical power: testing null effects or very small effects with small samples can lead to vast overestimations of effect sizes [27]. However, this is mostly true in the presence of *p*-hacking or publication bias, two issues that are eliminated by moving beyond the statistical significance filter and adopting registered reports.

References

[1] Amrhein, V., Korner-Nievergelt, F., and Roth, T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ 5*, 7 (2017), e3544.

[2] Baker, M. Is there a reproducibility crisis? *Nature 533*, 7604 (2016), 452–454.

[3] Begley, C. G., and Ellis, L. M. Raise standards for preclinical cancer research. *Nature 483*, 7391 (2012), 531.

[4] Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E., Berk, R., Bollen, K., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C., Clyde, M., Cook, T., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A., Forster, M., George, E., Ramadorai, T., Gonzalez, R., Goodman, S., Green, E., Green,

D., Greenwald, A., Hadfield, J., Hedges, L., Held, L., Hau Ho, T., Hoijtink, H., Jones, J., Hruschka, D., Imai, K., Imbens, G., Ioannidis, J., Jeon, M., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S., McCarthy, M., Moore, D., Morgan, S., Munafo, M., Nakagawa, S., Nyhan, B., Parker, T., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schonbrodt, F., Sellke, T., Sinclair, B., Tingley, D., Zandt, T., Vazire, S., Watts, D., Winship, C., Wolpert, R., Xie, Y., Young, C., Zinman, J., and Johnson, V. Redefine statistical significance. *PsyArXiv* (July 22 2017).

[5] Boisvert, R. F. Incentivizing reproducibility. *Commun. ACM 59*, 10 (Sept. 2016), 5–5.

[6] Boyd, D., and Crawford, K. Critical questions for big data. *Information, Communication & Society 15*, 5 (2012), 662–679.

[7] Bruns, S. B., and Ioannidis, J. P. A. p-curve and p-hacking in observational research. *PLOS One 11*, 2 (02 2016), 1–13.

[8] Cockburn, A., Gutwin, C., and Dix, A. HARK no more: On the preregistration of CHI experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 141:1–141:12.

[9] Collberg, C., and Proebsting, T. A. Repeatability in computer systems research. *Commun. ACM 59*, 3 (Feb. 2016), 62–69.

[10] Cristea, I. A., and Ioannidis, J. P. A. P values in display items are ubiquitous and almost invariably significant: A survey of top science journals. *PLOS One 13*, 5 (2018), e0197440.

[11] Cumming, G. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Multivariate applications series. Routledge, 2012.

[12] Cumming, G. The new statistics: Why and how. *Psychological Science 25*, 1 (2014), 7–29.

[13] Denning, P. J. ACM President's letter: What is experimental computer science? *Commun. ACM 23*, 10 (Oct. 1980), 543–544.

[14] Dragicevic, P. Fair statistical communication in HCI. In *Modern Statistical Methods for HCI*, J. Robertson and M. Kaptein, Eds. Springer International Publishing, Cham, 2016, pp. 291–330.

[15] Durik, A. M., Britt, M. A., Reynolds, R., and Storey, J. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology 27*, 3 (2008), 217–234.

[16] Franco, A., Malhotra, N., and Simonovits, G. Publication bias in the social sciences: Unlocking the file drawer. *Science 345*, 6203 (2014), 1502–1505.

[17] Goodman, S. N. A comment on replication, p-values and evidence. *Statistics in medicine 11*, 7 (1992), 875–879.

[18] Gundersen, O. E., and Kjensmo, S. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018* (2018), pp. 1644–1651.

[19] Ioannidis, J. P. A. Why most published research findings are false. *PLOS Medicine 2*, 8 (08 2005).

[20] John, L. K., Loewenstein, G., and Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science 23*, 5 (2012), 524–532. PMID: 22508865.

[21] Kaplan, R. M., and Irvin, V. L. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLOS One 10*, 8 (08 2015), 1–12.

[22] Kay, M., Nelson, G. L., and Hekler, E. B. 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* (2016), ACM, pp. 4521–4532.

[23] Kerr, N. L. Harking: Hypothesizing after the results are known. *Personality & Social Psychology Review (Lawrence Erlbaum Associates) 2*, 3 (1998), 196.

[24] Kosara, R., and Haroz, S. Skipping the Replication Crisis in Visualization: Threats to Study Validity and How to Address Them. In *Evaluation and Beyond - Methodological Approaches for Visualization* (Berlin, Germany, Oct. 2018).

[25] Krishnamurthi, S., and Vitek, J. The real software crisis: Repeatability as a core value. *Commun. ACM 58*, 3 (Feb. 2015), 34–36.

[26] Kruschke, J. K., and Liddell, T. M. The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review 25*, 1 (2018), 178–206.

[27] Loken, E., and Gelman, A. Measurement error and the replication crisis. *Science 355*, 6325 (2017), 584–585.

[28] McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. Abandon statistical significance. *The American Statistician 73*, sup1 (2019), 235–245.

[29] Monogan, III, J. E. A case for registering studies of political outcomes: An application in the 2010 house elections. *Political Analysis 21*, 1 (2013), 21.

[30] Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology 2*, 2 (1998), 175–220.

[31] Open Science Collaboration and others. Estimating the reproducibility of psychological science. *Science 349*, 6251 (2015), aac4716.

[32] Pashler, H., and Wagenmakers, E.-J. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science 7*, 6 (2012), 528–530.

[33] Perezgonzalez, J. D., and Frias-Navarro, D. Retract 0.005 and propose using JASP, instead. *Preprint available at https://psyarxiv.com/t2fn8* (2017).

[34] Rennie, D. Trial registration: A great idea switches from ignored to irresistible. *JAMA 292*, 11 (2004), 1359–1362.

[35] Rosenthal, R. The file drawer problem and tolerance for null results. *Psychological Bulletin 86*, 3 (1979), 638 – 641.

[36] Sanbonmatsu, D. M., Posavac, S. S., Kardes, F. R., and Mantel, S. P. Selective hypothesis testing. *Psychonomic Bulletin & Review 5*, 2 (Jun 1998), 197–220.

[37] Savalei, V., and Dunn, E. Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology 6* (2015), 245.

[38] Simmons, J. P., Nelson, L. D., and Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*, 11 (2011), 1359–1366.

[39] Simonsohn, U. Posterior-hacking: Selective reporting invalidates Bayesian results also. *SSRN* (2014).

[40] Simonsohn, U., Nelson, L. D., and Simmons, J. P. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General 143*, 2 (2014), 534–547.

[41] Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance — or vice versa. *Journal of the American Statistical Association 54*, 285 (1959), 30–34.

[42] Trafimow, D., and Marks, M. Editorial. *Basic and Applied Social Psychology 37*, 1 (2015), 1–2.