The Continued Prevalence of Dichotomous Inferences at CHI

Lonni Besançon, Pierre Dragicevic

The Continued Prevalence of Dichotomous Inferences at CHI

Lonni Besançon
Linköping University, Sweden
lonni.besancon@gmail.com

Pierre Dragicevic
Inria, France
pierre.dragicevic@inria.fr

ABSTRACT

Dichotomous inference is the classification of statistical evidence as either sufficient or insufficient. It is most commonly done through null hypothesis significance testing (NHST). Although predominant, dichotomous inferences have proven to cause countless problems. Thus, an increasing number of methodologists have been urging researchers to recognize the continuous nature of statistical evidence and to ban dichotomous inferences. We wanted to see whether they have had any influence on CHI. Our analysis of CHI proceedings from the past nine years suggests that they have not.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**;

KEYWORDS

Dichotomous inferences, NHST, p-values, confidence intervals, dichotomous thinking, statistics.

ACM Reference Format:
Lonni Besançon and Pierre Dragicevic. 2019. The Continued Prevalence of Dichotomous Inferences at CHI. In *Proceedings of CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts).* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3290607.3310432

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05…$15.00
https://doi.org/10.1145/3290607.3310432

INTRODUCTION

By *dichotomous inference* we refer to the classification of statistical evidence as either sufficient or insufficient, typically through the use of conventional cutoffs. Although dichotomous inference can be carried out using a variety of statistical methods (e.g., Bayes factors [21], posterior probabilities [2],...), the most popular procedure is by far null hypothesis significance testing (NHST). Thus this paper focuses on NHST. Using NHST, one first computes the probability p of "one particular test statistic being as or more extreme than observed in our particular study, given that the model it is computed from is correct" [2]. This model typically includes the hypothesis that there is no effect, also called the "null hypothesis". The p-value is then compared to a cutoff α (typically α=.05). If p is smaller than α, then the null hypothesis is rejected, which means there is sufficient evidence to conclude that there is an effect. Otherwise, the null hypothesis cannot be rejected—the evidence is insufficient to conclude.

Although advocated in many textbooks and broadly applied in HCI, NHST is a loose mix of two incompatible philosophies of statistical inference—the computation and reporting of exact p-values follows Ronald Fisher, while the use of an α cutoff to guide decision making is taken from Neyman and Pearson [14]. Although the Neyman-Pearson approach is thought to be well suited for automated decision-making (e.g., for deciding which batches to reject in a factory production line), Fisher and many others after him have rejected it as entirely inappropriate for carrying out scientific research [14]. Although Fisher initially suggested that researchers can distrust results with $p > .05$ as a rule of thumb, he later insisted that p-values should be seen as a continuous measure of strength of evidence against the null hypothesis and stated that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses" [11].

While dichotomous inference through NHST has been used for decades and is still in use today, many have recognized that the ritualistic application of a cut-off leads to a number of problems. First, it promotes dichotomous thinking, i.e., thinking about evidence as black and white [8]. This often results in researchers putting too much trust on results having a p-value less than the conventional .05 threshold, irrespective of how far it is from that threshold [28]. This in turn typically results in conclusions being overstated, or in sample means being interpreted as accurate, ignoring their uncertainty [11]. At the same time, researchers tend to distrust p-values that are above the threshold, even if they just missed the mark [28]. An extreme version of this, which involves the fallacy of taking the absence of evidence as evidence of absence, consists in taking a non-significant p-value (even $p = .06$) as a sign that there is no effect. A related error is considering that a significant effect and a non-significant effect are statistically different [12, 13], or that a stream of studies with significant and non-significant results is necessarily inconsistent or controversial [8]. Finally, dichotomous inference with NHST encourages practices that distort the scientific literature, such as publication bias (studies that do not achieve statistical significance are never published), outcome reporting bias (results that

do not achieve statistical significance are not reported in published studies), and significance chasing (researchers try many different analysis methods until they obtain a significant result) [29]. All of these issues contribute to making published studies less trustworthy and less likely to replicate.

Consequently, the practice of NHST-based dichotomous inference has been strongly discouraged by countless statisticians and methodologists, especially in the past few years [1–3, 8, 10–12, 15, 17, 25, 29]. In 2016, the executive director of the American Statistical Association stated that "in the post p<0.05 era, scientific argumentation is not based on whether a p-value is small enough or not. [...] Evidence is thought of as being continuous rather than some sort of dichotomy" [24]. Recently, Gerd Gigerenzer, a prominent psychologist and methodologist, suggested that "editors should no longer accept manuscripts that report results as "significant" or "not significant"" [15], while others went as far as qualifying NHST-based dichotomous inference as "scientifically destructive behavior" [2].

A commonly advocated alternative is to focus on effect sizes and their interval estimates rather than on p-values [2, 8, 11]. However, interval estimates do not offer a sure protection against dichotomous inference, as researchers still tend to classify results as statistically significant or not, depending on whether an interval estimate contains zero [2, 7]. Other methodologists argue that p-values should still be used, but only exact p-values should be reported and no mention of statistical significance should be made [1, 2, 15–17]. This stands in contrast to many guidelines and textbooks which recommend reporting exact p-values but without discouraging dichotomous inference [31]. Irrespective of the statistical tools used, many modern methodologists urge researchers to think of evidence as gradual rather than binary when interpreting their results [2, 8, 12, 25]. Peter Dixon introduced the "graded evidence" principle, according to which "similar results should lead to similar interpretations. In other words, if the results change a little bit, the evidence afforded by those results should only change a little bit" [10]. He added that "describing results in terms of a catalogue of significant and nonsignificant effects fails to satisfy this principle" and that "classifying results as either significant or nonsignificant is an impoverished, potentially misleading way to describe evidence" [10]. Similarly, it has been suggested that in HCI "a statistical analysis should [...] be designed so that similar experimental outcomes yield similar results and conclusions" [11]. Nevertheless, some methodologists believe that α cutoffs still have a place, and suggest for example that the issues of overconfident claims and irreplicable findings can be alleviated by switching to a more stringent cutoff of $\alpha = .005$ [5].

Although there is still an ongoing debate on whether dichotomous inferences should be banned from the researcher's toolbox, the past few years have seen a prolific literature and solid arguments against their use. Since HCI (like many other disciplines) has traditionally been dominated by NHST-based dichotomous inference, we wanted to examine whether the recent literature against dichotomous inferences has had any influence on CHI authors in the past few years. To this end, we analyzed all articles from the CHI proceedings between 2010 and 2018, using p-value inequalities (e.g., p<.01) and statistical significance language as indicators of dichotomous inferences.

CHI PROCEEDINGS ANALYSIS

We collected the CHI conference proceedings from 2010 to 2018 (4234 articles in total), and converted all the PDF files to text files using *pdfminer*[1]. We then analyzed the text files and extracted sentences using *NLTK*[2]. All scripts, results and plots are available as supplementary material[3].

We were interested both in how inferential statistics are reported, and in the use of significance language. For the former, we examined how often p-values were reported in the form of inequalities (e.g., $p < .05$, $p < .01$, $p > .05$), and how often they were reported as exact values (e.g., $p = .0412$). Although p-value inequalities are indicative of the use of NHST cutoffs, most guidelines that recommend reporting exact p-values also recommend reporting an inequality when the p-value is very small (e.g., $p < .001$ [31] or $p < .0001$). Therefore, we classified those cases as ambiguous. In addition, we looked at the reporting of confidence intervals, which are by far the most common interval estimates [8].

We therefore used the following search strings:

- To find p-value inequalities we looked for "p <", "p >", "p<", and "p>".
- To find exact p-values we looked for "p =" and "p=".
- To identify ambiguous cases we looked for occurrences of "p <X" and "p<X", with X < 0.01. These occurrences were eliminated from the list of p-value inequalities.
- To identify the reporting of confidence intervals we looked for "confidence interval", "% ci", and "%ci". All string searches were case-insensitive.

The use of significance language is more difficult to detect. Although the phrase "statistically significant" is univocal, many authors use the term "significant" without the qualifier "statistically", rendering the word ambiguous. For example, "a significant decrease" can be used to express effect magnitude, while occurrences of "a significant contribution" or "significant others" are unlikely to be related to statistical inference. Nevertheless, phrases such as "no significant effect" or "a significant interaction" are reasonably likely to refer to statistical significance. In order to gather a list of use cases of "significant" and "significantly" that are likely to refer to statistical significance, we listed all trigrams (sequence of three words) that contained either of these two words in the middle. We obtained a list of 10,334 trigrams, which we pruned by removing all trigrams occurring less than three times (the most common trigram was "a significant effect", with 1151 occurrences). This left us with 1250 trigrams to consider. Two coders (authors of this article) separately coded whether they considered that each of the 1250 trigrams was likely to refer to statistical significance. The two coders reached an agreement of Cohen's $\kappa = 0.74$. We considered that a trigram was likely to refer to statistical significance when both coders agreed it was. This was the case for 676 trigrams out of the 1250. Each of these 676 trigrams was then used as a search term.

In addition to looking for likely uses of significant language using the 676 trigrams, we searched the term "statistically significant" to identify sure occurrences of significance language.

[1] github.com/euske/pdfminer

[2] github.com/nltk/nltk

[3] www.aviz.fr/dichotomous

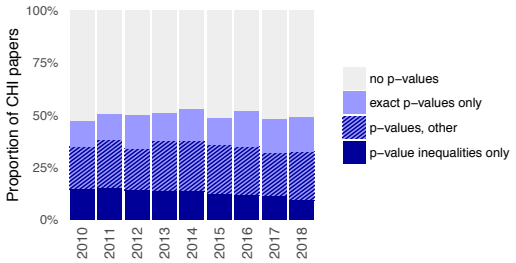

Figure 1: Report of p-values in CHI proceedings from 2010 to 2018.

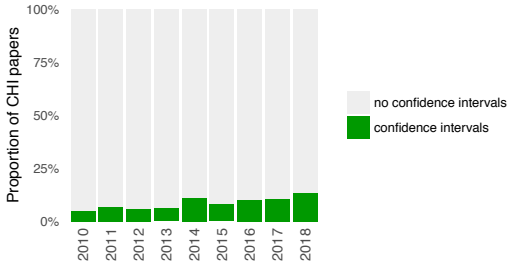

Figure 2: Report of confidence intervals in CHI proceedings from 2010 to 2018.

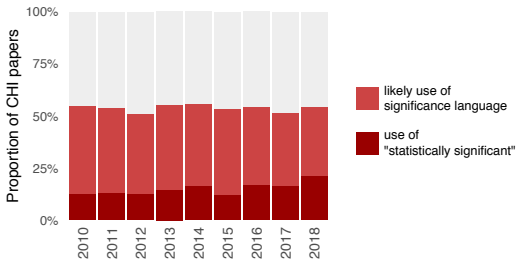

Figure 3: Use of significance language in CHI proceedings from 2010 to 2018.

REPORTING HABITS AND DICHOTOMOUS INFERENCES ACROSS YEARS

The results of our analysis across conference years (2010–2018) are reported in Fig. 1 to Fig. 3.

Fig. 1 shows the proportion of CHI papers that only report p-value inequalities (dark blue, bottom) and the proportion of papers that only report exact p-values (light blue, top). Both categories can contain ambiguous p-value formats (e.g., p<.0001), but all other p-values have to be in the same format. The category "other" (hatched bars, middle) represents papers with p-values that either mix the two formats, or whose status is undetermined because they only contain ambiguous p-value formats. While the proportion of papers reporting p-values seems stable (around 50% of all CHI papers, irrespective of whether they include a user study), the proportion of papers that exclusively report p-value inequalities seems to have decreased from 2010 to 2018. Meanwhile, the proportion of papers that exclusively report exact p-values seems to have slightly increased. This suggests that the recommendation to report exact p-values [31] is being increasingly endorsed at CHI, although the trend is rather modest and most papers report a mix of both.

Fig. 2 shows the proportion of CHI papers reporting confidence intervals, which has also seen an increase from 2010 to 2018 (from 6% to 15%). Thus, while p-values remain largely dominant, there is an increasing attention paid to effect sizes and the uncertainty around their estimates [2, 8, 11].

Fig. 3 paints a less optimistic picture about the prevalence of dichotomous inferences. The bottom bars (dark red) show the proportion of papers using the term "statistically significant", while the top bars (red) show the proportion of papers that are likely to employ other forms of significance language. Overall, the use of significance language is highly common (about 50% of papers) and has remained stable from 2010 to 2018. Nevertheless, the relative proportion of papers using "statistically significant" has been slightly increasing. Methodologists have often deplored that the term "significant" is easily confused with "important", and thus it has been recommended not to omit the term "statistically". It seems that CHI authors have been increasingly following this advice.

Overall, our results suggest that more and more CHI authors are embracing best reporting practices (i.e., reporting exact p-values, reporting interval estimates, and avoiding the term "significant" without the qualifier "statistically"). However, the trends are rather modest, and despite these slow changes in reporting habits, the prevalence of dichotomous inferences as captured by the use of significance language shows no sign of diminishing. It is then fair to assume that the numerous criticisms of dichotomous inference by prominent methodologists have had virtually no influence on CHI.

RELATIONSHIPS BETWEEN REPORTING HABITS AND DICHOTOMOUS INFERENCES

We wanted to examine whether reporting habits (i.e., p-value inequalities, exact p-values, and confidence intervals) have an influence on the use of significant language in CHI papers.

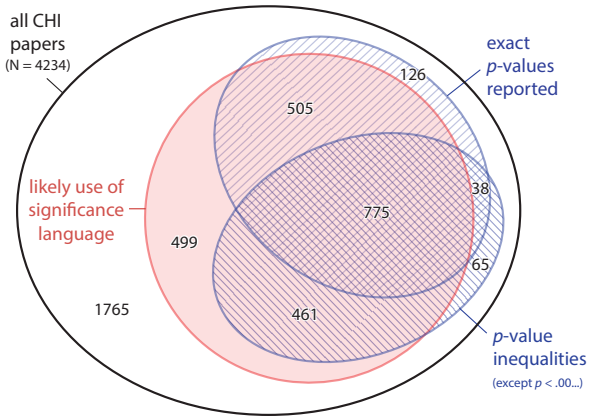

Figure 4: Euler diagram showing the relationship between p-value reporting style and use of significance language in the CHI proceedings from 2010 to 2018. The paper count is provided for each of the 8 mutually exclusive regions in the diagram. Made with EulerAPE [26].

The area-proportional Euler diagram in Fig. 4 shows the relationships between the use of significant language and p-value reporting format. The red ellipse shows the total number of CHI papers that likely employ significance language[4], all years confounded (2010–2018). The bottom hatched ellipse shows the total number of CHI papers that report p-value inequalities, while the top hatched ellipse shows the number of CHI papers that report exact p-values. Papers at the intersection report both. Papers that only report p-values whose format is ambiguous (e.g., $p < .0001$) are not shown.

This diagram confirms what Fig. 1 has already showed, that is, there are about as many papers reporting p inequalities as exact p-values, while the majority of papers report a mix of both. Crucially, most papers we found to be likely to use significance language report p-values, and conversely, the vast majority of papers reporting p-values likely use significance language. Specifically, the likely presence of significant language was found in 88% of papers which only report p-value inequalities, in 80% of papers which only report exact p-values, and in 95% of papers which report both. It would thus seems that the reporting of exact p-values does not help to reduce dichotomous inferences.

Fig. 5 shows a similar Euler diagram that includes data on confidence intervals. The blue hatched ellipse shows all CHI papers that report p-values in any form, while the green ellipse (top) shows all CHI papers that report confidence intervals. Again, significance language seems to be used across the board. However, out of the 22+40=62 papers that exclusively report confidence intervals, only 22 (35%) likely use significant language. Although few papers exlusively report confidence intervals, this trend stands in stark contrast with papers that only report p-values (87% of which likely employ significance

[4]Contrary to the top red bars in Fig. 3, papers containing the phrase "statistically significant" are not excluded here. In fact, many of the trigrams we retained during the coding process contained that phrase. However, since trigrams occurring less than 3 times were not coded, a few papers using the term "statistically significant" (about 6%) were not captured.

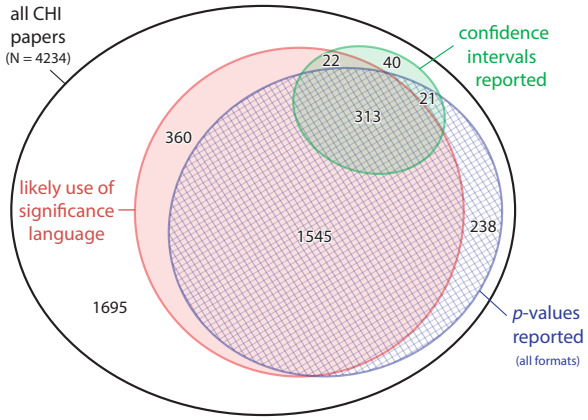

Figure 5: Euler diagram showing the relationship between the reporting of p-values vs. confidence intervals and the use of significance language in the CHI proceedings from 2010 to 2018. The paper count is provided for each of the 8 mutually exclusive regions in the diagram. Made with EulerAPE [26].

language) and papers that report both (94%). It therefore seems that CHI authors who only report confidence intervals are less likely to use dichotomous inferences in their communication, perhaps because they follow reporting principles from estimation statistics (sometimes dubbed "the new statistics") [8, 20], which require to focus on interval estimates and avoid dichotomous interpretations.

To get a sense of why some articles appear to use significance language without reporting statistics, we randomly sampled 10 articles from the uniform pink region in the Euler diagram of Fig. 4 (499 articles). Of the 10 articles, 4 were significance language false positives ("significant" was used colloquially), and 1 was a p-value false negative (it stated a "p-value of 0.0000984 << 0.05"). Of the remaining 5 that were correctly classified, 2 used ambiguous p formatting in that they only reported small p-values (e.g., "p < 0.0005", "p < 0.001"), 2 reported statistical significance without p-values (one reported "statistically significant (α = 0.05)", and the other one reported no numerical information), and 1 employed statistical significance language when discussing related work.

We again randomly sampled 10 articles from the uniform pink region in the second Euler diagram from Fig. 5, which only contains articles for which we found no p-value whatsoever (irrespective of the format) and no confidence interval (360 articles). Of the 10 articles sampled, 4 were significance language false positives, while there was no p-value or CI false negative. Of the remaining 6 that were correctly classified, 1 reported statistical significance without any numerical information, and 5 employed statistical significance language when discussing related work.

DISCUSSION, LIMITATIONS, FUTURE WORK

This work is only a quick investigation of the prevalence of dichotomous inferences at CHI, and it has of course a number of limitations. First of all, our classification of CHI papers into articles that use or do not use significance language is imperfect. Some of the trigrams we rated as likely to refer to statistical significance might lead to correct classifications in some papers, but to false positives in others. For example, we rated "is significant for" as likely to refer to statistical significance, but it led to one of the 4 false positives we identified in the diagram of Fig. 4. At the same time, we erred on the side of caution while classifying trigrams, and we ignored many infrequent trigrams (88% of all 10,334 trigrams, which account for 35% of all occurrences), so our dataset is also likely to contain false negatives. At this point we cannot easily assess the number of false positives and false negatives, and which are the most common. Nevertheless, the remarkable overlap between likely use of statistical language and reporting of p-values shown in Fig. 5 suggests that our classification is reasonably accurate overall. Thus there are reasons to think that the trends we have seen are not overly affected by the presence of false negatives and false positives.

Similarly, a few false positives and false negatives might have occurred in our analysis of statistical reporting formats. In particular, because we analyze only the text of the PDFs, we could not capture statistics that were reported in figures. However, it is reasonable to assume that most papers reporting p-values or confidence intervals in figures also mention them in the text. As for tables, the conversion to text seemed to have preserved table content in most cases, but we cannot ascertain that all tables were correctly converted. Other p-values or confidence intervals might have been missed because they were reported in a non-standard fashion (see, e.g., our previous example of an article reporting "p-value of $0.0000984 << 0.05$"). Conversely, paper authors may discuss confidence intervals or p-values without reporting them, for example in methodological articles. However, due to the prevalence of studies at CHI and the very standardized way of presenting their results, we believe that false positives and false negatives in our analysis of statistical reporting were not too common.

We did not try to distinguish between papers with a user study and papers without: our analyses include all CHI papers without distinction. Depending on the year and on the source, it has been estimated that between 78% and 91% of CHI papers report a user study [4, 6, 19]. Some of these papers focus on reporting qualitative observations and/or descriptive statistics (and are thus beyond the scope of this article), while others report inferential statistics. Our experience is that the overwhelming majority of the latter use frequentist inference (and thus report p-values and/or confidence intervals), while papers employing other methods (e.g., Bayesian inference [18]) represent a tiny minority. Therefore, there are good reasons to believe that the union of the hatched blue and green regions in Fig. 5 (about 50% of all papers) is indicative of the proportion of CHI papers between 2010 and 2018 reporting user studies with inferential statistics.

Overall, we found that the vast majority of CHI papers reporting inferential statistics make dichotomous inferences. Despite modest improvements in reporting habits (e.g., exact p-values are more frequently reported), the prevalence of NHST-based dichotomous inferences appears to have shown no sign of evolution since 2010. Thus the numerous calls for avoiding dichotomous inferences [1–3, 8, 10–12, 15, 17, 25, 29] seem to have had virtually no effect on the CHI community. Reassuringly, a small but increasing minority of papers focus their inferences on confidence intervals, and among these, dichotomous inferences seem less prevalent. However, among papers reporting both confidence intervals and p-values, dichotomous inferences are remarkably common. We also found that significance language is sometimes used to summarize previously published studies, a practice that can possibly oversimplify or mischaracterize the literature [22].

We have only looked at the presence of significance language by searching the terms “significant" and “significantly", but dichotomous conclusions can be made in many other ways, with statements like “we found that task has an effect on performance, but not technique”. By presenting statistically significant results as sure findings or by implicitly accepting the null hypothesis, such statements are also diagnostic of dichotomous inference. Conversely, hedges and terms such as “likely”, ”possibly”, and “evidence” could be indicative of nuanced conclusions, which are recommended to faithfully communicate scientific findings [23, 30, 32] and to give readers the freedom to evaluate evidence and reach conclusions by themselves [27]. Though interesting to study as future work, the extent to which conclusions are binary or nuanced is likely hard to analyze using automated text processing tools.

While there is an overabundance of guidelines on NHST, guidance on how to interpret results in a non-dichotomous manner is harder to find. For advice on how to interpret p-values without using dichotomous language, see the recent blog post by Frank Harrel [16][5]. For advice on how to interpret confidence intervals without using dichotomous language, see Cumming [9] and Dragicevic [11][6].

[5]For examples of studies, see the references in Hurlbert and Lombardi [17] on top of p.314.

[6]For examples of studies in HCI and Vis, see aviz.fr/badstats#papers and aviz.fr/ci/.

ACKNOWLEDGEMENTS

We are grateful to Yvonne Jansen, Mickael Sereno, Xiyao Wang, and Valentin Amrhein for their help and feedback. We also thank the pseudonymous reviewers (Xiaojun Bi, Florent Cabric, Géry Casiez, Andy Cockburn, Geoff Cumming, Jessica Hullman, Theophanis Tsandilas, Chat Wacharamanotham, Shumin Zhai) and anonymous reviewers of our submission.

REFERENCES

[1] Valentin Amrhein, Fränzi Korner-Nievergelt, and Tobias Roth. 2017. The earth is flat (p>0.05): Significance thresholds and the crisis of unreplicable research. *PeerJ Preprints* 5 (June 2017), e2921v2.

[2] Valentin Amrhein, David Trafimow, and Sander Greenland. 2018. Inferential Statistics as Descriptive Statistics: There is No Replication Crisis if We Don’t Expect Replication. *The American Statistician* (2018).

[3] Thomas Baguley. 2012. *Serious stats: A guide to advanced statistics for the behavioral sciences.* Red Globe Press.

[4] Louise Barkhuus and Jennifer A. Rode. 2007. From Mice to Men - 24 Years of Evaluation in CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’07).* ACM, New York, NY, USA, Article 1.

[5] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, and others. 2018. Redefine statistical significance. *Nature Human Behaviour* 2, 1 (2018), 6.

[6] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI ’16).* ACM, New York, NY, USA, 981–992.

[7] Melissa Coulson, Michelle Healey, Fiona Fidler, and Geoff Cumming. 2010. Confidence intervals permit, but don’t guarantee, better inference than statistical significance testing. *Frontiers in psychology* 1 (2010), 26.

[8] Geoff Cumming. 2012. *Understanding the new statistics: effect sizes, confidence intervals and meta-analysis*. Routledge Taylor & Francis Group.

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

[10] Peter Dixon. 2003. The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology* 57, 3 (2003), 189.

[11] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer International Publishing, Cham, Switzerland, Chapter 13, 291–330.

[12] Andrew Gelman. 2017. No to inferential thresholds. Online. Last visited 04 January 2019. (2017).

[13] Andrew Gelman and Hal Stern. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* 60, 4 (2006), 328–331.

[14] Gerd Gigerenzer. 2004. Mindless statistics. *The Journal of Socio-Economics* 33, 5 (2004), 587–606.

[15] Gerd Gigerenzer. 2018. Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science* (2018), 2515245918771329.

[16] Frank Harrell. 2018. Language for communicating frequentist results about treatment effects. Online. Last visited 04 January 2019. (2018).

[17] Stuart H Hurlbert and Celia M Lombardi. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. In *Annales Zoologici Fennici*, Vol. 46. BioOne, 311–349.

[18] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the CHI 2016*. ACM, 4521–4532.

[19] Lisa Koeman. 2018. How many participants do researchers recruit? A look at 678 UX/HCI studies. Online. Last visited 06 January 2019. (2018).

[20] John K Kruschke and Torrin M Liddell. 2018. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review* 25, 1 (2018), 178–206.

[21] Daniel Lakens. 2016. Dance of the Bayes factors. Online. Last visited 03 January 2019. (2016).

[22] Joe Marshall, Conor Linehan, Jocelyn Spence, and Stefan Rennick Egglestone. 2017. Throwaway citation of prior work creates risk of bad HCI research. In *Proc. CHI Extended Abstracts*. ACM, 827–836.

[23] Anna Mauranen. 1997. *Hedging in Language Reviser's Hands*. Vol. 24. Walter de Gruyter. 115 pages.

[24] Alison McCook. 2016. We're using a common statistical test all wrong. Statisticians want to fix that. Online. Last visited 03 January 2019. (2016).

[25] Blakeley B. McShane and David Gal. 2017. Statistical Significance and the Dichotomization of Evidence. *J. Amer. Statist. Assoc.* 112, 519 (2017), 885–895.

[26] Luana Micallef and Peter Rodgers. 2014. eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. *PloS one* 9, 7 (2014), e101717.

[27] Greg Myers. 1989. The pragmatics of politeness in scientific articles. *Applied Linguistics* 10, 1 (1989), 1–35.

[28] Robert Rosenthal and John Gaito. 1963. The Interpretation of Levels of Significance by Psychological Researchers. *The Journal of Psychology* 55, 1 (1963), 33–38.

[29] Jeffrey C. Valentine, Ariel M. Aloe, and Timothy S. Lau. 2015. Life After NHST: How to Describe Your Data Without "p-ing" Everywhere. *Basic and Applied Social Psychology* 37, 5 (2015), 260–273.

[30] Kees van Deemter. 2010. *Not Exactly: in Praise of Vagueness*. Oxford University Press.

[31] Gary R. VandenBos (Ed.). 2009. *Publication Manual of the American Psychological Association* (6th ed.). American Psychological Association, Washington, DC.

[32] Ignacio Vázquez Orta and Diana Giner. 2008. Beyond mood and modality: epistemic modality markers as hedges in research articles. A cross-disciplinary study. In *Revistas - Revista Alicantina de Estudios Ingleses*. Vol. 21. Universidad de Alicante. Departamento de Filología Inglesa.

COMMENTARIES

This section contains the non-anonymous reviews posted during the alt.chi open reviewing process, and for which we obtained the reviewer's consent to add to this author's version.

Xiaojun Bi

This paper is thought-provoking. The authors investigated a common practice in the CHI community: the usage of dichotomous Inferences. The majority of the researchers in CHI take it for granted and heavily rely on dichotomous Inferences to draw conclusions. The authors point out the potential problems of this practice, and refer to the solutions that are practiced in other fields such as psychology and medical research.

I hope this paper would draw the CHI community's attention to the potential problems of using dichotomous Inferences, and promote discussion on how to solve this issue. They mentioned some approaches such as reporting effect size and confidence interval.s I hope these methods will serve as a start point and more methods will be suggested.

To further promote the discussion, it would be great if CHI could organize a panel to discuss this issue, and suggest possible actions CHI researchers can take to alleviate this issue. Maybe one of the authors could serve as the coordinator to invite panelists and moderate the discussion? I look forward to seeing such an event at CHI.

Relationship to Authors: (none specified)

Géry Casiez

First I find this is a very interesting and relevant work for the community addressing important questions that should be more discussed. Overall there should exist more work investigating how the community use and report statistical analyses. The approach followed by the authors to analyze so many papers is impressive. Below I provide some feedback to improve the quality of the paper.

The authors should clarify what they mean by exact p-values. What is an exact p-value? How many decimals are required to be exact? Is it allowed to do some rounding? How? I raise this question because $p < 0.001$ is qualified by the authors as ambiguous. What would it bring to report for example $p = 0.0008967$? I guess statistical software already report rounded p-values at some point.

As the authors state, HCI researchers do not seem to read papers from the statistical community but they can read stats for HCI. One example is the recent book "Modern Statistical Methods for HCI" 10.1007/978-3-319-26633-6. I recommend reading chapter 5 written by Koji Yatani. He says: "Another common misunderstanding is that the p value indicates the magnitude of an effect. For example, someone might say that the effect with $p = 0.001$ is larger than with $p = 0.01$. This is not true. The p value has nothing to do with the magnitude of an effect. The p value is merely the conditional probability of the occurrence of the data you observed given the null hypothesis. It does not give us any information of how large the effect is, and we need another metric." As a result I don't understand why exact p-values should be reported. At least everyone is not on the same line and the authors only expose one point of view, shared by a given number of statisticians. It is not that clear to me that we should report "exact" p-values because the direct consequence is that p-values indicate the magnitude of an effect. After reading chap 5, what can you reply to Koji Yatani?

Koji Yatani also says "Effect sizes and power analysis can mitigate over-reliance on the p value". With that regard the authors talk about confidence intervals but they do not say much about eta-square values, for example,

and how they are reported in papers. Regarding confidence intervals, I do not agree with the following sentence "few papers report p-values or confidence intervals in figures without ever mentioning them in the text". Instead I consider there are many papers that report confidence intervals ONLY in figures.

In summary, very interesting analysis of work published at CHI but I remain to be convinced about the reporting of exact p-values.

Relationship to Authors: (none specified)

Andy Cockburn

The following is my unaltered review for the original submission.

Great work.

One of the things that struck me while reading the paper is the issue who/what contributes to the continuance of reporting dichotomous outcomes? Often, but not always, the choice to report dichotomous outcomes reflects the authors' true desire. Sometimes, however, the author would prefer to NOT report dichotomous outcomes (for good reasons), but is compelled to do so by their fear/knowledge that if not included, reviewers will expect it and criticise its absence (I've certainly succumbed to this in past papers). Other times, the authors stick by their convictions and choose not to brand outcomes as "sig./not sig.", but get beaten up by reviewers for following through with their choice... then, in rebuttal authors can stick with their convictions to not report (which is likely acceptance suicide) or bow to the reviewers' "wisdom" and include it (elevating acceptance probability) – I've fallen victim to this on both sides.

Moving away from dichotomous testing seems to require a step-change from the whole community. It's not practical for individuals to change their practice while the community maintains its expectations... the individuals who move away will simply have their papers rejected. It pretty much requires everyone to agree simultaneously that we've had enough of the approach, and eliminate it in one fell swoop. But this seems unrealistic. Furthermore, there is more than one legitimate change in practice that would improve the situation:

- we might adjust what we mean (as authors and reviewers) with the word "significant" – I could imagine a scale of normative terminology indicating different levels of likelihood of observing the data (or more extreme) if the null were true: "suggestive", "significant", "...";
- we might alter the boundaries at which we assign these terms;
- we might develop more nuanced appreciation of what threshold terms mean (I think probably the key problem with p values is that a shocking proportion don't know the meaning);
- we might make the addition of further data a key part in results interpretation (CIs, effect sizes, etc., as is becoming more common in our field).
- ...

And problematically, while we have more than one possible path for modifying our behaviour as authors and reviewers, the likelihood that any one path will be chosen is commensurately reduced (like the step change required for banning dichotomous reporting).

Relationship to Authors: I have had several conversations with Pierre and I hope to work with his team in the future, but have not done so to date.

Geoff Cumming

It's worth reflecting on how bizarre the dominant logic for drawing conclusions from data is. A result is pronounced 'real' or not, 'nonzero' or not, 'existing' or not. The true size of the effect in question does influence what result is found, but sample size is also highly influential, and this should not be a factor in judging existence. I'm referring, of course, to dichotomous thinking, which usually focuses on statistical significance. Finding $p < .05$ brings joy, publication, funding, and fame, whereas $p > .05$ brings the opposite. NHST has dominated in many disciplines for more than half a century; the method has been justifiably excoriated by leading scholars for just as long. And yet, unaccountably, it persists! Advocacy of much better alternatives, notably estimation (effect sizes and confidence intervals) has become more widespread in recent years, along with the rise of Open Science practices designed to make published research more trustworthy. Moving beyond dichotomous thinking to these better practices would improve research greatly.

The first two pages of this paper review the issue well and explain how dichotomous thinking leads to big problems, especially selective publication, and also cherry-picking and other strategies to find small p values, somehow. We could add that a mistaken faith in statistical significance suggests replication isn't needed. Surprise: very few replications are conducted!

Computer scanning of text has drawbacks, which the authors acknowledge and discuss, but their main findings are strong and justified. It seems, sadly, that in HCI, at least as represented by CHI proceedings, the level of dominance of dichotomous thinking in statistical inference has remained pretty much constant over the last decade or so. There is encouragement in the increasing use of confidence intervals, but from a small base, and still only about 1% of papers report CIs without signs of dichotomous interpretation.

I hope this paper will serve as a wake-up call among HCI researchers and practitioners. Issues for discussion and action should include:

(1) Recognition of the damage that dichotomous thinking has done and is doing.
(2) Discussion of which changes (estimation? Open Science practices? Bayesian techniques?) hold most promise for HCI.
(3) Discussion of which are the most effective strategies for achieving widespread change.

P.S. For illustrations of just one terrible feature of p values, at YouTube search for 'dance of the p values' and 'significance roulette'

Relationship to Authors: I have known one author (Pierre) for many years. We have met a couple of times. We have discussed many things, but we have never published or conducted research together.

Jouni Helske

The paper's goal seem to be two-fold: First, it raises the issue of dichotomous thinking on the table for the CHI community, which is no doubt important goal. The second part of the paper complements the more general discussion with data analysis of past CHI papers, and concludes that the transition to nuanced interpretation of statistical analysis has not been as fast as one could have hoped for (although I am not really surprised by the results).

Overall the paper is well written. The automatic text analysis has likely added some noise the results (for example the sensitivity checks done for some random samples suggest that there were several issues with the data processing and classification), but the authors acknowledge these issues well, and as they are not claiming

dichotomously(!) their results, I don't think these potential issues are crucial in delivering the main messages of the paper.

Some minor issues: There seems to be typo in the description of text mining, one the strings is just "p ", not "p >", but the actual script looks to be using the correct string.

For ambiguous case, were space on both sides of > tested? Definition of the actual threshold seems bit confusing, the text suggests 0.001 but later it is 0.01? And the scripts seem to suggest that 0.01 was used.

Authors write "use frequentist inference (and thus report p-values and/or confidence intervals), while papers employing other methods (e.g., Bayesian inference [15] or likelihood inference)". This is unclear. What is meant by likelihood inference? We can have (and often have) frequentist inference based on maximizing the likelihood function, but on the other hand likelihood has a crucial place in Bayesian inference as well. I assume authors are trying to make distinction between test-statistic based inference (e.g. t-tests) and model based inference (e.g. regression models etc)? Latter often report p-values and confidence intervals as well though.

Relationship to Authors: I work at the same department in Linköping University as the first author, and we have common research interests, but I have not been involved in makings of this paper.

Jessica Hullman

This paper presents the results of an analysis of over 4k CHI papers from 2010 - 2018 used to investigate whethe recommendations among statistical reformers to avoid dichotomous presentation of effects has had an influence in the CHI community. The paper concludes that it has not: 'We wanted to see whether they have had any influence on CHI. Our analysis of CHI proceedings from the past eight years suggests that they have not.'

The classification approach for designating different styles of p value usage as well as significance language were reasonable and clearly described. The random sample of 10 papers from several subsets of the papers indicate that the error rate is non-trivial on some classifications. A useful followup would be to examine each of the 4k papers.

The findings do not indicate evidence of a decrease of dichotomization overall, but do suggest that certain suggested practices within NHST, like exact p value reporting and use of the adjective statistically with significant have increased. Overall these are valuable findings to point out to the CHI committee, and provide a basis for speculation on why researches may be "selectively listening" to advice on statistical reform. I suspect the difficulty for many researchers to learn alternatives to NHST may be one cause.

It occurs to me that the abstract of the paper could allude more to at least one other finding in addition to the lack of perceived evidence that dichotomization is decreasing. It's interesting to think about the statement at the end of the abstract as an example of the nuances of dichotomization – one can still state that there does not appear to be an effect without using p values. It reminds me of how statistical reformers like Gelman have discussed the irrationalit of describing comparions in data anaysis, regardless of what inference paradigm, as indicating that there is "no difference" since in reality there is always some difference, if slight. A discussion of whether dichotomization can occur in more subtle ways outside of NHST would be an interesting follow up to this work.

This paper is an important data point in the transparent statistics movement and stands to call attention to ways that CHI researcher could be more mindful of hte way they report their results. I wholly endorse its acceptance to alt.chi.

Relationship to Authors: (none specified)

Arnaud Prouzeau

This paper presents a review about the use of dichotomous inference in the papers coming from the CHI conference. The authors did a text analysis in the last 8 proceedings of the conference looking for evidence of the use of p-values inequalities, exact p-values and significance language. They found that the use of p-values inequalities is decreasing, the use of exact p-values and confidence interval is increasing over the years. On the other hand, the use of significance language is stable and correlated with the use of exact p-value and confidence interval, meaning that their use didn't prevent the authors to do dichotomous inferences.

Overall, the paper is very well written, the analysis is well done and clearly presented. This is a very important topic in our community and this paper is a step in the direction of its more global understanding.

A global understanding of this issue by the entire community is probably key before seeing an important change in the papers. To start doing it now could avoid in HCI the replication crisis which happened in psychology.

One issue with this paper, which is minor, but if fixed could foster a bigger debate, is the future work, which is too weak in my opinion considering the issue. I was expecting some directions on how to report results without dichotomous inferences? The authors provided a lot of references, but we can't really expect all CHI authors to read them all (They should, but they won't). It would be good to point to one or two guidelines references or a summary of the different options, with concrete examples of reporting. It would also be interesting to know what is done in other fields like psychology and medicine which lived a confidence crisis that led to a change of methodology.

Overall, I strongly support this paper for its quality and the importance of the topic tackled in it. I would like to see a more practical approach of what we should do now, what would be the best practices for result reporting.

Relationship to Authors: (none specified)

Theophanis Tsandilas

The issue discussed by the paper is very relevant to the CHI community and touches a very important problem. I am aware that the authors, in particular the second author, have been working on these problems for several years. Pierre Dragicevic's page on "bad stats" (https://aviz.fr/badstats) has been very influential and has influenced my own way of thinking and my research methods. The alt.chi paper strengthens discussions about problems on how the HCI community reports on statistical results and draws conclusions.

The authors base their analysis on a large volume of past CHI publications. This is a nice contribution by itself. Although the analysis method has limitations (as it is largely based on the accuracy of an automatic language-processing system), I believe that its results are representative of real trends. I also find that this topic is very appropriate for alt.chi. It can generate a lot of discussions that may further encourage the community to rethink both about how we write and how we review papers. For all these reasons, I am very supportive of this work.

I would like to admit that I have been the author of several of those dichotomous-thinking CHI papers. And although I am well aware of the problems discussed by the paper, I find that making the transition to a pure "non-dichotomous" reporting style is not easy (although I try). In this sense, I would like to present my additional thoughts but also criticisms of this work, with the hope that they could further contribute to the discussion:

My personal interaction with HCI researchers is that very few people are even aware of the problems of NHST. I feel that related discussions in the HCI community have started only recently (4-6 years) within a very limited

circle of people. I would be surprised if practices had changed within so little time. That said, I would be very interested in a future study that explores our community' attitudes in more depth. Are HCI researchers really aware of these problems and to what extent? If yes, what are the difficulties of changing research strategies? The textbooks from which people learn about experimental design in HCI, the lack of examples, the reviewing process, the fact that thesis supervisors do not have the time to make the effort to adopt new approaches, confusion about what is correct and what is not, practical problems about how to tackle specific modeling/inference problems?

The authors have been careful in their writing and clearly state "that there is still an ongoing debate on whether dichotomous inferences should be banned." However, I would also expect a short discussion about why there is an ongoing debate (probably less about the use of dichotomous inference and more about the use of p-values). Why is the problem so complex and why do even statisticians not fully agree? The paper cites a statement of ASA's executive director, but I would also like to refer to the ASA's official statement about p-values, which shows a hard compromise of very different points of view, as the note below the paper and the supplementary material indicate (See: https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108)

Concerning the use of the wording "statistically significant." Unfortunately, such wording is very often used as a final "decision-making" statement. However, people often tend to simply use it as a "statistical jargon" to avoid subjectivity, as there is no experience with more nuanced vocabularies, while it is always easy to abuse language. I feel that some authors who use this terminology make an effort to interpret results and make conclusions based on the full set of evidence, including visualizations of the effect, variance, etc. Thus, I am a bit more optimistic, in the sense that the use of dichotomous language at the level of statistics does not always translate into dichotomous language at the level of overall conclusions. But as the paper mentions, this requires further research.

I personally feel that subjectively adapting the language to accommodate difference levels of statistical evidence is not an easy exercise. I am not sure what the best approach to this direction is, and I definitely look for advice and examples on this problem. Cohen's wording for effect sizes was also an attempt to apply nuances with words, but as words became a standard, the approach was widely criticized. I am also familiar with the textbook of Baguley [3], correctly cited by the paper as someone who has criticized dichotomous thinking. However, the term "statistically significant" is very often used (although not consistently) by the author for many of his examples throughout the textbook – the textbook would have been clearly classified into the category of "combining CIs + p values + dichotomous thinking" by the paper's analysis. The exercise can be especially difficult when reporting on ANOVA models (largely used in HCI), e.g., when summarizing overall effects or interactions. (Some would argue for abandoning such methods altogether, but it seems to me that there is no consensus on such a direction.) I think that the best approach might be to report statistics (e.g., in a table) without even trying to individually characterize their magnitude with subjective statements or standardized words.

I have some additional minor questions about the analysis:

The discussion section mentions that "probably few papers report p-values or confidence intervals in figures without mentioning them in the text." I would be more careful about this statement. I have seen (and used in the past) figures showing "statistical significance" on top of graphs with lines and stars. Also many figures show confidence intervals without even mentioning what they are.

I think that CIs are very commonly used to estimate individual means (often incorrectly constructed over the full set of repeated trials or observations) and less often for estimating differences. However, dichotomous statements are usually relevant to the second only case. To what extent is the observation that CIs lead less to

dichotomous thinking than p-values due to this fact? I am curious about whether the approach could differentiate between such types of uses.

ASA seems to suggest complete report of p-values, although it is still unclear what level of precision is enough. I think that this issue requires more careful examination, but this is certainly out of the scope of the present submission. However, most HCI researchers seem to follow APA's recommendations for reporting statistics (they were suggested to me by several reviewers in the past), which advise authors to report p values lower than .001 as $p < .001$. Thus, I find that the author's decision to ignore these cases from the analysis (as ambiguous) is a bit strict, given that the papers followed what was considered as a "good style of writing."

Relationship to Authors: I am in conflict with the second author, as we work at the same research institution, although we are not in the same team. We have collaborated together and have very frequent discussions about statistics.

Chat Wacharamanotham

The prevalence of dichotomous inference is known anecdotally among those who are concerned about statistical practices in the CHI community. This paper *quantifies* this phenomenon and show how it developed over time. Having this information will allow us—as a community—to determine whether we are willing to live with this with dichotomous inference or to move forward to adopt statistical methods that allow more nuanced understanding of the data.

The methodology is sound. I praise the authors of acknowledging the limitations of their methodology and sampled some data to estimate the impact of the limitations.

What can this paper improve:

To me, a take-home message of this paper is that "We tell people to report confidence intervals, and they increasingly report them. But their interpretation is are still focusing on p-values." The results in this paper indicate that the community needs to double down in educating authors and reviewers how to interpret beyond statistical significance. To this end, I think that our community still lacks educational resources for teaching how to interpret (and, for reviewers, to evaluate the interpretations) beyond dichotomous. I wished the authors could have pointed readers who are interested in learning to some resources, both from the field of statistics (e.g., Cumming's book) and exemplars in our field (e.g., one of the authors' collection of paper at "badstats" website).

Possible directions for future work

Let me mebe a devil's advocate: Despite the evidence presented in this paper, it is unclear how harmful the dichotomous interpretation is to CHI research. It could be that most of the p-values are so low and their effect sizes are so high that even if we change the interpretation to continuous, it wouldn't matter. I think it would be awesome if in the future we can say that X% of the papers would have a different conclusion if we change from dichotomous interpretation to continuous interpretation.

A missed opportunity is the absent of a reference point to compare CHI to other fields. Is CHI doing better or worse compared to a more rigor field (such as psychology) or neighbor fields (such as VIS)? I think that a comparison across fields is an interesting future direction.

Thoughts on dichotomous inference and incentive structure of the review process:

Dichotomous inference is easier for both the reviewers and the authors. This criterion is clear-cut and established. Assuming that there is a credible effect in the results, an author would feel safe to use dichotomous inference. At least, he can be sure that he won't get a disagreement from the reviewers. Since the authors don't

know whether their reviewers are knowledgable in non-dichotomous inference (and since it is expectable that the majority of the reviewers aren't knowledgable in this), it would be game-theoretically sensible for the authors to actually use the dichotomous inference.

The reviewing game is a single-shot game (in the sense of game theory). The next time that an author submits a paper, he will have another set of reviewers. Even if he learned to use continuous inference from one submission, the incentive structure is against him to use a continuous inference in the next submission.

One way to break this game is to provide a channel to communicate information. I could imagine that reviewers can tick a set of checkboxes in PCS to specify their statistical expertise. Each submission metadata can also contain an optional field that specifies the statistical methods used. In the reviewing process, in addition to the usual set of reviewers, one may add a statistical reviewer to weigh in on the inferential method used in the paper (without commenting on other parts). This mechanism could help alleviate anxiety from the side of authors. It could break the current incentive structure that promotes dichotomous interpretation as a "safe way".

Relationship to Authors: I previously co-authored with Pierre, and he was also in my PhD committee.

Shumin Zhai

The authors bring up (again) a really important topic in HCI research. We have to accept this paper and its key messages (I would have recommended accepting them as a paper but alt.chi is fine too). We as a community should use this as another opportunity to improve our research analysis practice.

Dichotomous statistics, NHST, or inferential statistics in general are convenient, relatively easy, simplified and often misleading standardization of evidence-based logic and acceptance criteria. They had been criticized from the beginning of these methods' existence to no avail. Alternative methods have always ran into similar challenges. Personally I think we should move to descriptive statistics and visualization that can give the reader an accurate reading of the data distribution in the context of the research question or system design, and in comparison to other relevant variables, designs, and individual differences. The challenge is that reviewers and readers will have to be sophisticated enough to assess the strength of research findings without the convenience of applying a common "community standard" in NHST.

In the more theoretical fields, for example in psychology, the traditional thinking is that the author can "prove" a theoretically motivated hypothesis or a conceptual construct if the effect of that construct in an experiment is statistically significant - meaning better than chance due to sample size, or individual difference or any sort of noise in measurements, based on a pre-set criterion (e.g. $p < 0.05$ in a repeated measure F test). The absence of meeting such a criterion means either the construct/hypothesis was misconceived, or the experiment was not well designed, therefore there would be nothing worth reporting (as a footnote, this is not necessarily true. A well reasoned hypothesis that does not pan out empirically often is very informative. Furthermore discouraging "null results" publication biases meta-studies later on the field).

The use of dichotomous inferences is even more problematic in HCI. In (many parts of) HCI, we are not often testing a single or simple theoretical construct. Accurately (confidently) knowing how similar in usability of an UI is to another can be just as valuable than rejecting the Null Hypothesis that NHST is designed to do (another footnote: it is very jarring to see so many HCI papers listing their good research questions as hypotheses, often multiple incoherent hypotheses in one study altogether without a priori reasoning).

As a very first step, reporting exact p values and removing languages such as "statistically significant" seemed very practical to me. Before computer programming and software became widely available, researchers had to

look up p values in statistical tables which had to be segmented in the value range. With software, the p values can be as exact as needed. But stop at thousandths decimal place (e.g. p = 0.002) in reporting would appear more modest in implying the precision of the measurement.

Definitely remove statements such as p < 0.05 or "statistically significant". Let's do that! Let's push back on those naive reviewers who demand < signs and significance languages they learned in textbooks (or worse yet, only saw and heard of them but never really seriously understood or reasoned along with textbooks).

PS. Don't know what presentation format is most effective, perhaps inviting a commentator / discussant?

Relationship to Authors: nothing more than knowing one of the two authors for a long time.

Authors' response

We are very grateful to our reviewers for their time and their excellent feedback. Many reviewers (including anonymous reviewers not quoted above) ask for examples and guidance on how to interpret results without using dichotomous significance language. In this new revision we provide pointers in the last paragraph of our conclusion, both for interpreting p-values and confidence intervals.

Many of our reviewers agree that dichotomous inferences are a problem, but there is also some skepticism. **Géry Casiez** remains unconvinced that exact p-values should be reported, and argues that "the direct consequence [of doing so] is that p-values indicate the magnitude of an effect." By effect magnitude most people mean the point estimate of the population effect size, which often corresponds to the sample effect size (e.g., a difference in sample means, or the sample Cohen's d). This point estimate is a "best bet" of the population effect size. Although we fully agree that p-values say nothing about the most likely population effect size, we disagree that reporting exact p-values implies endorsing this misconception. p-values simply capture something different, most often the strength of evidence against the hypothesis that the population effect size is exactly zero – which does say something about effect magnitude, although many would admit it is rarely what we want to know. Regardless, this strength of evidence is on a continuous scale and this is what justifies the reporting of exact p-values.

We would like to clarify that by "exact p-value" we follow common terminology, which does not imply that p should be reported with arbitrarily high precision (see also comments from **Theophanis Tsandilas** and **Shumin Zhai**). Different sources recommend different levels of precision, and although we think that p-values are so noisy that a single significant digit should suffice, how exactly p-values should be reported is a topic beyond the scope of our paper. Only relevant to our paper is that many recommend reporting very small p-values as inequalities, which introduces ambiguities in our analysis. This is because we do not know whether a statement such as $p<.001$ stems from an "exact p" reporting practice, or indicates dichotomous inference with $\alpha=.001$, or alternatively (and perhaps more plausibly) categorical inference (e.g., reporting $p<.001$ as *** or very highly significant, $p<.01$ as ** or highly significant, $p<.05$ as * or significant, and $p>.05$ as non-significant). To us, this type of categorical inference barely differs from dichotomous inference, but we should have perhaps covered it more explicitly.

Some reviewers agree that dichotomous inferences are misguided but are uncertain to what extent they pose a problem. According to **Theophanis Tsandilas**, it remains to be seen if dichotomous inferences are really harmful in practice, since CHI authors may use significance language to please reviewers while remaining aware that they are an oversimplification. Presumably, such authors would still strive to convey the full complexity of their results, and would end up drawing nuanced conclusions. We remain unconvinced that this represents a common case. Our experience with CHI papers is that many authors completely fall into dichotomous thinking, and among authors who do not, many present their findings in a way that can easily lead their readers to do

so. Similarly, **Chat Wacharamanotham** points out that dichotomous inferences may not be so problematic if most CHI papers report very small p-values and large effect sizes. This is an excellent point but again, from our experience, this is very far from representing the majority of CHI papers, where many reported p-values are in the same order of magnitude as (and often close to) .05. We do agree that it would be very informative to look more systematically at the distribution of p-values and effect sizes in the CHI literature.

Theophanis Tsandilas asks the important and difficult question of why is there still an ongoing debate about dichotomous inference. We think status quo bias and motivated reasoning are one reason, and these tend to produce weak arguments in defense of dichotomous inference. But we think there may also exist sophisticated arguments in favor of dichotomous inference. In all honestly, we are currently unable to cite a paper that offers such arguments, but we have not gone through all of the literature. There definitely are excellent researchers who endorse dichotomous testing. We already mentioned the proposal to redefine statistical significance by changing the α threshold to .005, a paper with many authors. A response to this paper, also written by excellent methodologists, suggests to use custom α thresholds defined and justified ahead of time. Any use of α implies an endorsement of dichotomous inference, but the authors do not explicitly engage with arguments against dichotomization of evidence. Some prominent Bayesian statisticians and methodologists are comfortable with dichotomous testing, while others prefer to stay away from it. We are fairly certain, however, that no serious statistician or methodologist defends NHST, i.e., the most prevalent way of doing dichotomous inference that involves an incoherent mixture of Fisher and Neyman-Pearson methods. We do not think there is any serious debate about NHST-based dichotomous inferences specifically. If there is any dichotomous inference method that makes sense, it is not NHST.

We did not discuss possible causes and solutions to the continued prevalence of dichotomous inferences at CHI, but we enjoyed reading the reflexions offered by our reviewers. Concerning possible causes, we fully agree with **Andy Cockburn**, **Theophanis Tsandilas** and **Chat Wacharamanotham** that the incentive structure in the current publication system encourages dichotomous inferences, since many reviewers demand them and it is much safer to comply. Our reviewers also point out that dichotomous inferences are popular because they are easy. They are simple and objective decision making rules, even though as Andrew Gelman eloquently puts it, they are nothing more than an "uncertainty laundering" machine meant to "create a sense of certainty where none should exist". Not relying on mechanical decision rules and using our judgment is naturally harder. **Jessica Hullman**, **Theophanis Tsandilas** and **Chat Wacharamanotham** correctly point out at there is currently not much guidance for doing so: there is a lack of educational resources, and there is no simple recipe to follow. We however think that writing nuanced and non-dichotomous interpretations is not as difficult as many of our reviewers imply. Perhaps it is intimidating and unfamiliar only because researchers got addicted to the false certainty provided by NHST. Many papers have been already published that interpret results without resorting to significance language, for example using a so-called "NeoFisherian" interpretation of p-values, or using a "new statistics" interpretation of interval estimates (see the last paragraph of our paper). Actually, we think that *any* subjective interpretation of statistical results is fine as long as results are reported clearly and readers can judge the evidence by themselves. When this is the case, we do not think subjective language can abuse readers as **Theophanis Tsandilas** seems to fear. Only false objectivity can do this.

Our reviewers offer excellent suggestions for possible next steps. We are very sympathetic to **Shumin Zhai**'s sentiment that we should be bold, ban the use of statistical significance language in our papers, and push back against reviewers. Meanwhile, we also agree with **Andy Cockburn** and **Theophanis Tsandilas** that a step

change needs to happen before authors can safely do away with statistical significance, and that such a transition will not be easy. **Chat Wacharamanotham** has good ideas on how to accelerate the transition by better educating authors and reviewers, and by improving the ways statistics are reviewed. We also like **Geoff Cumming**'s call for more methodological discussions within the community and **Xiaojun Bi**'s suggestion to organize a panel at CHI (so far only workshops and SIG meetings have been organized, see transparentstatistics.org). At the same time, we also agree with **Theophanis Tsandilas** that we should not expect rapid changes, since such methodological discussions are relatively recent at CHI and they have been involving a relatively small circle of people.

What should replace dichotomous inference? Both **Andy Cockburn** and **Theophanis Tsandilas** wonder whether a normative terminology with graded degrees of evidence (presumably used to interpret p-values) could replace the normative terminology of statistical significance. Such a terminology would be fascinating to consider but as **Theophanis Tsandilas** points out, canned interpretations like the ones suggested by Cohen for his d metric tend to be abused once widely adopted. How strong or weak evidence is often depends on the context, and a badly calibrated or overly strict normative terminology can do more harm than good. As **Geoff Cumming** suggests, moving away from p-values (e.g., using frequentist or Bayesian estimation) could make it easier to avoid dichotomous inferences, and can cure us from our obsession of disproving the null (see **Shumin Zhai**'s comment on why this should not be HCI's sole focus). But as **Jessica Hullman** recalls and as we have stressed in our introduction, dichotomization of evidence can easily occur without p-values. We did observe that CHI papers that report confidence intervals without p-values tend to rely less on dichotomous inferences. **Theophanis Tsandilas** has an interesting explanation of why this may be the case. We however think that the main reason is the recent popularity of the "new statistics" or "estimation statistics" philosophy, which simultaneously encourages the reporting of interval estimates and discourages dichotomous interpretations. It is definitely possible to derive non-dichotomous conclusions from p-values (see the last paragraph of our paper), but it just happens that papers promoting this approach have not reached the popularity of papers about estimation statistics.

Ultimately, we wholeheartedly agree with **Theophanis Tsandilas**'s recommendation to focus on reporting results clearly and faithfully, perhaps with a particular emphasis on detailed descriptive statistics as suggested by **Shumin Zhai**. This is in agreement with suggestions from several methodologists to give the reader the freedom to reach their own conclusions. We also fully agree with **Geoff Cumming** that full transparency can only be achieved if open science practices are adopted.

We thank our reviewers for their excellent suggestions for follow-up work. In particular **Theophanis Tsandilas** for suggesting to explore our community's attitudes in more depth, and **Chat Wacharamanotham** for suggesting to compare CHI with other fields. This paper is only a first step. It does not go deep into details but as **Arnaud Prouzeau** points out, it provides an initial overview of the problem for a CHI audience and offers useful pointers for people who want to learn more. We also agree with **Jouni Helske**, **Jessica Hullman**, and **Theophanis Tsandilas** that the errors in our automated analysis are not trivial, and a follow-up study with more reliable classification or coding methods would be an excellent goal to pursue.

Finally, we thank our reviewers for pointing out minor issues. We corrected the typo mentioned by **Jouni Helske**. Concerning ambiguous p-value formats, $p<.001$ and $p<.0001$ were only examples of occurrences: a $p<X$ inequality was considered ambiguous *iif* $X<0.01$. Concerning likelihood inference ("likelihoodist" is more correct), Chapter 1 of Sober's book "Evidence and Evolution" has an excellent introduction. We however removed our mention of likelihoodist inference, because it is likely very uncommon at CHI. We agree with **Géry Casiez** and **Theophanis Tsandilas** that many papers report statistics only in figures, so we reworded our sentence.