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, 10 pages. https://doi.org/10.1145/3290607.3310432

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

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 [2, 3, 8, 10, 11, 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

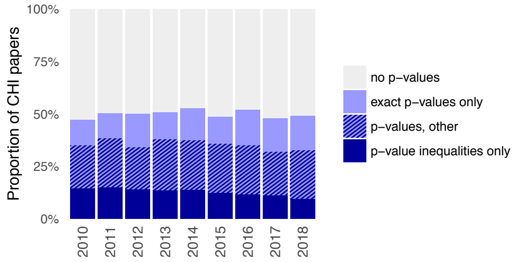

Figure 1: Report of p-values in CHI proceedings from 2010 to 2018.

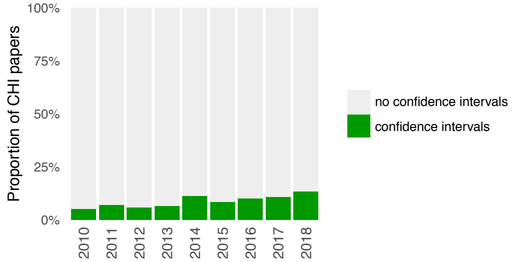

Figure 2: Report of confidence intervals in CHI proceedings from 2010 to 2018.

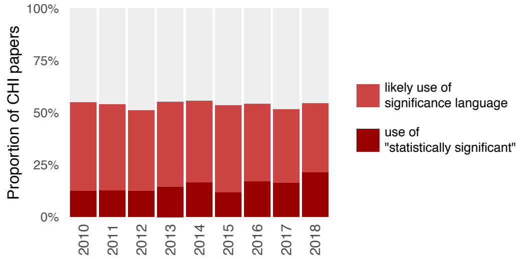

Figure 3: Use of significance language in CHI proceedings from 2010 to 2018.

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.

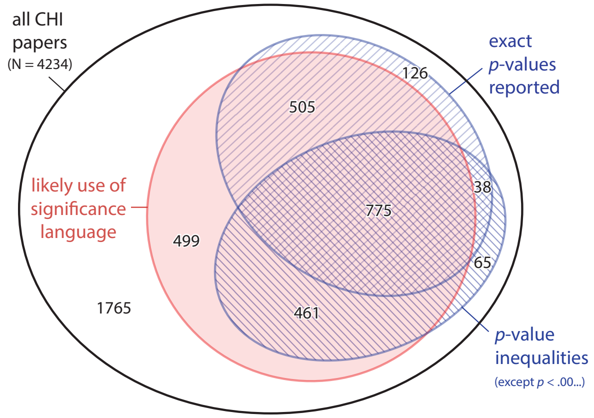

Figure 4: Euler diagram showing the relationship between p-value reporting style and use of significance language in the CHI proceedings from 2010 to 2018. The paper count is provided for each of the 8 mutually exclusive regions in the diagram. Made with EulerAPE [26].

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.

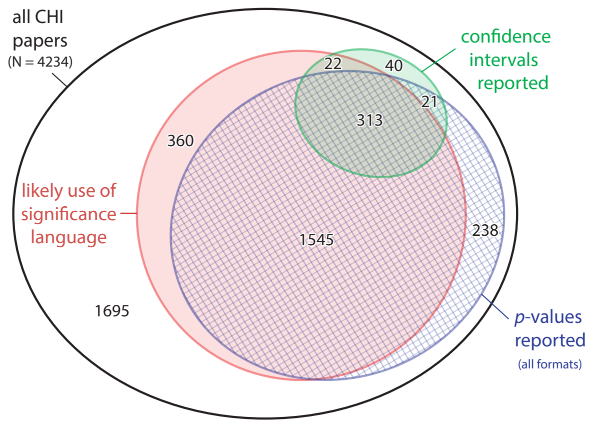

Figure 5: Euler diagram showing the relationship between the reporting of p-values vs. confidence intervals and the use of significance language in the CHI proceedings from 2010 to 2018. The paper count is provided for each of the 8 mutually exclusive regions in the diagram. Made with EulerAPE [26].

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 wish to 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.

Commentary

For alt.chi paper
The Continued Prevalence of Dichotomous Inferences at CHI

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

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).