as expected from the classical definition of probability. In contrast, under $H_1$ most *P* values were concentrated leftward (**Fig. 1c**), even though only 4,812 of them were declared positive. Beyond model validation, *P* value histograms can also provide an estimate of the proportion of truly null ($\pi_0$) or truly alternative ($\pi_1$) hypotheses among multiple tests, based on the balance between features found respectively on the left or right side of *P* value histograms[3]. Analyzing empirical distributions of *P* values can thus provide much insight into the structure of data sets comprising a large number of tested features.

In summary, the *P* value provides valuable information if interpreted correctly. Replacing *P*-value-based decisions with CI-based ones is delusive, as the results will be identical. Instead of dismissing the *P* value by itself, we should avoid any blind trust in pre-established thresholds, and this should hold for CIs as well. Indeed, there is no obvious justification for setting the confidence to 95% rather than 99% or 92.35%. A more suitable approach is to promote the interpretation of the *P* value as "a continuous variable to aid judgment," as originally proposed by Ronald Fisher and relevantly quoted by Halsey *et al.*[1].

As in the American Statistical Association's recent statement[4], the *P* value should be taken for what it is—no more, no less. It indicates only the probability of obtaining—under $H_0$—a result at least as extreme as the observation. As such, it is used to control the risk of false positives, and thereby ensure specificity, but its role has never been to measure the strength of an effect under the alternative hypothesis, or to ensure experimental reproducibility or achieve a desired power.

Recently the editors of the journal *Basic and Applied Social Psychology* announced their decision to ban *P* values, hypothesis testing and CIs[5]. This is in my opinion the worst way to address the situation: renouncing estimation of the risk of false positives would delegate the interpretation entirely to a subjective evaluation of the importance of the observation. Moreover, this is impractical in high-throughput biology, where a single analysis encompasses thousands (transcriptome analysis), millions (genome-wide association studies) or billions (similarity searches in sequence databases) of tests.

Rather than banning the *P* value, let us tame it: promote its understanding and appropriate use, combine it with unit-based statistics (e.g., effect size or CI), define relevant controls, estimate the robustness of the results via replicated experiments, or, when these are not possible, use resampling tests.

**Code availability.** The R library stats4bioinfo (**Supplementary Data 1**) and the script used to produce the results and **Figure 1** (**Supplementary Data 2**) are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

**ACKNOWLEDGMENTS**
I am grateful to G. Lima-Mendez for detailed corrections of the successive versions of the manuscript, and for numerous suggestions that contributed to its improvement. I also thank M. Zytnicki, L. Spinelli, P. Rihet, D. Puthier, A. Saurin, G. Nuel and A. Stevens for constructive comments and discussions.

**COMPETING FINANCIAL INTERESTS**
The author declares no competing financial interests.

**Jacques van Helden**

Aix-Marseille Univ, INSERM, TAGC, Marseille, France.
e-mail: Jacques.van-Helden@univ-amu.fr

1. Halsey, L.G., Curran-Everett, D., Vowler, S.L. & Drummond, G.B. *Nat. Methods* **12**, 179–185 (2015).
2. Open Science Collaboration *Science* **349**, aac4716 (2015).
3. Storey, J.D. & Tibshirani, R. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
4. Wasserstein, R.L. & Lazar, N.A. *Am. Stat.* **70**, 129–133 (2016).
5. Trafimow, D. & Marks, M. *Basic Appl. Soc. Psych.* **37**, 1–2 (2015).

**Halsey *et al.* reply: van Helden argues an old point: that the mathematics underlying *P* and confidence intervals (CIs) are the same, and thus the two variables give the same information. But in our original Commentary, although we offered CIs as an alternative, we specifically mentioned other options. Our paper was not about CIs but about the fickleness of *P*, and having criticized *P* we wished to broach other, arguably better analysis methods that readers might consider. Further, although CIs could be used to make *P*-value-like threshold decisions as we acknowledged in our paper, this would be an unfortunate application. In other words, the point is missed that our 'suggestion to use CIs' is really a suggestion to focus data interpretation on the size of the estimated effect rather than on whether the results are 'significant' or 'not significant'. With the focus on the effect size, CIs provide a way to assess the 'margin of error' around that effect estimate. This approach to data analysis moves things away from significance testing, and that is our main recommendation.**

**van Helden discusses other reasons for variability in the *P* value. However, the simulation we conducted (and which he repeated) in fact avoids all but one of the problems in real experiments. Because we used a theoretical 'perfect' set of data, we were studying merely the inability of insufficient samples to yield representative results—so *P* is fickle even when an experiment is 'perfect'. The problem with running the test many times is that this virtually never happens in practice. With these simulations we become 'all-seeing' about how experimental results can pan out. Real life gives only one chance at a study, and the fickleness of *P* indicates that whether we end up with a winning or losing hand has much to do with luck.**

**COMPETING FINANCIAL INTERESTS**
The author declares no competing financial interests.

**Lewis G Halsey[1], Douglas Curran-Everett[2,3] & Gordon B Drummond[4]**

[1]Department of Life Sciences, University of Roehampton, London, UK. [2]Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado, USA. [3]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado, USA. [4]University of Edinburgh, Edinburgh, UK. e-mail: l.halsey@roehampton.ac.uk