no other statistic fills this particular niche. Moreover, the alternatives (such as estimates, confidence intervals, false discovery rates, etc.) are also subject to random variation and, like *P* values, can behave badly if experiments are poorly designed or implemented. Nonetheless, we do not suggest that researchers rely on *P* values alone. Parameter estimates and confidence intervals, in particular, can describe data in a more detailed, contextual way. The *P* value gained its unique prominence because it is simple and interpretable across a variety of settings, despite the fact that it is sometimes misunderstood. *P* values are variable, but this variability reflects the real uncertainty inherent in statistical results. Thus, we believe *P* values will continue to have an important role in research, but an explicit understanding of *P*-value uncertainty can improve their interpretation.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3741).

ACKNOWLEDGMENTS
We thank the reviewers for their helpful comments and suggestions. This work was supported by the US National Institutes of Health (grant MH086135 to L.C.L. as part of the Consortium on the Genetics of Schizophrenia (COGS)) and the Cooperative Studies Program of the US Department of Veterans Affairs.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Laura C Lazzeroni[1], Ying Lu[2,3] & Ilana Belitskaya-Lévy[2]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California, USA. [2]Cooperative Studies Program Palo Alto Coordinating Center, Department of Veterans Affairs, Mountain View, California, USA. [3]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to L.C.L. (lazzeroni@stanford.edu).

1. Halsey, L.G., Curran-Everett, D., Vowler, S.L. & Drummond, G.B. *Nat. Methods* **12**, 179–185 (2015).
2. Goodman, S.N. *Stat. Med.* **11**, 875–879 (1992).
3. Cumming, G. *Perspect. Psychol. Sci.* **3**, 286–300 (2008).
4. Boos, D.D. & Stefanski, L.A. *Am. Stat.* **65**, 213–221 (2011).
5. Lazzeroni, L.C., Lu, Y. & Belitskaya-Lévy, I. *Mol. Psychiatry* **19**, 1336–1340 (2014).
6. Nuzzo, R. *Nature* **506**, 150–152 (2014).

Halsey *et al.* reply: We agree with Lazzeroni *et al.* that researchers often believe *P* values are infallible[1]. If the intervals Lazzeroni *et al.* propose were obligatory with each presentation of *P*, the unthinking use of the unqualified *P* value would be undermined[2]. In theory this would be an excellent outcome.

However, in practice, simply providing tools for quantifying the fickleness of *P* will highlight an endemic problem without offering any treatment. Whereas Lazzeroni *et al.* suggest providing information to support *P*, we have suggested using measures that supersede *P* for interpreting data[3,4]. Effect sizes can be standardized, are not based on dichotomous decision making (the flaws of which severely limit the value of statistical power[5]) and address the more natural research question of how big the effect is, rather than simply asking whether there is an effect[3,6]. And 95% confidence intervals for the effect size provide a more consistent indication of the true (population-level) condition than does *P*. Thus comparing the effect sizes and confidence intervals of several similar studies typically uncovers a coherent pattern that is masked when only the *P* values of those studies are compared[2]. Furthermore, and crucially, the sample effect sizes and confidence limits of multiple studies can be combined for meta-analysis, enabling researchers to home in on the true effect.

Although we do not encourage the use of power analysis, Lazzeroni *et al.*'s figure supports our own illustration of the variability in *P*. As both our models[7] and Lazzeroni *et al.*'s models demonstrate, unless the results of an experiment show a very marked pattern in the data, the reported *P* value will be accompanied by limits so broad as to render *P* uninterpretable. Put simply, *P* is untrustworthy unless the statistical power is very high (above 90%), which offsets advantages of *P* such as its simplicity. As researchers better appreciate the typically artificial nature of the null hypothesis[3] and the limited capacity of *P* to support hypothesis testing, we believe that *P* will become much less highly valued.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Lewis G Halsey[1], Douglas Curran-Everett[2,3], Sarah L Vowler[4] & Gordon B Drummond[5]

[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]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [5]University of Edinburgh, Edinburgh, UK.
e-mail: l.halsey@roehampton.ac.uk

1. Open Science Collaboration *Science* **349**, aac4716 (2015).
2. Cumming, G. *Perspect. Psychol. Sci.* **3**, 286–300 (2008).
3. Cohen, J. *Am. Psychol.* **49**, 997–1003 (1994).
4. Johnson, D. *J. Wildl. Mgmt.* **63**, 763–772 (1999).
5. Rosnow, R. & Rosenthal, R. *Am. Psychol.* **44**, 1276–1284 (1989).
6. Nakagawa, S. & Cuthill, I. *Biol. Rev. Camb. Philos. Soc.* **82**, 591–605 (2007).
7. Halsey, L.G., Curran-Everett, D., Vowler, S. & Drummond, G. *Nat. Methods* **12**, 179–185 (2015).

Estimation statistics should replace significance testing

To the Editor: For more than 40 years, null-hypothesis significance testing and *P* values have been questioned by statistical commentators, their utility criticized on philosophical and practical grounds[1]. Luckily, the preferred statistical methodology is accessible with modest retraining. An obstacle to the adoption of this alternative seems to be the lack of a widely used name; we suggest the term 'estimation statistics' to describe the group of methods that focus on the estimation of effect sizes (point estimates) and their confidence intervals (precision estimates). Estimation statistics offers several key benefits with respect to current methods.

Estimation is an informative way to analyze and interpret data. For example, for an experiment with two independent groups, the estimation counterpart to a *t*-test is calculation of the mean difference (MD) and its confidence interval[2]. One calculates the MD by subtracting the mean for one group from the mean for the other, and its confidence interval falls between MD – (1.96 × SEMD) and MD + (1.96 × SEMD), where SEMD is the pooled standard error of the MD[3]. For quantitative science, it is more useful to know and think about the magnitude and precision of an effect than it is to contemplate the probability of observing data of at least that extremity, assuming absolutely no effect. An old joke about study-