Why Hypothesis Tests Are Essential for Psychological Science: A Comment on Cumming

Richard D. Morey[1], Jeffrey N. Rouder[2], Josine Verhagen[3] and Eric-Jan Wagenmakers[3]

[1] University of Groningen

[2] University of Missouri

[3] University of Amsterdam

Word count: 896

Address Correspondence to:

Richard D. Morey

Grote Kruisstraat 2/1

9712TS Groningen, The Netherlands

r.d.morey@rug.nl

Author note:

This work was supported by the starting grant "Bayes or Bust" awarded by the European Research Council, and National Science Foundation grants BCS-1240359 and SES-102408.

Abstract

Psychological Science recently announced changes to its publication guidelines (Eich, in press). Among these are many positive changes that will increase the quality of the scientific results published in the journal. One of the changes emphasized by Cumming (in press) is an increased emphasis on estimation, as opposed to hypothesis testing. We argue that estimation alone is ill-suited to science, in which testing predictive models of phenomena are a key goal. Both estimation and hypotheses testing methods are essential.

Why Hypothesis Tests Are Essential for Psychological Science: A Comment on Cumming

In a much anticipated move, Psychological Science announced important changes in its publication practices (Eich, in press); specifically, the changes promote open science (e.g., open data, open materials, and preregistration) and recommend that researchers report confidence intervals instead of *p*-values. Inference by confidence interval, it is argued, constitutes a "new statistics" in psychology (Cumming, in press; Grant, 1962) and avoids the pitfalls of null hypothesis significance testing. To embrace the new statistics, Cumming recommends a shift to estimation as "usually the most informative approach," and states that "interpretation [of data] should be based on estimation."

We broadly agree with the recommendations for more quantitative thinking, more openness and transparency, and an end to *p*-values. Nonetheless, the benefits of estimation have been overstated, and the mistaken idea that estimation is superior to hypothesis testing becoming the conventional wisdom in psychology. This conventional wisdom is perpetuated by the APA Manual, in the new statistical guidelines for journals of the Psychonomic Society, the Society for Personality and Social Psychology Task Force on Publication and Research Practices, and now also by Psychological Science. However, estimation alone is insufficient to move psychology forward as a science; proper hypothesis testing methods are crucial.

Scientific research is often driven by theories that unify a diverse number of previous observations and make clear predictions. For instance, in physics one might test for the existence of the Higgs boson (Low, Lykken, & Shaughnessy, 2012). In biology, one might compare various phylogenetic hypotheses about how species are related (Huelsenbeck & Ronquist, 2001). In psychology, one might test whether fluid intelligence can be improved by training (Harrison et al., 2013).

Testing hypotheses is not as simple as looking at data to see whether they agree with the theory. There are three necessary components to testing a theory: first, one must know what one would expect of the data if the theory were true; second, one must know what one would expect if the theory were false; and third, one must have a principled method for using the data to make an inference about the theory. The second and third components are crucial. Inferring support for a theory on the sole basis of agreement between the observations and the theory is a logical fallacy (known as the converse error; Popper, 1959); having no principled method for inferring support leaves one with only ad hoc rules subject to one's own biases.

The difficulties inherent in using estimation to test theory can be illustrated by the example chosen by Cumming (in press, p.15). Velicer et al. (2008) tested a theoretically-motivated model of smokers' readiness to stop smoking. The authors predicted the strength of association between 15 inventory scales and the “Stage of Change” from the Transtheoretical model of behavioral change (Prochaska & Velicer, 1997). To assess the model they determined whether confidence intervals contained their predictions. Eleven of the 15 predictions were included in 95% confidence intervals, which Velicer et al. state “provid[es] overall support for the theoretical model.” There are two issues with this assessment. First, it is arbitrary. Would 10 out of 15 also support the theory, or instead refute it? If we instead used 99% CIs, would 12 out of 15 be enough to support the theory? This arbitrariness arises because CIs offer no principled method for generating an inference regarding the theory. Second, no indication is given of what one would expect if the theory were false. If one would expect similar data even when the theory is false, then the observed data cannot be said to support the theory. In Cumming's example, two out of three necessary conditions for testing theory are missing.

We are solidly in agreement with Cumming, however, that any reliance on null hypothesis significance testing (NHST) should be avoided. NHST also fails to consider predictions when the null hypothesis is false and thus also cannot provide support for theory. If traditional hypothesis testing is inappropriate, what should replace it? One possibility that we advocate is Bayesian model comparison (Kass & Raftery, 1995; Rouder, Morey, Speckman, & Province, 2012; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), which meets the three necessary conditions for theory testing without suffering from the problems that plague NHST. In Bayesian model comparison, the probability of the observed data is compared across various models. This comparison is justified by Bayes' theorem, yielding a principled, continuous measure of relative evidence called the Bayes factor. Non-Bayesian methods for model comparison exist as well (Burnham & Anderson, 2002; Grünwald, 2007), though a discussion of the specifics of the different approaches is outside the scope of this comment.

For psychological science to be a healthy science, both estimation and hypothesis testing are needed. Estimation is necessary in pre-theoretical work before clear predictions can be made, and in post-theoretical work for theory revision. But hypothesis testing, not estimation, is necessary for testing the quantitative predictions of theories. Neither hypothesis testing nor estimation is more informative than the other; rather, they answer different questions. Using estimation alone turns science into an exercise in keeping records about the effect sizes in diverse experiments, producing a massive catalog devoid of theoretical content; using hypothesis testing alone may cause researchers to miss rich, meaningful structure in data. For researchers to obtain principled answers to the full range of questions they might ask, it is crucial for estimation and hypothesis testing to be advocated side by side.

References

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information theoretic approach (second ed.).* New York, NY: Springer-Verlag.

Cumming, G. (in press). *The new statistics: Why and how.* Psychological Science.

Eich, E. (in press). *Business not as usual*. Psychological Science.

Grant, D. A. (1962). *Testing the null hypothesis and the strategy and tactics of investigating theoretical models*. Psychological Review , 69 , 54-61.

Grünwald, P. (2007). *The minimum description length principle.* Cambridge, MA: MIT Press.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). *Working memory training may increase working memory capacity but not fluid intelligence*. Psychological Science , 24 (12), 2409-2419.

Huelsenbeck, J. P., & Ronquist, F. (2001). *MRBAYES: Bayesian inference of phylogenetic trees*. Bioinformatics, 17 (8), 754-755.

Kass, R. E., & Raftery, A. E. (1995). *Bayes factors*. Journal of the American Statistical Association, 90 , 773-795.

Low, I., Lykken, J., & Shaughnessy, G. (2012). *Have we observed the Higgs boson (imposter)?* Physical Review D - Particles, Fields, Gravitation and Cosmology, 86.

Popper, K. (1959). *The logic of scientific discovery.* (2002 e-Library ed.) London: Routledge Classics.

Prochaska, J. O., & Velicer, W. (1997). *The transtheoretical model of health behavior change.* American Journal of Health Promotion , 12 , 38-48.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). *Default Bayes* *factors for ANOVA designs*. Journal of Mathematical Psychology, 56, 356-374.

Velicer, W. F., Cumming, G., Fava, J. L., Rossi, J. S., Prochaska, J. O., & Johnson, J. (2008). *Theory testing using quantitative predictions of effect size.* Applied Psychology, 57 (4), 589-608.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). *Why psychologists must change the way they analyze their data: The case of psi. A comment on Bem (2011).* Journal of Personality and Social Psychology, 100, 426-432.