frontiers in PSYCHOLOGY

IMPACT 2.8 FACTOR

Cognition

Is the call to abandon p-values the red herring of the replicability crisis?

Victoria Savalei and Elizabeth Dunn

Journal Name:	Frontiers in Psychology
ISSN:	1664-1078
Article type:	Opinion Article
Received on:	03 Nov 2014
Accepted on:	17 Feb 2015
Provisional PDF published on:	17 Feb 2015
Frontiers website link:	www.frontiersin.org
Citation:	Savalei V and Dunn E(2015) Is the call to abandon p-values the red herring of the replicability crisis?. *Front. Psychol.* 6:245. doi:10.3389/fpsyg.2015.00245

Is the call to abandon p-values the red herring of the replicability crisis?

Victoria Savalei[1]*, Elizabeth Dunn[1]

University of British Columbia

Correspondence:

Dr. Victoria Savalei
Department of Psychology
University of British Columbia
2136 West Mall
Vancouver, BC V6T 1Z4
v.savalei@ubc.ca

[1] Department of Psychology, University of British Columbia, Vancouver, Canada

Abstract

Cumming (2014) argues that the field should abandon the reliance on p-values in favor of reporting confidence intervals (CIs) only. We show that Cumming overstates the consensus in the field when it comes to the harms done by null hypothesis significance testing (NHST). We also reject the implied connection made by Cumming between psychology's reliance on NHST and the current crisis in the field. Finally, we point out that there is no evidence that Bayes factors would be any less prone to misinterpretation and misuse. We argue that any statistical reform within psychology should be evidence-based, a criterion psychologists demand of any other domain where our science is applied.

Keywords: null hypothesis significance testing (NHST), p-values, confidence intervals (CIs), Bayes factors, crisis of replicability

In a recent article, Cumming (2014) called for two major changes to how psychologists conduct research. The first suggested change—encouraging transparency and replication—is clearly worthwhile, but we question the wisdom of the second suggested change: abandoning p-values in favor of reporting confidence intervals (CIs) only in all psychological research reports. This article has three goals. First, we correct the false impression created by Cumming that the debate about the usefulness of NHST has been won by its critics. Second, we take issue with the implied connection between the use of NHST and the current crisis of replicability in psychology. Third, while we agree with other critics of Cumming (2014) that hypothesis testing is an important part of science (Morey, Rouder, Verhagen, & Wagemakers, 2014), we express skepticism that alternative hypothesis testing frameworks, such as Bayes factors, are a solution to the replicability crisis. Poor methodological practices can compromise the validity of Bayesian and classic statistical analyses alike. When it comes to choosing between competing statistical approaches, we highlight the value of applying the same standards of evidence that psychologists demand in choosing between competing substantive hypotheses.

Has the NHST Debate Been Settled?

Cumming (2014) claims that "very few defenses of NHST have been attempted" (p. 11). In a section titled "Defenses of NHST," he summarizes a single book chapter by Schmidt and Hunter (1997), which in fact is not a defense but another critique, listing and "refuting" arguments for continued use of NHST[2]. Thus, graduate students and others who are new to the field might understandably be left with the impression that the debate over NHST has been handily won by its critics, with little dissent. This impression is wrong. Indeed, the book that published Schmidt and Hunter's (1997) chapter (Harlow, Mulaik, & Steiger, 1997) included several defenses (e.g., Abelson, 1997b; Mulaik, Raju, Harshman, 1997), and many contributions

[2] See Krantz (1999) for a criticism of the faulty logic in this chapter.

with more nuanced and varied positions (e.g., Harris, 1997; Reichardt & Gollob, 1997). Defenses have also appeared in the field's leading peer-reviewed journals, including *American Psychologist* (Krueger, 2001, with commentaries) and APA's quantitative psychology journal *Psychological Methods* (Frick, 1996; Cortina & Dunlap, 1997; Nickerson, 2000). Nickerson (2000) provided a particularly careful and thoughtful review of the entire debate and concluded "that NHST is easily misunderstood and misused but that when applied with good judgment it can be an effective aid to the interpretation of experimental data" (abstract). Perhaps the most famous critique of the use of NHST in psychology (Cohen, 1994), published in the *American Psychologist,* has seen several defending commentaries (Baril & Cannon, 1995; Parker, 1995; Frick, 1995), plus a lengthier retort (Hagen, 1997). We do not believe that the debate about the appropriate use of NHST in psychology has been decisively settled. Further, the strong NHST-bashing rhetoric common on the "reformers" side of the debate may prevent many substantive researchers from feeling that they can voice legitimate reservations about abandoning the use of *p*-values.

Is the Replicability Crisis Caused by NHST?

Cumming (2014) connects the current crisis in the field (e.g., Pashler & Wagenmakers, 2012) to "the severe flaws of null-hypothesis significance testing (NHST)." In our opinion, the reliance of psychologists on NHST is a red herring in the debates about the replicability crisis (see also Krueger, 2001). Cumming cites Ioannidis (2005) to draw the connection between NHST and the replicability crisis. Yet, Cumming does not explain how the fundamental problems articulated by Ioannidis (2005) could be resolved by abandoning NHST and focusing on CIs. Ioannidis (2005) described the intersecting problems that arise from running

underpowered studies, conducting numerous statistical tests, and focusing only on the significant results. There is no evidence that replacing p-values with CIs will circumvent these problems[3]. After all, p-values and CIs are based on the same information, and are thus equivalently susceptible to "hacking."

While Cumming warns that using CIs in the same way we use NHST (to reach a binary decision) would be a mistake and advocates not focusing on whether a CI includes 0, it is difficult to imagine researchers and editors ignoring this salient information. In fact, we feel that all claims about the superiority of one statistical technique over another in terms of facilitating correct interpretation and reasoning should be supported by evidence, as we would demand of any other claim made within our discipline. The only experimental study evaluating whether presenting data in terms of CIs reduces binary thinking relative to NHST did not find this to be the case[4] (Hoekstra, Johnson, and Kiers, 2012; see also Poitevineau, & Lecoutre, 2001). Another purported advantage of abolishing p-values is that using CIs may make it easier to detect common patterns across studies (e.g., Schmidt, 1996). However, a recent experiment found that presenting the results of multiple studies in terms of CIs rather than in NHST form did not improve meta-analytic thinking (Coulson, Healey, Fidler, & Cumming, 2010)[5]. It has also been argued that CIs might help improve research practices by making low power more salient, because power is directly related to the width of the confidence interval. There is some evidence that presenting data in terms of CIs rather than p-values makes people less vulnerable to

[3] For instance, Ioannidis's (2005) main example (Box 1) is a hypothetical study with the goal to test whether any of the 100,000 gene polymorphisms are associated with susceptibility to schizophrenia, with the prior odds for any one polymorphism set to be .0001, and with the power of 60% to detect any one association. It is unclear how this intersection of problems, which plagues all exploratory research, can be solved with CIs.

[4] We recognize the irony of drawing a binary inference of no evidence from this study, but the authors also reach this conclusion (and they also present both CIs and p-values to support their conclusions).

[5]Participants were from three different fields with varying statistical practices. As Coulson et al. (2010) noted: "Confidence intervals have been routinely reported in medical journals since the mid-1980s, yet our MED [medical] respondents did not perform notably better than BN [behavioral neuroscience] and PSY [psychology] respondents" (p. 8).

interpreting non-significant results in under-powered studies as support for the null hypothesis (Fidler & Loftus, 2009; Hoekstra et al., 2012). Unfortunately, our reading of this research also suggests that using CIs pushed many participants in the opposite direction, and they tended to interpret CIs that include 0 as moderate evidence for the alternative hypothesis. It is worth debating which of these interpretations is more problematic, a judgment call that may depend on the nature of the research. Finally, existing data do not support the notion that CIs are more intuitive. Misinterpretations of the meaning of CIs are as widespread as misinterpretations of p-values[6] (Belia, Fidler, Williams, & Cumming, 2005; Hoekstra, Morey, Rouder, & Wagenmakers, 2014). Abolishing p-values and replacing them with CIs, thus, is not a panacea.

Successfully addressing the replicability crisis demands fundamental changes, such as running much larger studies (Vankov, Bowers, & Munafo, 2014; Button et al., 2013), directly replicating past work (Nosek, Spies, & Motyl, 2012), publishing null results, avoiding questionable research practices that increase "researcher degrees of freedom" (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, and Simonsohn, 2011), and practicing open science more broadly. To the extent that replacing p-values with CIs appears to be an easy, surface-level "solution" to the replicability crisis—while doing little to solve the problems that caused the crisis in the first place—this approach may actually distract attention away from deeper, more effective changes.

Are Bayes Factors the Solution to the Replicability Crisis?

Bayes factors have gained some traction in psychology as an alternative hypothesis-testing framework (e.g., Dienes, 2011; Kruschke, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009). This approach may be logically superior in that Bayes factors directly address the

[6] In fact, Cumming (2014) himself gives some decidedly Bayesian interpretations of CIs.

relative evidence for the null hypothesis versus the alternative. Another major advantage is that Bayes factors force researchers to articulate their hypotheses in terms of prior distributions on the effect sizes. A simple "H_1: $\mu>0$" will no longer do the trick, and the answer to the question "Is my hypothesis supported by the data?" will depend on the exact form of that hypothesis. Decades ago, Meehl (1990) argued that such a development was needed to push the science of psychology forward.

In the wake of the replicability crisis, some have argued that switching to Bayesian hypothesis testing can help remedy the bias against publishing non-significant results because, unlike NHST, Bayes factors allow researchers to establish support for the null (Dienes, 2014). More evidence is needed, however, that the switch to Bayes factors will have this effect. To the extent that the real source of publication bias is the pressure felt by journal editors to publish novel, striking findings, the rate of publication of null results will not increase, even if those null results are strongly supported by a Bayesian analysis. Further, when it comes to questionable research practices, one can "b-hack" just as one can "p-hack" (Simonsohn, 2014; Yu, Sprenger, Thomas, & Dougherty, 2014; Sanborn & Hills, 2014). In fact, Bayes factors and the values of the classic t-test are directly related, given a set sample size and choice of prior (Rouder et.al., 2009; Wetzels et. al., 2011). Although some have argued that the options for "b-hacking" are more limited (e.g., Dienes, 2014; Wagenmakers, 2007, in an online appendix; Rouder, 2014), no statistical approach is immune to poor methodological practices.

Furthermore, as pointed out by Simmons, Nelson, and Simonsohn (2011), using Bayes factors further increases "researcher degrees of freedom," creating another potential QRP, because researchers must select a prior—a subjective expectation about the most likely size of the effect—for their analyses. Although the choice of prior is often inconsequential (Rouder et.

al., 2009), different priors can lead to different conclusions. For example, in their critique of Bem's (2011) article on pre-cognition, Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) have devoted much space to the reanalysis of the data using Bayes factors, and less to pointing out the exploratory flexibility of many of Bem's (2011) analyses. Bem's response to this critique (Bem, Utts, & Johnson, 2011) was *entirely* about the Bayesian analyses—debating the choice of prior for psi. Given that the publication of Bem's (2011) article was one of the factors that spurred the current crisis, this statistical debate may have been a red herring, distracting researchers from the much deeper concerns about QRP's.

Conclusion

We agree with Cumming (2014) that raw effect sizes and the associated CIs should routinely be reported. We also believe that Bayes factors represent an intriguing alternative to hypothesis testing via NHST. But, at present we lack empirical evidence that encouraging researchers to abandon p-values will fundamentally change the credibility and replicability of psychological research in practice. In the face of crisis, researchers should return to their core, shared value by demanding rigorous empirical evidence before instituting major changes.

References

Abelson, R. P. (1997a). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12-15.

Abelson, R. P. (1997b). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (117-144). Mahwah, NJ: Lawrence Erlbaum.

Baril, G. L., & Cannon, J. T. (1995). What *is* the probability that null hypothesis testing is meaningless? *American Psychologist, 50,* 1098-1099.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389-396.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407-425.

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology, 101,* 716-719.

Button, K. S., Ioannidis, J. P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., & Munafo, M.R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience, 14,* 365-376.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49,* 997-1003.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2,* 161-172.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology, 1*.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7-29. doi:10.1177/0956797613504966

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274-290.

Dienes, Z. (2014). *How Bayes factors change scientific practice*. Unpublished manuscript. Available at http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/publications.html

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology, 217,* 27–37.

Frick, R. W. (1995). A problem with confidence intervals. *American Psychologist, 50,* 1102-1103.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379–390.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15-24.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (145-174). Mahwah, NJ: Lawrence Erlbaum.

Hoekstra, R., Johnson, A.,& Kiers, H. A. L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement, 72,* 1039–1052.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (in press). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2,* 696-701.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524-532.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 44,* 1372-1381.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56,* 16-26.

Kruschke (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299-312.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806-834.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195-244.

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science, 25,* 1289-1290.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (65-116). Mahwah, NJ: Lawrence Erlbaum.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631. doi:10.1177/1745691612459058

Parker, S. (1995). The "Difference of means" may not be the "effect size". *American Psychologist, 50,* 1101-1102.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528–530.

Poitevineau, J., & Lecoutre, B. (2001). Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated. *Psychonomic Bulletin & Review, 8*, 847-850.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical significance tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (259-286). Mahwah, NJ: Lawrence Erlbaum.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225-237.

Rouder, J. N. (2014). Optional Stopping: No Problem For Bayesians. *Psychonomic Bulletin & Review, 21,* 301-308.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review, 21,* 283-300.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366.

Simonsohn, U (2014). *Posterior-Hacking: Selective Reporting Invalidates Bayesian Results Also* (January 2, 2014). Available at SSRN: http://ssrn.com/abstract=2374040

Vankov, I., Bowers, J., & Munafo, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology, 67,* 1037-1040.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779-804.

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: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426–432. doi:10.1037/a0022790

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology an Empirical Comparison Using 855 t-tests. *Perspectives on Psychological Science, 6,* 291-298.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. Psychonomic Bulletin & Review. doi:10.3758/s13423-013-0495-z