Journal of International Business Studies (2020) 51, 274–278

www.jibs.net

Open

INVITED COMMENTARY

New-day statistical thinking: A bold proposal for a radical change in practices

Arjen van Witteloostuijn

School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Correspondence:
Arjen van Witteloostuijn, School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
e-mail: a.van.witteloostuijn@vu.nl

Abstract
In this commentary, I argue why we should stop engaging in null hypothesis statistical significance testing altogether. Artificial and misleading it may be, but we know how to play the *p* value threshold and null hypothesis-testing game. We feel secure; we love the certainty. The fly in the ointment is that the conventions have led to questionable research practices. Wasserstein, Schirm, & Lazar (Am Stat 73(sup1):1–19, 2019. https://doi.org/10.1080/00031305.2019.1583913) explain why, in their thought-provoking editorial introducing a special issue of *The American Statistician*: "As 'statistical significance' is used less, statistical thinking will be used more." Perhaps we empirical researchers can together find a way to work ourselves out of the straitjacket that binds us.
Journal of International Business Studies (2020) **51**, 274–278.
https://doi.org/10.1057/s41267-019-00288-8

Keywords:: null hypothesis testing; statistical significance; questionable research practices

Received: 19 June 2019
Accepted: 22 October 2019
Online publication date: 2 December 2019

BACKGROUND

In a 2017 editorial in the *Journal of International Business Studies* (*JIBS*), Klaus Meyer, Sjoerd Beugelsdijk, and I proposed a number of research guidelines that we believed would help in meeting the journal goal of enhancing the rigor of the empirical hypothesis-testing work published. One of those was to abandon the asterisk threshold *p* value; and hand-in-glove with that, another made it imperative to report and discuss actual effect sizes. With the adoption of the new guidelines, a new measuring stick was put in place at *JIBS*. Editors and reviewers should expect the authors of empirical papers, for instance, to report actual *p* values, to show real impact calculations, to include a genuine discussion of effect sizes, and to provide robustness analyses. This is a huge step forward, and one that was badly needed (van Witteloostuijn, 2016). *JIBS* can be proud of being among those in the vanguard of improving the quality of hypothesis testing and statistical reporting. *Management and Organization Review* and *Strategic Management Journal* are implementing similar changes, and in other Business and Management journals there are calls for change, too, such as by

Schwab, Abrahamson, Starbuck, & Fidler (2011) in *Organization Science*, by Lockett, McWilliams, & Van Fleet (2014) in the *British Journal of Management*, and by Starbuck (2016) in *Administrative Science Quarterly*. Considerable credit goes to Bill Starbuck who has argued passionately for more than a decade in articles, on websites, and at workshops against null hypothesis significance testing (e.g., https://sites.google.com/site/nhstresearch/).

In academics, as in so much else in this world, inertia abounds (Starbuck, 2016; van Witteloostuijn, 2016). International Business scholars are fully aware of this. Individuals do not like change. The organizations and the systems we formulate change slowly – if at all. Statistical significance is the flagship of the quantitative research methodology, testing for the statistical significance of a null hypothesis an unquestioned part of our research work. We have been trained to do things in a certain way, and we fall back on that training when we are conducting research or reviewing an article, or editing a journal. We perpetuate the system by training our own students to do as we do and have always done. Everyone expects an empirical paper to report, discuss, and interpret findings using statistical significance. It is more than just routine behavior; there is an ideology behind it, promoting a single, right way to go about quantitative hypothesis-testing research. How can we change? And for what?

Artificial and misleading it may be, but we know how to play the *p* value threshold and null hypothesis-testing game. We feel secure; we love the certainty. The fly in the ointment is that the conventions have led to questionable research practices, which we now seek to beat to death by introducing new guidelines, such as those regarding the discussion of effect sizes and running robustness checks. We do know that we should change, and that we need access, openness and transparency. It will take time for everyone to realize that, but it will come. *JIBS* and a few other influential journals have had the courage to cut the moorings. Those already doing away with asterisk threshold *p* values, and that are already reporting and discussing effect sizes genuinely, are doing just fine. So far so good, but when we articulated in *JIBS* *p* value and effect-size reporting guidelines, my co-authors and I also listed another eight (Meyer, van Witteloostuijn, & Beugelsdijk, 2017). If a new standard research practice is to be ushered in, the changes outlined in both *JIBS* editorials need to be adopted. But in this commentary, I take yet another step: we have to let go of "statistical significance" once and for all as "the important research question is not whether any effects occur, but whether these effects are large enough to matter" (Schwab et al., 2011, p. 1108).

PROGRESS

In this short commentary, I argue that the timely and important steps already taken by *JIBS* should be followed by still others. Specifically, I believe that we should do away with the notion of statistical significance and null hypothesis testing altogether. I am not alone. Wasserstein, Schirm, and Lazar, (2019, p. 1) explain why, in their thought-provoking editorial introducing a special issue of *The American Statistician*: "As 'statistical significance' is used less, statistical thinking will be used more." The special issue has 43 articles and all of them, in one way or another, argue that current statistical significance practices, if not the modern statistical significance obsession altogether, are just plain wrong. Some of the most prominent statisticians of our day have concluded that "it is time to stop using the term 'statistically significant' entirely" (Wasserstein et al., 2019, p. 2). Why? "Regardless of whether it was ever useful, a declaration of 'statistical significance' has today become meaningless … And so the tool has become a tyrant" (Wasserstein et al., 2019, p. 2).

I am encouraged in my own thinking by seeing that such august company shares the opinion that the way *p* value is used is a mistake. In fact, it was never supposed to become the end-all and be-all of empirical social science. It cannot do what many think it can, and indeed believe it does. It cannot provide "support" for hypotheses nor "confirm" a theory. It does not speak to the truth, importance, or relevance of an association or an effect. To paraphrase the pithy words of Gelman & Stern, (2006), the difference between what is claimed to be "significant" and what is said to be "not significant" is not statistically significant. Therefore, the point I would like to make in the current commentary is that, in addition to openness and transparency (cf. Beugelsdijk, van Witteloostuijn, & Meyer, 2019), we must also embrace uncertainty. As Tukey (1991, pp. 101–102) has said, "The worst, i.e., most dangerous feature of 'accepting the null hypothesis' is the giving up of explicit uncertainty."

After a full century of the old way of doing things (Boring, 1919), there is a new way that calls for really looking at the data and evaluating the degree

of compatibility with different theories – and that does not mean the near universal use of "no effect" as an alternative theory. Amrhein, Trafimov, & Greenland (2019) and Greenland (2019) suggest replacing confidence with "compatibility" intervals, an idea similar in spirit to Matthews (2019) suggestion of adopting an "analysis of credibility", Calquhoun (2019) notion of a "false positive risk", and Goodman (2019) proposal of a "confidence index". Turning blindly to Bayes rule is not the solution, as that theorem is also associated with dichotomizing threshold-like factors and priors – a difficulty in and of itself in the absence of replication. What we must do is scrutinize the data for the degree of compatibility with different theories, using effect sizes, actual *p* values (if any), power analyses, confidence (or compatibility) intervals, data visualization, sign consistency, robustness checks, and so on without resorting to "statistical significance" or "rejecting" or "supporting" (null) hypotheses. There is no denying that this means making subjective judgments, but that is inevitable. After all, everything in this world is inherently uncertain. Our reward for confronting – even embracing – uncertainty will be access, openness, and transparency. It is the only way forward.

What will "New Reporting" look like? We should not give in to the temptation to look backward for some pat answer. We need to take the time to gradually develop a menu of New Reporting guidelines. At this point, I can but introduce a few of my own ideas. I suggest five principles. The first is to deeply engage with the data. Data should be carefully described, including the specific context (Delios, 2017), using data visualization tools wherever that proves insightful (Greve, 2017). The second is to drop the no-effect null as a standard benchmark. It is patently meaningless. Rather, we should focus on compatibility with alternative hypotheses. In so doing, we can compare the explanatory power of alternative theories. The third is NO HARKing. There should be an explicit distinction between ex ante hypotheses and ex post inferences. This means using methodologies other than hypothetico-deduction ones, including abduction (cf. Lockett et al., 2014; Starbuck, 2016). The fourth is to focus on substantive effects. We need to experiment with alternative metrics to replace the banned "statistical significance". Some ways are suggested in *The American Statistician* 2019 special issue. Such experimentation aligns well with the guidelines proposed in Meyer et al. (2017) – i.e., an open discussion about uncertainty and the addition of robustness analyses. The fifth and final one is to do away with the obsession with "groundbreaking uniqueness". There is real value in replicating – exactly, and through different types of extensions (Starbuck, 2016; Walker, Brewer, Lee, Petrovsky, & van Witteloostuijn, 2019).

Challenging all this may be, it is at the same time very exciting. We are researchers after all. I for one welcome the challenge. We will have to be creative in how we analyze, present, and interpret findings – all of us, authors, reviewers, and editors alike. There is quite a bit already out there to work with. Many of our current practices are just fine, provided we are able to use them differently – e.g., reporting actual *p* values, but without reference to statistical significance, and genuinely discussing effect sizes. Moreover, we can borrow from Statistics, where suggestions of how to move away from statistical significance abound. The 43 contributions to this 2019 special issue of *The American Statistician,* to which I have referred a number of times, are a rich source of inspiration.

IMPLEMENTATION

Decades ago, I myself was trained in old-school statistical significance. With a background in Economics and Psychology, I am fully and deeply socialized in the tradition of null hypothesis significance testing. I have questioned over the years the way things are done, but the 2017 *JIBS* editorial was largely written with the idea of tweaking the null hypothesis significance testing paradigm by adopting corrective guidelines. Two years further down the road, after many discussions with co-authors and colleagues, following the ongoing debate in Statistics, and turning the issue over and over in my mind, I have become convinced that Sjoerd Beugelsdijk, Klaus Meyer, and I (2017, 2019) did not go far enough in our editorials. That opinion was confirmed by what I recently read in *Nature* (20 March 2019) by Amrhein, Greenwald, and McShane, endorsed by more than 800 signatories: "We're frankly sick of seeing nonsensical 'proofs of the null' and claims of non-associations in presentations, research articles, reviews and instructional materials." They go on

to "call for the entire concept of statistical significance to be abandoned", and conclude that not to do so would allow "these errors [to continue to] waste research efforts and misinform policy decisions." I can only agree.

Without a doubt, deviating from long-established ways is risky. Many worthy attempts to handle data differently or to replicate and build on prior work have ended up in the round file, as has been the case since long before I entered academia. Like me, the majority of my co-authors and the colleagues with whom I have discussed the issue are unhappy with the current state of affairs and would like to see a change. Is it possible for us to go against the tide without sabotaging our careers? Perhaps we empirical researchers can together find a way to work ourselves out of the straitjacket that binds us. Will it happen before I retire? Maybe. Change is in the air. I detect some momentum. By inviting editorials on the topic, *JIBS* opened an important discussion. If the top journal in International Business continues to lead by example, calls for experimenting with a variety of new lenses to describe, analyze, and interpret data, other academic journals may well follow suit. Who knows? After all, science mirrors as well as attempts to explain real life, and that inevitably means honestly dealing with uncertainty.

AFTERWORD

Dozens of colleagues read a first draft of this commentary, and offered critique and support. Listing them all is undoable, but that does not imply that I am not thankful – I am.

REFERENCES

Amrhein, V., Greenland, S., McShane, B. and 800-plus signatories. 2019. Scientists rise up against statistical significance. Comment in *Nature*. 20 March 2019, https://doi.org/10.1038/d41586-019-00857-9.

Amrhein, V., Trafimow, D., & Greenland, S. 2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1): 262–270. https://doi.org/10.1080/00031305.2019.1583913.

Beugelsdijk, S., van Witteloostuijn, A., & Meyer, K. E. 2019. The evolving rules of data management in the publication process: Data Access and Research Transparency (DART). *Journal of International Business Studies*, 49: forthcoming.

Boring, E. G. 1919. Mathematical vs. scientific significance. *Psychological Bulletin*, 16(10): 335–338.

Calquhoun, D. 2019. The false positive risk: A proposal concerning what to do about *p*-value. *The American Statistician*, 73(sup1): 192–201. https://doi.org/10.1080/00031305.2019.1583913.

Delios, A. 2017. The death and rebirth (?) of international business research. *Journal of Management Studies*, 54(3): 391–397.

Gelman, A., & Stern, H. 2006. The difference between 'significant' and 'not significant' is not itself statistically significant. *The American Statistician*, 60(4): 328–331.

Goodman, S. 2019. Why is getting rid of *p*-values so hard? Musing on science and statistics. *The American Statistician*, 73(sup1): 26–30. https://doi.org/10.1080/00031305.2019.1583913.

Greenland, S. 2019. Valid *p*-values behave exactly as they should: Some misleading criticisms of *p*-values and their resolution with *s*-values. *The American Statistician*, 73(sup1): 106–114. https://doi.org/10.1080/00031305.2019.1583913.

Greve, H. R. 2017. From the Editor. *Administrative Science Quarterly*, 62(2): v–vi.

Lockett, A., McWilliams, A., & Van Fleet, D. D. 2014. Reordering our priorities by putting phenomena before design: Escaping the straightjacket of null hypothesis significance testing. *British Journal of Management*, 25(4): 863–873.

Matthews, R. 2019. Moving toward the post $p < 0.05$ era via the analysis of credibility. *The American Statistician*, 73(sup1): 202–212. https://doi.org/10.1080/00031305.2019.1583913.

Meyer, K. E., van Witteloostuijn, A.. & Beugelsdijk, S. 2017. What's in a *p*? Reassessing best practices for conducting and reporting hypothesis-testing research. *Journal of International Business Studies*, 48(5): 535–551.

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. 2011. Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, 22(4): 1105–1120.

Starbuck, W. H. 2016. 60th anniversary essay: How journals could improve research practices in social science. *Administrative Science Quarterly*, 61(2): 165–183.

Tukey, J. W. 1991. The philosophy of multiple comparisons. *Statistical Science*, 6(1): 100–116.

van Witteloostuijn, A. 2016. What happened to Popperian falsification? Publishing neutral and negative findings: Moving away from biased publication practices. *Cross Cultural and Strategic Management*, 23(3): 481–508.

Walker, R. M., Brewer, G. A., Lee, M. J., Petrovsky, N., & van Witteloostuijn, A. (2019). Best practice recommendations for replicating experiments in public administration. *Journal of Public Administration Research and Theory*, *29*(4), 609–626.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. 2019. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1): 1–19. https://doi.org/10.1080/00031305.2019.1583913.

ABOUT THE AUTHOR

Arjen van Witteloostuijn is Professor of Business and Economics at the Vrije Universiteit (VU) Amsterdam and Dean of the VU School of Business and Economics in the Netherlands, as well as Research Professor in Business, Economics and Governance at the University of Antwerp and Antwerp

Management School in Belgium. He is Fellow of the Academy of International Business and Area Editor of the Journal of International Business Studies. He has published in journals such as the *Academy of Management Journal*, *Academy of Management Review*, *American Journal of Political Science*, *American Journal of Sociology*, *American Sociological Review*, *Journal of International Business Studies*, *Management Science*, *Organization Science*, and *Strategic Management Journal*.

Accepted by Alain Verbeke, Editor-in-Chief, 22 October 2019. This article has been with the author for one revision and was single-blind reviewed.