Why did it take so many decades for the behavioral sciences to develop a sense of crisis around methodology and replication?[1]

Andrew Gelman[2] and Simine Vazire[3]

10 Apr 2021

For several decades, leading behavioral scientists have offered strong criticisms of the common practice of null hypothesis significance testing as producing spurious findings without strong theoretical or empirical support. But only in the past decade has this manifested as a full-scale replication crisis. We consider some possible reasons why, on or about December 2010, the behavioral sciences changed.

"On or about December 1910 human character changed." — Virginia Woolf (1924).

Woolf's quote about modernism in the arts rings true, in part because we continue to see relatively sudden changes in intellectual life, not merely from technology (email and texting replacing letters and phone calls, streaming replacing record sales, etc.) and power relations (for example arising from the decline of labor unions and the end of communism) but also ways of thinking which are not exactly new but seem to take root in a way that had not happened earlier. Around 1910, it seemed that the literary and artistic world was ready for Ezra Pound, Pablo Picasso, Igor Stravinsky, Gertrude Stein, and the like to shatter old ways of thinking, and (in a much lesser way) the behavioral sciences were upended just about exactly 100 years later by what is now known as the "replication crisis."

We have occasion to consider this in the context of a previously unpublished article by psychologist Lewis Petrinovich from 1990, raising concerns about the state of behavioral research in ways that anticipate many ideas currently present in the science reform movement. But for reasons that are still not well understood, the behavioral sciences did not move toward serious reform until a few years ago, following the "replication crisis," as exemplified by three articles in leading psychology journals: a methodological criticism of "voodoo correlations" in neuroscience (Vul et al., 2009); a notorious paper claiming evidence for extra-sensory perception (Bem, 2011), and an article that introduced the terms "researcher degrees of freedom" and "p-hacking" (Simmons et al., 2011). Gelman (2016) reviews some of the tumult that followed. The behavioral sciences are at a different stage of methodological self-awareness than they were in 1990 or 1970, when Campbell, Cronbach, Meehl, and others appeared to represent only a small minority of the profession when lamenting the mismatch

[1] Discussion of "What behavioral scientists are unwilling to accept," by Lewis Petrinovich, for Journal of Methods and Measurement in the Social Sciences. We thank Margaret Echelbarger, Paul Smaldino, and several other people for helpful comments, Alex Weiss for soliciting this article, and the U.S. Office of Naval Research for partial support of this work.

[2] Department of Statistics and Department of Political Science, Columbia University, New York.

[3] Melbourne School of Psychological Sciences, University of Melbourne.

between statistical methods and substantive theory in the social and behavioral sciences.

From the 1960s onward, the methodological reform movement included leading academic researchers who made compelling arguments, which leads us to ask, from our perch in the third decade of the twenty-first century: What went wrong? Why were these solid arguments set aside? Why, despite the pleas of Petrinovich and others, did the behavioral sciences take so long to internalize these critiques? And then what happened to make the field suddenly recognize the need for change and respond to the crisis?

We offer here no systematic historical study, and our views arise from a mix of perspectives: one of us has done research in theoretical and applied psychology, and the other has published articles in behavioral science, but only on methodology.

To start with, internal change is difficult. Success within a field such as publications and academic honors, and external success such as media exposure and bestselling books, provide signals that reinforce persistence of the status quo. In addition, even those who find research critiques to be persuasive cannot do much with negative advice. Recommendations to develop stronger theory or better measurement often do not provide any particular way forward, and even when they do, it is not clear how to incentivize adoption of these practices (Smaldino, 2021). Third, the behavioral sciences were not in stasis during the last decades of the twentieth century: major progress was made in subfields including developmental psychology, cognitive behavioral therapy, and heuristics and biases, just to name a few. Sometimes an academic area of study is forced to change because of developments from neighboring fields (as, for example, statistics has been altered by machine learning during the past decade), but the flow of ideas in cognitive science has largely gone the other way, with research in neuroscience and the psychology of judgment and decision making influencing computer science and behavioral economics, respectively. So, despite some missteps, psychology was doing very well, both in public perception and in research advances, during the decades when the warnings of Meehl and others were largely disregarded. This suggests that the statistical methods being used at the time, while suboptimal, were working acceptably well, or at least not getting in the way of many useful developments.

In that case, why the big changes since 2010? Why have the ideas of Meehl and other reformers suddenly seemed so relevant to so many? We have pointed to some specific events taking place between 2009 and 2011, but presumably these only had an impact because the field was ready for change, with a deep reservoir of dissatisfaction among psychology researchers at all levels, including some senior academic researchers who took the lead in the open science movement and many junior scholars who were willing to speak out in an effort to avoid dedicating their careers to producing trivial and often unreplicable results.

What was different in 2010 and the years that followed that made the same criticisms resonate in a way they hadn't before? We offer some speculations. First, the ability to connect on social media could have made a big difference. In the past, skeptical researchers were isolated—at most they might find one or two other sympathetic voices in their home department. With social media, geographically sparse critics could more easily connect and advance each others' ideas. Communicating with others who were independently harboring the same concerns may have

emboldened the early critics in the latest wave of the reform movement. Moreover, social media presented an avenue to raise awareness among colleagues who may not have been skeptical to begin with but were interested in hearing the arguments for reform.

Other technological advances also likely facilitated the spread of the reform movement (Spellman, 2015). Sharing data, code, and materials has become much easier in the last 20 years. This makes it more reasonable and more socially acceptable to ask authors to let us see what went into their publications. In addition, collecting some kinds of data (e.g., survey experiments conducted on convenience samples) has become much easier and cheaper, making it easier to run replication studies and increase sample sizes. Without these advances, most of the activities of today's reform movement would have been very difficult if not impossible.

It's also possible that awareness of problems with diversity and representation, which had been growing for decades and was crystallized in the term WEIRD (people from Western, educated, industrialized, rich and democratic societies; Henrich, Heine, and Norenzayan, 2010), contributed to an atmosphere of self-scrutiny in the field (Schiavone, Bottesini, and Vazire, 2020). Attention to the problem of sample diversity may have contributed to more general concerns about generalizability of findings, including the narrowness of the research questions, researchers, participant populations, settings, and stimuli represented in the published literature.

One other possible factor is the role of key personalities, and particularly the kinds of contributions that often come from people with "insider-outsider" roles (people who have left the field or left their academic position, people in adjacent fields, etc.). These people have enough specialized knowledge and status to offer useful techniques that can be taken seriously, but not so much that it becomes difficult for them to see or speak out about the problems in their own field. On the other hand, earlier critics such as Meehl and Campbell had insider-outsider perspectives, and this was enough for them to be influential but it did not allow them to change the core patterns of behavior in behavioral research.

Finally, it may be that the situation in some subfields of behavioral science got worse between 1990 and 2010. This may be the case if, for example, the popularization of social psychology in the previous decade by writers like Malcolm Gladwell and Daniel Goleman led to increased public interest in and demand for behavioral research to provide answers to pressing problems of everyday life. This could have presented new opportunities and pressure for social psychologists to meet this demand. If researchers and journal editors became more tolerant of dramatic conclusions without stronger evidence, this may have amplified existing concerns regarding questionable research and publishing practices. Perhaps a reckoning regarding the credibility of social and behavioral sciences was inevitable, and 2010 was simply when we reached a breaking point.

These are surely not the only factors, and may not even be the most important, but in any case we believe it is worth trying to understand the starkly different trajectory of the current reform movement compared to past efforts.

References

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.

Gelman, A. (2016). What has happened down here is the winds have changed. Statistical Modeling, Causal Inference, and Social Science blog, 21 Sept. https://statmodeling.stat.columbia.edu/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world. Behavioral and Brain Sciences 33, 61-83.

Petrinovich, L. (1990). What behavioral scientists are unwilling to accept. Reprinted in Journal of Methods and Measurement in the Social Sciences (2021).

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

Smaldino, P. (2021). How to build a strong theoretical foundation. Psychological Inquiry 31, 297-301.

Spellman, B. A. (2015). A short (personal) history of revolution 2.0. Perspectives on Psychological Science 10, 886-899.

Vazire, S., Schiavone, S. R., and Bottesini, J. (2020). Credibility beyond replicability: Improving the four validities in psychological science. PsyArXiv preprint: https://psyarxiv.com/bu4d3/

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on Psychological Science 4, 274-290.

Woolf, V. (1924). Mr. Bennett and Mrs. Brown. Hogarth Press.