Open access • Journal Article • DOI:10.1016/J.LEAQUA.2017.01.006

On doing better science: From thrill of discovery to policy implications

John Antonakis

Institutions: University of Lausanne

Published on: 01 Feb 2017 - Leadership Quarterly (JAI)

Topics: Applied research

On doing better science: From thrill of discovery to policy implications

John Antonakis

Faculty of Business and Economics

University of Lausanne

john.antonakis@unil.ch

in press

The Leadership Quarterly

Acknowledgements:

I am grateful to all my colleagues, collaborators, and students who have helped shaped my thinking and have made me into a better researcher. The following individuals provided me with helpful comments in the development of this article: Nicolas Bastardoz, Valérie Chavez-Demoulin, Michael S. Cole, José M. Cortina, David V. Day, Alice H. Eagly, Jeffrey R. Edwards, Olga Epitropaki, Saskia Faulk, W. Page Faulk, William L. Gardner, S. Alexander Haslam, John P. A. Ioannidis, Mikko Ketokivi, James LeBreton, Cameron N. McIntosh, Michael D. Mumford, Steven G. Rogelberg, Mikko Rönkkö, Marianne Schmid Mast, Seth M. Spain, Rolf van Dick, Roberto A. Weber, Victoria Wetherell, Rolf van Dick, Mark van Vugt, Patrick M. Wright, Francis J. Yammarino, and Christian Zehnder.

Abstract

In this position paper, I argue that the main purpose of research is to discover and report on phenomena in a truthful manner. Once uncovered, these phenomena can have important implications for society. The utility of research depends on whether it makes a contribution because it is original or can add to cumulative research efforts, is rigorously and reliably done, and is able to inform basic or applied research and later policy. However, five serious "diseases" stifle the production of useful research. These diseases include: *Significosis*, an inordinate focus on statistically significant results; *neophilia*, an excessive appreciation for novelty; *theorrhea*, a mania for new theory; *arigorium*, a deficiency of rigor in theoretical and empirical work; and finally, *disjunctivitis*, a proclivity to produce large quantities of redundant, trivial, and incoherent works. I surmise that these diseases have caused immense harm to science and have cast doubt on the role of science in society. I discuss what publication gatekeepers should do to eradicate these diseases, to stimulate the undertaking of more useful and impactful research, and to provide the needed incentives to better align the interests of researchers with those greater good. Finally, I highlight where technical improvements are needed to enhance research quality, and call on deeper reflection, transparency, and honesty in how we do research.

In an accidental conversation during my postdoc, I learned about the importance of different scientific modes of discovery and their implications for policy. I interacted often with one of my neighbors, Lisa, who was a postdoc too; we never really talked about work because most of the time we met, we had our hands full with our young children who were play friends. I knew she did research in molecular biology or something like that. One day, when the moment was opportune, I asked Lisa what she was working on precisely and she replied "protein trafficking." I jokingly asked whether that involved smuggling proteins across country borders! She responded, "No silly, but you are not far off—I am trying to understand how proteins are transported in and out of cells."

"Why study that?" I asked her. Given my ignorance, I did not immediately see the point of what Lisa was doing and inquired how research of this sort could inform practice, whether in medicine or elsewhere. She said that she did not care about practical implications for the time being; those would come as the phenomenon was better understood. After several discussions with her, and many similar discussions with other researchers, I began to appreciate the importance of what she did. Many years later, this conversation sprang to mind when the 2013 Nobel Prize in physiology/medicine was awarded to researchers working on just this topic—one that has provided the foundation for many important medical applications.

The conversation I had with Lisa has often intrigued me and is a relevant starting point to set the stage for this position paper, and for better understanding the defining characteristics of useful science.[1] Whether reading an article about neutrinos, the evolutionary origins of our decision-making, or whether leadership can be taught, the answers to the following three generic questions should provide some indication of the potential importance of the article, and whether it will go on to have an impact:

[1] Note I use the terms "science" and "research" as well as "scientist" and "researcher" interchangeably.

1. So what?
2. Is it rigorous?
3. Will it make a difference?

These questions are vital ones to ask because their answers can inform us if the article is useful to the scientific record. I realize too that these are judgment calls and that as editors we are not perfect prognosticators—yet, I will lay out a case in this article regarding how we might better determine if an article will be impactful. The answer to the "so what" question informs us if the theoretical or empirical contribution is original, or if the finding—even if not original—is one that would contribute to cumulative research efforts. The "rigor" answer informs us about the robustness, accuracy, and reliability of the research, and if it reflects the actual description, process, or causal relation uncovered. Finally, the "will it make a difference" answer gives an indication concerning the extent to which the finding can inform basic research, so that we can better understand the building blocks of the phenomenon; if the research is more applied in scope, the research should inform policy or practice.

Even if relevancy may not immediately be evident does not imply that practice cannot be informed at a later stage of the scientific process. At this point, I wish to be clear on one key issue: There is no tradeoff between rigor and relevance as some management scholars have suggested (e.g., Bennis & O'Toole, 2005). Research that is *not rigorous* simply cannot be *relevant* (F. Vermeulen, 2005). The relevance-rigor debate is a false dichotomy. Yet, we as leadership scholars must do more to ensure that our research enters public policy debates because what we study has important practical implications. Our conduit to public policy should not be only via behavioral economics, as it is for some psychology theories that have been formalized by economists (Chetty, 2015). We must show leadership!

In this position paper I reflect on how science *is* currently being done and on how science *should* be done to ensure it is accurate and ultimately useful. Of course, there is a great deal of good research that has been published across all fields of science, including by leadership researcher. Yet, merely looking at a mainstay metric of research utility—citations—shows that many articles simply do not receive many citations (as has been long established: D. Hamilton, 1991) and the leadership field is not exempt from this criticism (Antonakis, Bastardoz, Liu, & Schriesheim, 2014). I am also deeply troubled when findings are not accurately reported (Atwater, Mumford, Schriesheim, & Yammarino, 2014), even on simple things like *p*-values (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016) or degrees of freedom (Cortina, Green, Keeler, & Vandenberg, 2016) . Not all of these errors should be attributed to fraud because most scientists are well-intentioned; but because of institutional incentive structures and cultural transmission of research practices, oftentimes research gets published that does not help advance science or practice (Nosek, Spies, & Motyl, 2012; Smaldino & McElreath, 2016). There are many reasons why this phenomenon may occur, but suffice it to say that it should be unconscionable to well-intended scientists that resources—in the form of salaries, institutional support, or time spent by journals to publish papers—are squandered on research that may not be useful.

We cannot hope these problems away. Science is not necessarily self-correcting because the process of publication may not be correctly managed by the publication gatekeepers. For whatever reasons, including failure to robustly replicate research, there are simply too many findings that become "received doctrines" (Barrett, 1972) or "unchallenged fallacies" (Ioannidis, 2012). We have to collectively make an effort to address this problem and to ensure that what is published is robustly done, tentatively accepted, and then actively challenged. Only then can we claim to approach the truth. How we do our science therefore matters much, which is why we

need more Socratic-type academic gadflies who do "meta-research" and reflect on how we can improve our scientific practice (Ioannidis, Fanelli, Dunne, & Goodman, 2015).

On another front, high-profile scientific scandals—whether about retractions or replication failures—have been picked-up by the mass media; such exposure is necessary, but at the same time it sullies the reputation of scientists and raises doubts about the place of science in this world. At the same time, we are in an era where any kind of information can come across as being credible if put on a website. It is no wonder we have anti-vaxxers, climate change deniers, evolution doubters, even flat-earthers, and this despite broad consensus from experts about the actual state-of-the-science regarding these issues (Achenbach, 2015). Sadly, this anti-intellectualism is evident in many walks of life (Jacoby, 2009); worse, even mainstream political parties actively undermine science (Mooney, 2005).

To correct the status quo, it will take effort and courage, and most importantly, a different way of managing the publication process. I will thus openly discuss challenges that science faces, not just in our field, but across a broad array of fields. Although I will borrow from general discussions and debates from a variety of disciplines, I will focus my paper on how we can improve on the scientific study of leadership.

Diseases that threaten the viability of science

We are in an era where submitting an article is a click away, where there is fierce competition to publish in top-ranked outlets, and where journals compete to draw what they believe will be impactful papers in terms of future citations; every researcher is vying to publish in high impact journals and every journal wants to have a high impact factor. Sexy and statistically significant findings are the talk of the day, particularly in psychology (Bakker, van Dijk, & Wicherts, 2012; Ellemers, 2013). But journals are not always publishing robust and policy-relevant findings. Defective studies often slip through (B. H. Hamilton & Nickerson,

2003), and our field is no exception (see Antonakis, Bastardoz, et al., 2014; Antonakis, Bendahan, Jacquart, & Lalive, 2010; Fischer, Dietz, & Antonakis, 2016). This problem is not just about inattentive peer reviewers or editors. Although I cannot provide a full and definitive diagnosis of the problem, the most likely explanation may have to do with the conditions and the incentives that are in place to reward success in the production of research. It involves institutions, journals, and authors. What constitutes good research performance for institutions and what determines whether it is published by journals sets the goals for authors.

It all begins with well-intentioned scientists who sometimes may have, what is nicely called, "hypothesis myopia"—that is, a desire to seek evidence to confirm what they are looking for (Nuzzo, 2015), the well-known problem of confirmation bias (Nickerson, 1998). Consider the case of Jacques Benveniste, an immunologist. He and his team made the revolutionary discovery that water had "memory," which proved to him that homeopathy—whose core belief is to dilute the active ingredient to infinitesimal levels to the point where no molecule of the ingredient can be detected—had real effects. Aficionados of homeopathy cheered. However, a controlled *replication* demonstrated that the results reported by Benveniste and his team—results that violated the laws of physics and chemistry and which were theoretically impossible—were flawed. The explanation was simple: The technicians, who counted how many cells reacted to homeopathically-treated or normal water, were not blinded to the treatments and "found" the effect that they were expecting (Maddox, 1988; Maddox, Randi, & Stewart, 1988).

Problems of this sort—that is, publishing research that may not be accurate—are happening in our field too not only because of ideologically-motivated or badly-trained scientists, but because the incentive structure in the publication game is wrong, particularly with respect to what constitutes a contribution. When contribution is equated to novel and statistically significant, and when replication or null findings are frowned upon and hardly ever undertaken

(Franco, Malhotra, & Simonovits, 2014; Makel, Plucker, & Hegarty, 2012)—and by replication I mean also by independent researcher groups and not just "replications" by the same authors in the same article—it is no wonder that the research landscape is littered with erroneous results (Ioannidis, 2005). Studies that are wrong, and consequently the theories that they have helped build, have to be decommissioned. Only findings or theories that stand the test of time, because they have survived attempted refutation (Popper, 1989), should be used to guide research and policy. Tavris and Aronson (2007, p. 108) put it nicely by noting that ''the scientific method consists of the use of procedures designed to show not that our predictions and hypotheses are right, *but that they might be wrong*."

There is better way to do science and editors are duty-bound to their editorial discretion to serve the greater good. Recently, I have also been involved in an initiative where we have laid out a sensible agenda to start dealing with the issues our discipline faces (Grote, 2016)—I use ideas from here, as well as many other sources (e.g., Bakker, et al., 2012; Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016; Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016; Ellemers, 2013; Ioannidis, 2014; Nosek, et al., 2012), my previous methodological work, and the many conversations I have had with colleagues on the topic over the years, whether in our field or others. Distilling the general problems science faces, and those of our field in particular, was not easy because I had to read across disciplines; however, I hope that my efforts here will make the core issues and challenges we collectively face more understandable so that the measures I propose make sense.

Because of the incentive structures that institutions and journals set as well as the propagation of particular types of research practices, progress in our science is impeded by five major "diseases." These diseases may not have unique causes and they intertwined. I named them as such to group together, in a simple way, the conditions that stifle the production of good

science whether in the social sciences or beyond. I have called them diseases and given them pernicious sounding names to draw attention to how serious these impediments are to science. The diseases include: (a) a fixation on statistically significant results ("*significosis*"), (b) an obsession for novel work ("*neophilia*"), (c) a fetish for new theory ("*theorrhea*"), (d) a lack of rigor in developing theory and undertaking empirical work ("*arigorium*"), and (e) a penchant to accumulate large amounts of disparate factoids, salami sliced output, and trivial works, with little attempt at theoretical integration ("*disjunctivitis*").

Disease 1: Significosis

Even if we assume that most scientists are competent, ethical, and are motivated to do their jobs correctly, and even if good research is being published, the system still produces an overall biased distribution of estimates. It is usually statistically-significant results that get published (H. Cooper, DeNeve, & Charlton, 1997). This problem is pervasive and occurs "when the probability that a result is published depends on the estimates produced by the study, holding the methodological quality of the study fixed" (Gerber & Malhotra, 2008, p. 4). In other words, the distribution of effects—because of chance or other causes—is clearly biased and does not represent what is actually out there (Banks, Rogelberg, et al., 2016; Ioannidis, 2016a). The problem of publication bias is ubiquitous and evident across many fields (Pfeiffer, Bertram, & Ioannidis, 2011) making it difficult to reconstruct the distribution of effect sizes. Yet it is easy to shoot down bad research via critical analysis, re-analysis of results, simulation, or replication. Biased effect-size distributions are used as foundations for future theory and feed into meta-analyses. Perhaps the effects claimed are real; perhaps they are not. The problem is that because they are statistically significant and published, they become legitimized and part of the research canon, and may be used to guide future research and decide policy.

Disease 2: Neophilia

It is the discovery of the truth that we should be rewarding "and not about getting spectacular, but wrong results" (Ioannidis, 2012, p. 652). For instance, "power posing" (Carney, Cuddy, & Yap, 2010)—wherein a brief pose can apparently significantly increase testosterone levels and risk taking—captured the imagination of the public because these highly novel and groundbreaking findings suggested that anyone could become powerful in a few minutes. The study is the research pillar for the most viewed TED talk of all time and a best-selling book by one of the coauthors (i.e., Cuddy). However, subsequent findings indicate that blinding experimenters to treatment nullifies the power-posing effect (Ranehill et al., 2015) in addition to other statistical issues that show that this effect does not exist (Simmons & Simonsohn, in press).

It is a plain fact that top empirical journals insist on findings that are novel, innovative and interesting (Cortina, 2016; Mathieu, 2016; C. C. Miller & Bamberger, 2016; Nosek, et al., 2012). Null results and replications as well as exploratory research is simply not seen as interesting or innovative enough by top journals (Cortina, 2016; Ferguson & Heene, 2012; Mathieu, 2016; C. C. Miller & Bamberger, 2016; Nosek, et al., 2012)—some journals, however, are finally breaking ranks (Bettis, et al., 2016; Hollenbeck & Wright, 2017; R. S. Landis, James, Lance, Pierce, & Rogelberg, 2014; C. C. Miller & Bamberger, 2016). We need to radically shift prevailing wisdom to make a difference, and more progressive thinking is needed across the board. We can learn much from null results and replications; empirical exploratory work, which is not, as many think, the province of qualitative researchers, is needed too (Lapierre, Edwards, Oswald, Shockley, & Landis, 2017).

Disease 3: Theorrhea

In addition to management journals like the *Academy of Management Review* (*AMR*), which only publishes theory articles, many of the top empirical journals expect that articles make

a theoretical contribution (Hambrick, 2007; Locke, 2007). Yet, ironically, 90% of theories/models published in *AMR* are never tested (Kacmar & Whitfield, 2000). More recent evidence, looking at the top-cited articles in *AMR* confirms that almost none of the theoretical propositions are directly tested in empirical research (Edwards, Berry, & Stewart, 2016). It is rather odd that we have "theory envy" when the best theories our field produces are never tested. Of course theory—in terms of explaining a phenomenon by describing how variables are causally related, why they are related, and the boundary conditions under which the relations hold (Bacharach, 1989; Dubin, 1976)—is very useful for guiding future research. I therefore strongly support efforts for further development of good-quality theory in our field. Although theory is the pinnacle of our efforts in science (Kerlinger, 1986), the building blocks of a theory-focused research do not require a hypothetico-deductive approach to research; identifying compelling and interesting empirical relations that can help scientists understand a phenomenon are also useful (Hambrick, 2007; D. Miller, 2007). Ironically too, we would like to think we "do" theory; however, as compared to other fields we do not *really* do theory. I return to this point when discussing the next disease, arigorium.

Disease 4: Arigorium

There is a general lack of rigor in our field, whether in theory development or testing. As compared to other disciplines that are more formalized (e.g., economics), our field operates from a weak paradigm, in terms of widely agreed upon assumptions, models, and methods (Pfeffer, 1993). For instance as regards theory, on a very basic level there is a general lack of precision in definitions, assumptions, and in expounding on the variables constituting the theory as well as their causal impact. This problem is evidenced in several theories including three of the most dominant theories in our field: Transformational leadership (van Knippenberg & Sitkin, 2013), charismatic leadership (Antonakis, Bastardoz, Jacquart, & Shamir, 2016), and leader-member

exchange theory (House & Aditya, 1997). This problem is one of theorizing that lacks in rigor, transparency, and precision—if we are to do theory properly we need to follow the examples of other disciplines and formalize theory (Adner, Polos, Ryall, & Sorenson, 2009); or at least be more precise with words, fleshing out assumptions, arguing counterfactually, clearly identifying the hypothesized causal mechanisms at play (Durand & Vaara, 2009) as well as the limits of the theory (Bacharach, 1989; Dubin, 1976).

Apart from weak theorizing, our field does not pay the needed attention to testing phenomena rigorously: This problem is manifold and has to do with study design as well as with estimation. On the simplest level, the statistical threshold for what constitutes a significant result (i.e., $p < .05$) may be too lax and may need to be lowered to .001 (Johnson, Payne, Wang, Asher, & Mandal, in press), which may help lower false positive findings (Ioannidis, 2014). There are many more complex issues making it hard to document them all. Put simply, our field lags in using the appropriate statistical methods particularly with respect to modeling data in non-experimental research (Antonakis, et al., 2010). Unless one only does basic experimental research, the usual estimation procedures in our field, ANOVA or OLS regression (coupled with the usual software i.e., SPSS), are insufficient to deal with the challenges researchers face today; in non-experimental settings the assumptions of these estimators are usually violated with respect to the exogeneity of the independent variables. Quantitative researchers are more cognizant of how to improve research standards because it is straightforward to demonstrate with analytical proofs or simulations what makes for a correct estimator. Only recently have top management journals expected authors to deal decisively with endogeneity (Bettis, Gambardella, Helfat, & Mitchell, 2014; Guide Jr & Ketokivi, 2015; Reeb, Sakakibara, & Mahmood, 2012), which is a welcome sign. As concerns qualitative research, it is thus hard to judge the rigor of such studies and opinions regarding what constitutes a good use of qualitative methods vary widely

(Antonakis, Bastardoz, et al., 2014; Dixon-Woods, Shaw, Agarwal, & Smith, 2004; Gibbert & Ruigrok, 2010; Gioia, Corley, & Hamilton, 2013).

Disease 5: Disjunctivitis

Because of the conditions that have allowed all the other diseases to proliferate, we witness a large volume of research produced. Researchers must enhance their reputations and stake out a territory, whether in terms of theory, a method, or a measure (cf. Barrett, 1972). They produce what they are mostly incentivized to do: Quantity (Bergh, Perry, & Hanke, 2006; Fanelli & Larivière, 2016; Grote, 2016; Larivière & Costas, 2016; Nosek, et al., 2012). Collaboration with coauthors—one failsafe way to boost output—is on the increase (Ductor, 2015). Given that articles are the unit of analysis counting toward productivity, there is a proliferation of publications via another pathway: A trend to produce short and rapid publications (Bertamini & Munafò, 2012; Ellemers, 2013). In some fields, this increase in output has occurred across-the-board and even affects the most prized of evidence: Systematic reviews and meta-analysis (Ioannidis, 2016a). Moreover being highly prolific increases the likelihood that a subset of papers go on to be highly cited (Larivière & Costas, 2016; Sandström & van den Besselaar, 2016). The end result is a voluminous set of articles that are fragmented both within and between disciplines (Ellemers, 2013; Wilson, 1998).

Symptoms of the diseases

When only statistically significant and novel results are published, individuals may game the system expressly; they may *p-hack* or engage in other questionable research practices. Interestingly, the proportion of support hypotheses receive increase as an article is transformed from dissertation to journal article (O'Boyle, Banks, & Gonzalez-Mulé, 2017). Of course, dissertations and published articles have different incentive structures, which may explain why when in a competitive publication game, researchers may, intentionally or not, use whatever

means (e.g., adding or removing controls, "cleaning" data so as to remove observations that are not helpful), both in experimental and non-experimental research, to make results statistically significant and thus publishable (Banks, Rogelberg, et al., 2016; O'Boyle, Banks, Walter, Carter, & Weisenberger, 2015; I. Vermeulen et al., 2015).

As Nobel economist Ronald Coarse said: "if you torture the data long enough it will confess" (Tullock, 2001, p. 205). Banning the reporting of *p*-values, as *Basic and Applied Social Psychology* recently did, is not going to solve the problem because it is merely treating a symptom of the problem. There is nothing wrong with hypothesis testing and *p*-values per se as long as authors, reviewers, and action editors use them correctly (Abelson, 1997; Bettis, et al., 2016; Cortina & Dunlap, 1997; García-Pérez, 2016; Murtaugh, 2014). Much of the argumentation against classical hypothesis testing framework is misplaced particularly because substantive effects beyond nil effects can be tested. That is, this framework does not preclude testing specific hypotheses against particular values or estimating models where parameters are constrained based on prior knowledge (Edwards & Berry, 2010). The source of the problem is that the winners in the publication race are those who have statistically significant—and ideally snazzy—results. Thus, changing the incentive structure, what constitutes "winning" (Bakker, et al., 2012) should reduce the prevalence of this disease to some extent. As a prime example, registered trials wherein the research design is first reviewed prior to gathering data, substantially increase the prevalence of null results and provides a more accurate assessment of treatment effects (Kaplan & Irvin, 2015).

Another symptom resulting from theorrhea, and first discussed explicitly in our field by Kerr (1998) is HARKing—"hypothesizing after the results are known." HARKing appears to be highly prevalent in our field (Bosco, Aguinis, Field, Pierce, & Dalton, 2016). Briefly, because research must be grounded in original theory to get published in top journals, researchers will do

what it takes to make it seem "as if" they tested some novel theoretical insights. They first dig into their data to find statistically significant results that look potentially interesting and new, and then invent a plausible theory to be used upfront in the paper to give the impression that this theory drove the data-gathering and data-testing efforts of the researchers. Thus, authors are making up theory in a post-hoc way while pretending the theory was ex-ante. Capitalizing on chance in this way, and coupled with the fact that non-novel findings are hard to publish creates a situation where findings are rarely if ever replicated, and theories never trimmed or refined, resulting in theory that is not very useful (Cortina, 2016).

Because of neophilia, significosis, and theorrhea, it is no wonder why null findings and replications are suppressed and exploratory work is ill-regarded (see Cortina, 2016; Ferguson & Heene, 2012; C. C. Miller & Bamberger, 2016). Some may think rightly so because exploratory work is about *p-hacking* (or maybe HARKing too). Well, some may see it this way or do research this way. As I see it, *p-hacking* is about capitalizing on chance by overfitting models to data (Leigh, 1988; Maccallum, Roznowski, & Necowitz, 1992)—of course, it is possible that some researchers HARK and then pass off their research as a theory-driven exercise. But, useful exploratory or inductive work is about relying on theory-free observation, and saying that at the outset, to discover important causal relations and then testing these observations in further experimentation. Theoretical explanations can come in later work and it is useful to know what the causal mechanism is that explains the relation.

Medicine is a good example wherein observation has led to important serendipitous discoveries (Klein, 2008). Arguably one of the greatest discoveries in medicine, and for humankind, is anesthesia. Since first discovered in 1846, there have been many different anesthetic elements used to induce sleep and manage pain in patients (Robinson & Toledo, 2012). The most widely used anesthetic today is propofol (Yip et al., 2013) and it was discovered by

accident (DeMonaco, Ali, & Von Hippel, 2005). Doctors can be rather sure that if given the appropriate dosage of propofol, the patient will not wake up during surgery. However, the exact mechanism by which propofol works is not fully clear (DeMonaco, et al., 2005), though inroads continue to be made to understand how it works (Haeseler & Leuwer, 2003; Yip, et al., 2013).

The lesson here is that exploratory work can lead to important discoveries (see also Hollenbeck & Wright, 2017), which need to be confirmed causally in subsequent research. More importantly is that *the absence of a causal explanation does not preclude a causal effect*; to influence policy and practice it is causal effects that we should going after (Antonakis, et al., 2010). Of course, providing a causal explanation for a known causal effect is nice to have so that we can fully understand the nature of the phenomenon; but it is not required to affect practice or to make a difference to society.

Furthermore, despite our focus on generating theory, too much of it makes it hard to put it all together (Ellemers, 2013). Moreover, the general lack of rigor in theory production is rather problematic; for example, constructs are oftentimes left undefined or not properly defined (e.g., charisma is a gift or mystical ability), and propositions are clearly tautological or defined by their outcomes (e.g., charismatic leaders are highly effective; see Antonakis, et al., 2016, for detailed discussions). Hazy theories and circular arguments cannot properly influence policy; only a clearly identified causal relation can (ideally with the mechanism explained). We need sharper definitions and to the extent possible to try figure out how to manipulate constructs of interest; as Kurt Lewin was credited to have said, “The best way to understand something is to try to change it.”

For instance, take the case of leader-member exchange theory (LMX), which, like the charisma construct, is riddled with issues[2]. Two decades ago, House and Aditya (1997) mentioned: "While it is almost tautological to say that good or effective leadership consists in part of good relationships between leaders and followers, there are several questions about such relationships to which answers are not intuitively obvious . . . . High-quality relationships may be influenced by a host of situational factors, follower attributes and behaviors, and leader behaviors. LMX theory does not specify these, but implies that the behavior of subordinates influences superiors to show support, delegate to subordinates a substantial amount of discretion in conducting their work, engage in open communication, and encourage mutual influence between themselves and their subordinates" (pp. 431-432). Thus, apart from LMX theorists defining and measuring LMX in terms of trust in the leader or leader-follower relationship quality (i.e., defined by its outcome), LMX per se depends on omitted causes that would predict the modelled dependent variable too, making it hard to know if LMX actually causes anything[3]. LMX is clearly an endogenous variable that shares many common causes with outcomes that it is meant to predict; if these causes are omitted, the effect of LMX on outcomes will be plagued by endogeneity and thus confounded. Has this theory been reformulated since the House-Aditya critique? Can LMX be manipulated in consequential settings? Unfortunately not; and, several meta-analyses have been conducted since, fed by endogeneity-rife estimates (Antonakis, Bendahan, Jacquart, & Lalive, 2014).

---

[2] Readers from management can simply replace LMX with any other endogenous variable (e.g., organizational commitment/citizenship, perceptions of job enrichment), which is often used as a predictor of other outcomes.

[3] An endogenous variable, y_1, studied as such (i.e., an outcome) is not problematic. What does cause problems in estimation is if the endogenous variable is modelled as a predictor of another endogenous variable, y_2. In this case, the coefficient of $y_1 \rightarrow y_2$ will be biased because y_1 will correlate with omitted causes (i.e., the disturbance) of y_2. The only way around this endogeneity problem is to randomize y_1 (i.e., make it exogenous with respect to y_2), to control for all known causes of y_2 (which is difficult to do), to use instrumental variable estimation, where an exogenous cause x, is used to lock in the causal direction of y_2 on y_2, or to use a quasi-experimental procedure (see Antonakis et al., 2010).

LMX theory is not the only one at fault. Many leadership models, including the transformational-transactional leadership model, are on weak theoretical scaffolding (van Knippenberg & Sitkin, 2013), and composed of endogenous variables (Antonakis, et al., 2016). For instance, leader intellectual stimulation of followers, a key part of transformational leadership, is conceptualized and measured to be an important predictor of follower performance. Yet, a leader may show more or less intellectual stimulation to a follower depending on follower skills or performance, because of other leader- or follower-level variables (e.g., personality), or because of organizational-level variables[4]. If any of these variables correlate with follower outcomes and are omitted from the model, estimates for intellectual stimulation will be biased.

The lack of demonstrable rigor in theorizing is matched by a lack of rigor in testing. Reviews have shown that there are many threats to the validity of quantitative leadership studies, which means that much of it is not informative to policy (Antonakis, Bastardoz, et al., 2014; Antonakis, et al., 2010; Fischer, et al., 2016). As concerns qualitative research, given the idiosyncratic methods used and a general lack of clear description of how data are extracted and analyzed, makes it hard to understand what was done and whether it is replicable (Antonakis, Bastardoz, et al., 2014; Dixon-Woods, et al., 2004; Gibbert & Ruigrok, 2010; Gioia, et al., 2013). There are, however, many appropriate ways to conduct qualitative research bearing in mind sampling issues (Geddes, 2003; Gerring, 2007), measurement reliability (Wright, 2016), and counterfactual thinking (Geddes, 2003; Gerring, 2007; Gerring & McDermott, 2007).

Given the diseases and their consequences, it is no wonder that there appears to be a proliferation of salami-sliced and marginal research in the field of management and business research (Karabag & Berggren, 2016); this problem is evident too even in research-synthesis efforts (Ioannidis, 2016a) and across disciplines (Wilson, 1998)! We need to proceed in

[4] In this case the endogeneity problem is more complicated and includes simultaneity (see Antonakis et al., 2010).

paradigm-driven steps, replicate, and incrementally extend, not try to radically reinvent (Nosek, et al., 2012). At this time, we have too many theories, many of which may be redundant (Banks, McCauley, Gardner, & Guler, 2016; Hoch, Bommer, Dulebohn, & Wu, 2016)—an atomization of science not only within but between disciplines (Wilson, 1998).

We witness too an increase in number of authors on papers in management research (Acedo, Barroso, Casanueva, & Galán, 2006). More authors on papers is not necessarily bad and does predict better statistical conclusion validity (Antonakis, Bastardoz, et al., 2014). Also, trends suggest that research done by teams goes on to be more cited than is solo work (Wuchty, Jones, & Uzzi, 2007). Beyond having more knowhow on the team, that articles with more co-authors go on to be more cited could be due to a variety of mechanisms, include a wider dissemination network (Antonakis, Bastardoz, et al., 2014). Still, there is the possibility of negative externalities, including the free-rider problem (cf. Ductor, 2015). Authorship issues of this sort may have an impact on the integrity of research too (Marušić, Bošnjak, & Jerončić, 2011). More research is required to better understand whether there is a problem in our field with respect to authorship misuse (Banks et al., 2016).

Consequences: Crises, lack of quality and impact

Because of the above diseases, some branches of psychology, a key foundational discipline of leadership, are currently facing difficulties with respect to replicability (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). What precisely brought the problem to the fore is not entirely clear to me, but anonymous post-publication peer review (e.g., www.pubpeer.com), science-oriented blogging (e.g., see www.datacolada.org), and the media have probably played a role (a useful one too). Journals might not have an interest to be first movers because many editors may believe that replication failures or retractions harm the reputations of their journals. *Au contraire*! Journal editors that show the needed intestinal

fortitude strengthen the reputations of their journals (cf. Cortina, 2015). Misreported or manipulated data, or findings that fail reanalysis efforts have, depending on their severity, to be corrected or retracted (Atwater, et al., 2014).

We are starting to see theories falling: High-profile cases include power posing (Ranehill, et al., 2015; Simmons & Simonsohn, in press; K. M. Smith & Apicella, 2016), which is relevant for leadership, and ego depletion theory (Hagger & Chatzisarantis, 2016; Lurquin et al., 2016) to name a couple. The leadership field too is replete with theories that may have been reformulated or fallen, if the field did not suffer from the five diseases. As with calls made for the field of industrial-organizational psychology in general, it is time that we begin to see how to trim or even cull theories that have overpopulated the research landscape (Cortina, 2016; Grote, 2016; Leavitt, Mitchell, & Peterson, 2010); we cannot sit and wait for the demise of their illustrious masters (Azoulay, Fons-Rosen, & Zivin, 2015).

As concerns qualitative research, researchers appear to be distrustful of such research findings (Antonakis, Bastardoz, et al., 2014; Gephart, 2004; Gerring, 2012; Wright, 2016), thus making it difficult to get such research into top journals (Wright, 2016). Those qualitative studies that do get published are significantly less cited as compared to work informed by other modes of inquiry (e.g., quantitative, review, or theory articles, see Antonakis, et al., 2016; Antonakis, Bastardoz, et al., 2014); qualitative papers also do not figure in the top cited papers relative to the base rate of articles produced (Antonakis, Bastardoz, et al., 2014). But it does not have to be so as I discuss later regarding the kinds of qualitative articles I seek for this journal.

Moreover, given the general lack of a unified paradigm, our field of social science research as compared to other disciplines does not get the attention or resources that it should—this is not a new issue of course (Barrett, 1972; Pfeffer, 1993). We also do not affect public policy as much as we could (Amir et al., 2005). Next, although we like to measure impact in

terms of citations garnered, it appears that academic impact of the sort we do does not translate to practical impact on external stakeholders, that is, those outside of academia (Aguinis, Suárez-González, Lannelongue, & Joo, 2012; Chan et al., 2014). In the field of leadership we face a similar problem; research findings, whatever their quality, are not used much by practitioners, who instead turn to unscientific popular books and fads (Zaccaro & Horn, 2003).

Finally, the emphasis on producing quantity places a sword of Damocles over the heads of researchers, particularly early career researchers: They must ratchet up the output, despite the fact that common sense would suggest that quantity and quality are two performance indicators that might make for unusual bedfellows (Antonakis & Lalive, 2008; Bergh, et al., 2006). Though it is possible to be highly productive and have high quality work, it is unusual; and, having low productivity does not mean that the work is of high quality. Fetishizing statistical significance, favors the production of trivial research that risks being pointless: Statistical and theoretical significance are not isomorphic (Haslam & McGarty, 2001). Yet, the system favors the selection of such research because of incentives for publication quantity (Nosek, et al., 2012; Smaldino & McElreath, 2016), which creates a publication volume that is putting a strain on the review system (Bertamini & Munafò, 2012; Grote, 2016).

We need better quality research that is cohesive and integrative—where new research is preceded by systematic reviews so that the territory is fully charted and known effects are used as benchmarks (Ioannidis, 2016b). A system that is producing a large volume of bad research (Smaldino & McElreath, 2016), disjunctive and marginal findings (Ellemers, 2013; Haslam & McGarty, 2001; Karabag & Berggren, 2016), many of which are ignored and fail to get cited (Antonakis, Bastardoz, et al., 2014), is simply not ideal. Instead we should be thinking in terms of (a) paradigm-driven research, starting in our field first (Nosek, et al., 2012), and then later, (b) consilience—the blending of the sciences and how bridges can be built between disparate

findings not only within the social sciences but also between disciplines (e.g., biology and social sciences, Wilson, 1998).

How to make a more useful contribution to the research record

Because of the five diseases, we are faced with a research production system that is not optimal. We need to change publication incentive structures and rethink how we evaluate research standards and what makes for a useful contribution. For this reason, and as documented in my "hello editorial," our journal will now accept a wider range of article types (Antonakis, 2017). These will include *registered reports* and *results-masked articles*, *replication* and *null results studies*, and *exploratory studies*, in addition to several other types of articles, like *commentaries* and *critiques* as well as *adversarial collaborations*. The general idea here is that articles that make a contribution can do so not only because they are statistically significant or novel but because they can also make important discoveries, or contribute to cumulative scientific work; the latter is currently heavily biased by our collective failure to publish non-significant findings.

At the same time, our field must focus on producing more rigorous and useful research, in terms of theorizing, study design, testing, and reporting. Fully dealing with the issues below will help ensure that manuscripts submitted to *The Leadership Quarterly* will not be desk rejected; I am reasonably confident too that considering the below suggestions will improve one's chances to publish in other top journals.

Theorize clearly

Theories in our field are very imprecise, even by the standards of management (Edwards & Berry, 2010). We have to do better by being more exact in our predictions, which can help in testing for specific effects (Edwards & Christian, 2014); in this way, we can test stronger theory by using prior information, which is possible within the classical hypothesis testing framework (e.g., by constraining parameter, comparing nested models, or testing parameters against specific

values via Wald tests). Precision and formalization of theory has many benefits for later research (Adner, et al., 2009). We can also aspire to more theoretical sophistication. Models do not have to be complex (e.g., even "toy models" with simple utility functions will do), and the steps toward building a model are not that difficult to undertake (Varian, 1997).

We should not mind "word theories," of the sort usually published in our field (and as typically done in the *Academy of Management Review*), whether arguing using (a) propositions to describe causal relations, (b) narrative arguments to identify patterns, connections, and process theories, or (c) identifying typologies (Delbridge & Fiss, 2013). What is most important is that the theory is clearly thought out (Durand & Vaara, 2009), terms are precisely defined (MacKenzie, 2003; P. M. Podsakoff, MacKenzie, & Podsakoff, 2016), and the relations clearly identified along with boundary conditions (Bacharach, 1989; Dubin, 1976).

What is most disturbing is tautological theorizing, reminiscent of what Wicklund (1990) calls "zero-level" theories, those where the *explanandum* redescribes the *explanans*. A more subtle version of such theorizing is "endogenous theorizing," which specifies endogenous variables (e.g., good leader-member relations) as causes of other endogenous variables (e.g., follower satisfaction). "Upstream causes" (e.g., stable individual difference, or higher-level factors like contextual factors) must be identified in such models, and the causal effects of endogenous variables on other endogenous variables must be clearly articulated along with their common causes (which has implications for causal testing, discussed later).

Be more creative in measurement

The typical cross-sectional survey design has been a workhorse for our field, but it has reduced policy implications because of its limit in measurement, its static nature and obvious endogeneity threats; it is also becoming rather overused in our field and in management in particular. However, depending on how effects manifest themselves over time it can be policy

relevant (see Fischer, et al., 2016). There is so much we still need to learn and it will take more than cross-sectional snapshots to get there. If one of our goals is to affect practice, we also need to figure out better ways to understand the leadership production pipeline, from selection to development and from individual to organizational-wide leadership (Day & Dragoni, 2015; Day, Fleenor, Atwater, Sturm, & McKee, 2014).

There are many ways to get to leadership by measuring different kinds of variables from various sources, including data that is archival-historiometric (Barnes, Dang, Leavitt, Guarana, & Uhlmann, 2016; Friedrich et al., 2014; Simonton, 2003, 2009), neurological (Waldman, Wang, & Fenters, 2016), genetic (De Neve, Mikhaylov, Dawes, Christakis, & Fowler, 2013), hormonal (van der Meij, Schaveling, & van Vugt, 2016), facial (Todorov, Mandisodza, Goren, & Hall, 2005), economic (Zehnder, Herz, & Bonardi, in press), big data (Tonidandel, King, & Cortina, in press), or agent-based simulation data (McHugh et al., 2016). Looking at leadership from evolutionary angles would also be very useful (J. E. Smith et al., 2016); even data from insects can help explain the need for leadership (Hodgkin, Symonds, & Elgar, 2014). Importantly, as we discover more about leadership from different angles, we can begin to identify, for instance, how to make links to the brain sciences, evolutionary biology, and cultural transmission; in this way, we can one day strive to provide a more unified explanation of human behavior, using biology, the social sciences, even humanities (cf. Wilson, 1998).

Getting to leadership via different means can also be accomplished by using different types of designs. For instance the regression discontinuity design is useful for studying a local treatment effect that is relevant to practitioners in a specific context or an entire system—it most closely mimics the experimental design in terms of validity of causal inference (Antonakis, et al., 2010; Mellor & Mark, 1998). Yet, it is seriously underused in our field. Another unknown design is the difference-in-differences design (Antonakis, et al., 2010; Meyer, 1995), which can be used

as an alternative to case studies for identifying causal effects over time in systems (e.g., two company sites having received a different treatment) or even large entities like countries. Both these designs are contextually rooted and provide practitioners with specific information about a causal effect relevant to them.

Moreover, data from unusual contexts—not typically thought of in organizational science—can help shed important light on the leadership phenomenon and will be particularly welcome (see Bamberger & Pratt, 2010). Historical-, archeological-, or geographic-inspired studies would be especially interesting (Diamond & Robinson, 2010).

Design realistic experiments

The sure way to deal with the endogeneity problem is to manipulate the modelled independent variables; thus, I would like to see more experiments published in our journal. Experiments have much to teach us (Brown & Lord, 1999) and there are many ways in which experimental settings can be creatively used to study important real-world analogs (Zelditch, 1969). Although some are skeptical of laboratory experiments, findings from laboratory and field settings concord pretty well (Anderson, Lindsay, & Bushman, 1999; Herbst & Mas, 2015; Mitchell, 2012). In addition, students who partake in experiments do not act that differently from individuals in the general population insofar as social preference-type decisions are concerned (Falk, Meier, & Zehnder, 2013). Of course, the sample and setting must be appropriate for the problem at hand and the population to which one wishes to generalize (Haslam & McGarty, 2001; Hogarth, 2005).

Although it is not a requirement, incentivizing participants when it makes sense to so do and is realistic could help improve the ecological validity of the findings (Hertwig & Ortmann, 2001; Zizzo, 2010). That is, decisions taken and performance outcomes should have clear payoffs; moreover, the more consequential the decisions and the outcomes are, and the more

meaningful and realistic the situation is for participants, the more valid the setting becomes. It is one thing measuring intentions and impressions of participants in settings that have hypothetical outcomes and where participants can engage in "cheap talk" or give socially-desirable answers to questions; it is another thing measuring actual choices, having real-life tradeoffs that affect participants' payoffs in situ.

Even "paper people"—vignette-type or scenario—experiments can be useful (Aguinis & Bradley, 2014; Woehr & Lance, 1991); in my book, such an experiment well done may beat an endogeneity plagued-field study. Nowadays it is not so difficult to add more realism to the stimulus materials (e.g., using video materials, actors; see: N. P. Podsakoff, Podsakoff, MacKenzie, & Klinger, 2013); even more high-tech options, like virtual reality are now affordable and have many broad applications (Bombari, Mast, Canadas, & Bachmann, 2015). I hope also to see more field experiments in our journal; consequential and ecologically-valid settings are particularly welcome because such designs are the "gold standard" for causal, and externally valid, evidence (Chatterji, Findley, Jensen, Meier, & Nielson, 2016; Eden, 2017; Harrison & List, 2004).

Of course, experiments are not a panacea and must be carefully designed. Manipulation checks should be used if necessary and appropriate (Sigall & Mills, 1998); they should be avoided or manipulation checks using another sample should be used if the checks in the original sample may induce suspicion or reveal the purpose of the study and possibly trigger demand effects (i.e., where participants act in a way that they think is expected or desirable). Attention must be paid to not making unfair comparisons (W. H. Cooper & Richardson, 1986), which could trigger demand effects too. For instance, typical power-priming studies (Galinsky, Gruenfeld, & Magee, 2003), in which the treatment group is told to think about a time they had power and the control group thinks about something banal (what they did yesterday), is highly problematic

(Sturm & Antonakis, 2015); such a design could trigger demand effects (Orne, 2009; Rosenthal & Rosnow, 2009; Zizzo, 2010) and consequently, confounded estimates. Or how about the problem of comparing a possibly demand-driven treatment stemming from subjects receiving training and hence attention (Barling, Weber, & Kelloway, 1996), to a control group that is given no attention (cf. W. H. Cooper & Richardson, 1986)? Experimenters should consider having placebo controls or alternative treatments to serve as strong counterfactuals.

Finally, I hope to see little or no deception used (Hertwig & Ortmann, 2008; Ortmann & Hertwig, 1997, 2002). If used, it should be as a last resort only. Institutionalized deception is hard to justify and it sets a bad example to the very students we train and who will later enter working life. With thorough planning and an appropriate budget I believe that most studies can be run without using deception. If they cannot, perhaps the nature of the experiment is unethical per se—for example, as has been often discussed with regard to the Milgram electric shock experiments (there are other, deeper ethical complexities involved in the Milgram experiments, see: Haslam, Reicher, Millard, & McDonald, 2015). Obfuscating the purpose of the experiment or the manipulations is not deception; however, lying, misleading, or purposefully misrepresenting the purpose of the experiment is (Ortmann & Hertwig, 2002). Repeated deception engenders suspicion in participants and could affect participants choices in experiments (Jamison, Karlan, & Schechter, 2008; Ortmann & Hertwig, 2002).

Interestingly, the best concordance between laboratory and field results comes from industrial-organizational psychology and not social psychology (Mitchell, 2012)—although the latter is known to use deception quite regularly (Ortmann & Hertwig, 2002), it is not clear what explains this difference in concordance. Interestingly too, economics experiments (Camerer et al., 2016) have a much higher replicability rate than do psychology experiments (Open Science Collaboration, 2015); again, the reason for this finding is not clear but it may have an explanation

in a combination of factors including that economists: (a) use more formal theory, (b) put a premium on causal identification, (c) incentivize participants, (d) do not use deception, and (d) are usually very transparent about methods and sharing data (cf. Camerer, et al., 2016). It is also possible too that behavioral economists have "borrowed" mostly from the best psychology theories.

Do quality qualitative research

Qualitative and small-*n* size research can be very useful, particularly when studying large, unusual, or elite entities (Gerring & McDermott, 2007); such research can provide important contextual insights about leadership phenomena. Thus, qualitative research plays a needed role in understanding leadership and I would like to see more well-done qualitative studies in our journal. Some obvious questions that I raise when reading qualitative papers are: Are data reliably coded and are the findings replicable? That is, if an independent team analyzed the data would they come to similar findings? Are the findings trustworthy? As Gerring (2012, p. 94) has noted: "replicability is an idea for which all research ought to strive. Arguably, it is even more important for qualitative work than for quantitative work, given the degree of authorial intervention that is usually involved in the latter (and hence the greater possibility of investigator bias)."

The key issue here has to do with measurement (Wright, 2016); human judgment, even expert judgment is fallible (Dawes, Faust, & Meehl, 1989) and processes information in expectancy-based ways (Fiske, 1995; Tversky & Kahneman, 1974)[5]. Thus, researchers using the qualitative method should be as transparent as possible in their procedures (Weiner-Levy & Popper-Giveon, 2013), to use mixed methods when possible and to triangulate (Bansal & Corley,

[5] This problem is not unique to qualitative research; though it is more difficult in qualitative research to establish the validity and reliability of measures used.

2011; Bluhm, Harman, Lee, & Mitchell, 2011). When feasible, data should be coded (Wright, 2016) and the reliability in the data codings be demonstrated (J. R. Landis & Koch, 1977; Lin, 1989). Also, an often ignored issue in qualitative research is the failure to obtain data on contrasting cases (Geddes, 2003). For example, to understand what drives high performance, intensively studying only what high performing entities have in common is not helpful—what if low performing entities also share these characteristics (Denrell, 2003, 2005)? I agree that studying extreme cases can be interesting for some types of research problems (Gerring, 2007), but understanding causes of outcomes, generating propositions about causes of outcomes for future testing, or understanding processes requires counterfactual thinking and observation, along space and time (Gerring & McDermott, 2007; Hastie & Dawes, 2001).

The medical sciences may be useful here particularly with respect to testing effects using very small samples wherein, for instance, a group having been exposed to some treatment or condition (e.g., having a particular brain lesion) is compared to a matched control group (J. Duncan, Emslie, Williams, Johnson, & Freer, 1996). Matching of this sort can be achieved in a variety of ways, even on a basic level (Grimes & Schulz, 2005), though ideally should be undertaken with a statistical algorithm (Li, 2013; Niccolai et al., 2007). Any type of entity can be matched in this manner (e.g., individual, company, country). Having controls of this sort can be helpful in improving causal inference and in building better theories (Gerring & McDermott, 2007).

Do not ignore endogeneity

Although there is a lot of good published work on leadership, unfortunately there are a lot of methodologically weak studies because of endogeneity issues (Antonakis, Bastardoz, et al., 2014; Antonakis, et al., 2010; Fischer, et al., 2016). Endogeneity, which is a matter of degree, biases estimates in unknown ways; the higher the endogeneity, the more the bias to the point

where results cannot inform policy. Yet, endogeneity-threatened results have been used to inform research and practice; way too many meta-analyses have been done that fail to fully consider and transparently discuss the endogeneity-riddled data that feed them. Meta-analysis are supposed to be at the highest rung of the hierarchy of evidence (Evans, 2003); but if not well done how useful can they be to policy (Ioannidis, 2016a)?

Much has been written on the topic of endogeneity and how to correct for it but journals in management and applied psychology have been slow to adopt more robust standards in causally testing relations. When independent variables have not been manipulated a researcher must ensure that the estimator used is consistent (i.e., as the sample size increases coefficients get closer to the true effect) and unbiased (i.e., the mean coefficient approaches the true effect in repeated sampling). It is essential that authors reflect critically about the exogeneity of their measures; and it is important to note that the exogeneity of a variable is firstly a theoretical question (Antonakis, et al., 2010; Stock & Watson, 2011; Wooldridge, 2002).

For instance, a common issue I see is to justify using a particular design—splitting measurement periods into two or more times (or splitting rater sources)—as a good remedy for reducing common source/method effects. Depending on the nature of what is being measured, such a procedure could reduce some of these effects (P. M. Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; P. M. Podsakoff, MacKenzie, & Podsakoff, 2012); however, doing so is not sufficient to ensure exogeneity of the modeled independent variable (Antonakis, et al., 2010)[6].

Another common issue I see is modeling of only one measured leadership style (e.g., only empowering leadership) in predicting an outcome. Beyond failing to control for key predictors of

[6]Simply measuring x at time 1 (x_{t1}) does not guarantee that it is a likely cause of y at time 2 (y_{t2})—the well-known *post hoc ergo propter hoc fallacy* (Kerlinger, 1986). If x is not exogenous, both x and y could be explained by an omitted common cause, z, which will usually correlate with itself over time, that is, $\text{cov}(z_{t1},z_{t2} \neq 0)$ and consequently explain the $x_{t1} \rightarrow y_{t2}$ relation. It is also possible that y is a cause of x and $\text{cov}(y_{t1},y_{t2} \neq 0)$. Similar problems are evident too when splitting rating sources.

outcomes (at the follower, leader, or unit level), which may correlate with the modelled style, it is important to also control for competing correlated leadership styles in predicting an outcome (e.g., see Banks, McCauley, et al., 2016; Fischer, et al., 2016; Hoch, et al., 2016); failure to control for these competing constructs will engender omitted variable bias and does not inform us of the incremental validity of the construct either.

There are so many other endogeneity-related issues too that are often not adequately dealt with; interested readers should refer to some introductory sources that discuss endogeneity and how to deal with it (Antonakis, Bendahan, et al., 2014; Bascle, 2008; B. H. Hamilton & Nickerson, 2003; Larcker & Rusticus, 2010). I list a few more major sources of endogeneity bias here; these concern typical empirical articles, though they can be used in the context of an exploratory framework too (though paying particular attention to issues like capitalization on chance and out of sample prediction are very important here, Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015):

- Using HLM (random-effects) incorrectly (Halaby, 2004; McNeish, Stapleton, & Silverman, 2016) or failing to partial out fixed effects by including the cluster means of the within variables (Antonakis, et al., 2010; Rabe-Hesketh & Skrondal, 2008);
- Comparing groups when selection to group is endogenous (Bascle, 2008; Certo, Busenbark, Woo, & Semadeni, 2016; Li, 2013) or based known a cut-off but not modeling this assignment via a regression discontinuity design (Cook, 2008);
- Estimating mediation models via Baron-Kenny or the Preacher-Hayes type method, but not comparing estimates to an instrumental-variable estimator (Antonakis, et al., 2010; Shaver, 2005)—bootstrapping an inconsistent estimate does not make it consistent;

- Sampling on the dependent variable and not accounting for survival bias or other forms of selection effects (Denrell, 2003, 2005);
- Using self-selected samples or so-called "snowball samples" (Marcus, Weigelt, Hergert, Gurt, & Gelléri, 2016) and not reflecting on how samples can produce bias (Fiedler, 2000);
- Using partial least squares (PLS)—editors are desk rejecting articles that use PLS (Guide Jr & Ketokivi, 2015) and I will do likewise because PLS should ever be used in applied work (Rönkkö, McIntosh, Antonakis, & Edwards, 2016). Authors should use 2SLS if they require to use a limited information estimator;
- Ignoring overidentification tests (i.e., the χ^2 test of fit) in simultaneous or structural equation models (Ropovik, 2015)—such models have biased estimates (Bollen, Kirby, Curran, Paxton, & Chen, 2007; Hayduk, 2014; McIntosh, 2007); indexes of fit (CFI, RMSEA) should not be trusted because they depend on idiosyncrasies of the model (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Marsh, Hau, & Wen, 2004; Savalei, 2012). Note, *I will consider articles that fail the χ^2 overidentification test*, as long as the potential source of failure is acknowledged with a caveat too that estimates may be biased, and that recommendations are made to develop and test a better model in future research.

The list goes on (for some checklists see Antonakis, et al., 2010). Thus authors should find creative ways to harness exogenous variance in models. Beyond the introductory sources I referred to earlier, there are also several other accessible sources that will be useful to authors who wish to read more about endogeneity and quasi-experimental designs (see Angrist & Pischke, 2014; Antonakis, et al., 2010; Bascle, 2008; Clougherty, Duso, & Muck, 2016; Cook, Shadish, & Wong, 2008; G. J. Duncan, Magnusson, & Ludwig, 2004; Gennetian, Magnuson, & Morris, 2008; B. H. Hamilton & Nickerson, 2003; Shadish & Cook, 1999; Shadish, Cook, &

Campbell, 2002). And, some might be tempted to say that the problems of endogeneity are overblown and all this is a matter of opinion. It is not. Estimators that have been mathematically proven to be consistent for particular data and conditions must be used when relevant (e.g., Rubin, 1977).

Note too that methodological rigorous articles that deal correctly with endogeneity issues tend to be more cited (Antonakis, Bastardoz, et al., 2014; Bergh, et al., 2006). Still, that is not to say that journals should never consider publishing studies that report relations that are just "correlational"; *they should consider these for publication if the correlations observed have important implications for future research* and if authors clearly identify correlational evidence for what it is (do not try to sell it as causal).

Be transparent with data, methods, and reporting

Transparency is a virtue that should be rewarded in the publication process. Transparency, whether for qualitative or quantitative studies concerns the data gathered, the design, the analyses methods, and the results. Transparency begins with declaring all data that were gathered and their stopping rules, along with all treatments, and all experiments, and not selectively reporting statistically significant findings (see Nosek, et al., 2012; Simmons, Nelson, & Simonsohn, 2011).

Summary data for manuscripts using a quantitative design should be reported in the form of a correlation matrix, ideally at the item level, with means and standard deviations (for experimental data, binary variables for conditions should be included in the correlation matrix), or in another appropriate form. For studies that do not have a large sample size, authors should include the data in an Appendix table. To the extent possible, authors should consider posting data on permanent repositories (Stodden et al., 2016). Authors of meta-analysis should include in an Appendix all studies used in the meta-analysis along with sample sizes and effect sizes of

these articles; they should also conduct appropriate tests to determine if there is possible publication bias (Kepes, Banks, & Oh, 2014).

In terms of reporting estimates, it is important to fully report the size of the effect and the uncertainty of the estimate. The standard error carries the key information needed and must be reported along with the exact *p*-value (unless the *p*-value is very small in which case $p < .0001$ will suffice); confidence intervals can be reported too in tables, as well as in text for key findings. As for measure of effect, elasticities or semi-elasticities, or the like (depending on the metrics used, e.g., a change of 1 unit in *x* increases *y* by 33.2%), standardized estimates (if warranted and if measurement errors are removed from *y* and *x*), or some other metric of effect (odds ratio, incidence response ratio, etc.) can help establish the economic or practical significance of the results, as is plotting results and showing marginal effects, particularly for non-linear models.

Reporting transparency should be such that for quantitative studies, researchers should be able to reanalyze the same data and produce the same results. As concerns this journal, in all cases, raw data whether quantitative or qualitative must be made available to editor if requested; thus, authors should make appropriate provisions to have the necessary authorization from institutional bodies or sponsors to share the data with the editor in an anonymized/redacted manner. The study should fully describe its methods and procedures so that other researchers are able to faithfully reproduce it (see Atwater, et al., 2014; Cortina, et al., 2016; Goodman, Fanelli, & Ioannidis, 2016; Nosek et al., 2015); computational steps taken to arrive at analyzed datasets and results, as well as programing and needed computer code must be fully disclosed (Atwater, et al., 2014; Stodden, et al., 2016). Finally, the limitations of the article should be honestly assessed and discussed.

Declare conflicts of interest

Oftentimes researchers in our field do not declare conflicts of interest defined as "a set of conditions in which professional judgment concerning a primary interest [e.g.,] . . . the validity of research . . . tends to be unduly influenced by a secondary interest (such as financial gain)" (Thompson, 1993, p. 573). Authors must declare conflicts of interest if they are reporting on a questionnaire measure from which they stand to gain commercially at the time of review or in the future (by selling it via a publisher). Other instances can include reporting on consulting methods, training programs or methods, or software. This issue is very salient in other disciplines like in the medical sciences (Bekelman, Li, & Gross, 2003; Ioannidis, 2014; Papanikolaou et al., 2001; Thompson, 1993); it is becoming more prevalent in the social sciences as attested by ethics polices and guidelines for authors at many journals.

Be an honest broker

Our goal as scientists is to do good science and to report fully and accurately; it is not to push an ideological agenda. Sometimes we may not like what we find because it contradicts what we were expecting or hoping to find (Eagly, 2016). Take the case of the effect of women on corporate boards. Advocates for having more women may want to believe one thing (i.e., that more women means better performance); however, well-done science shows that it may not matter or may even reduce performance (Adams, 2016; Matsa & Miller, 2013). As mentioned by Eagly (2016, p. 213): "research literatures are often much more extensive than anticipated by most advocates, who may fix on particular studies that support their favored policy positions, with little concern for how typical, generalizable, or scientifically valid their findings are. Psychologists and other social scientists may be swept up by the excitement of seeing their findings used in advocacy and policy contexts." In other words, research must be done, reported, and synthesized in a responsible and honest way.

Conclusion

I have attempted a *tour de force* for how we can improve our scientific practices. Much of what I have written has focused on how better to manage the publication process, and how to best incentivize researchers. I focused on technical and methodological elements that we must improve on. I believe that the actions I suggest are a serious step forward to "vaccinating" our field against significosis, neophilia, theorrhea, arigorium, and disjunctivitis. In doing so, I believe that we will publish articles that will make useful contributions, that will be rigorous, and that will make a difference to research or practice.

Finding different ways to study leadership is what will take our knowledge base to the next level. In reflecting on the recent discovery of gravitational waves, an editorial in the journal *Science* stated: "measure what is measurable, and make measurable what is not so" (attributed to Galileo Galilei, see: Turner, 2016, p. 1243). Of course we are not physicists nor will we be able to elaborate eloquent theories like that of general relativity or to make precise enough measures to detect outcomes akin to gravitation ways. However, in most instances we can accomplish the following steps for a particular phenomenon:

1. conceive of and describe a construct, its different states and forms
2. explain how the construct is affected or can affect other constructs, whether causally or in a process
3. identify how the system of relations is bounded by space or time
4. demonstrate how the studied phenomenon can solve problems
5. build an explanatory bridge of this phenomenon to another discipline

Developing different ways to measure constructs and their outcomes in a rigorous and reliable way, coding or quantifying the variables, exploring and observing them in situ, linking them to other variables, experimenting with them, testing their limits, showing how they affect practice, as well as understanding their proximal or ultimate causes is what we should be

pursuing, and this a cohesive way. Each of these steps could be a study in its own right, would take years to fully understand, and might not have immediate applications; yet the steps together—moving from basic to applied research and then reaching out to another discipline—each help illuminate the nature of the phenomenon.

Moreover, beyond doing our science better and more creatively, we should also reflect on what we are doing, on whether we are asking the right questions, and trying to solve the right problems (cf. Ellemers, 2013). Such reflection must challenge what we do so that we never lose sight of whom we are serving. Our leitmotif should always be: So what? Is it rigorous? Will it make a difference?

Finally, by improving on how we do our science I believe that what we publish will make a difference and have an impact; it is not just about this journal but collectively what we do as scientists that matters. We must do our science *better*; never mind how impact is measured, rescaled, or conceptualized, and whether it considers impact on social media, policy, or elsewhere. Doing better on journal impact factor or other individual-level metrics are not goals in themselves; doing better science is.

The editorial team of this journal is passionate about doing good science and ensuring that what we publish matters. But, if we chase success we might never find it. If we all follow our passion and do *good science*, success will find its way to us.

References

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*(1), 12-15.

Acedo, F. J., Barroso, C., Casanueva, C., & Galán, J. L. (2006). Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis. *Journal of Management Studies, 43*(5), 957-983.

Achenbach, J. (2015). Why do many reasonable people doubt science. *National Geographic, 14*(5).

Adams, R. B. (2016). Women on boards: The superheroes of tomorrow? *The Leadership Quarterly, 27*(3), 371-386.

Adner, R., Polos, L., Ryall, M., & Sorenson, O. (2009). The case for formal theory. *Academy of Management Review, 34*(2), 201-208.

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351-371.

Aguinis, H., Suárez-González, I., Lannelongue, G., & Joo, H. (2012). Scholarly impact revisited. *The Academy of Management Perspectives, 26*(2), 105-132.

Amir, O., Ariely, D., Cooke, A., Dunning, D., Epley, N., Gneezy, U., et al. (2005). Psychology, behavioral economics, and public policy. *Marketing Letters, 16*(3), 443-454.

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in psychological laboratory: Truth or triviality? *Current Directions in Psychological Science, 8*(1), 3-9.

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton: Princeton University Press.

Antonakis, J. (2017). Editorial: The Future of The Leadership Quarterly. *The Leadership Quarterly*.

Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. (2016). Charisma: An ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organizational Behavior, 3*(1), 293-319.

Antonakis, J., Bastardoz, N., Liu, Y., & Schriesheim, C. A. (2014). What makes articles highly cited? *The Leadership Quarterly, 25*(1), 152-179.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*, 1086-1120.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), *The Oxford Handbook of Leadership and Organizations* (pp. 93-117). New York: Oxford University Press.

Antonakis, J., & Lalive, R. (2008). Quantifying scholarly impact: IQp versus the Hirsch h. *Journal of the American Society for Information Science and Technology, 59*(6), 956-969.

Atwater, L. E., Mumford, M. D., Schriesheim, C. A., & Yammarino, F. J. (2014). Retraction of leadership articles: Causes and prevention. *The Leadership Quarterly, 25*(6), 1174-1180.

Azoulay, P., Fons-Rosen, C., & Zivin, J. S. G. (2015). *Does science advance one funeral at a time?* : National Bureau of Economic Research.

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review, 14*(4), 496-515.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554.

Bamberger, P. A., & Pratt, M. G. (2010). Moving forward by looking back: Reclaiming unconventional research contexts and samples in organizational scholarship. *Academy of Management Journal, 53*(4), 665-671.

Banks, G. C., McCauley, K. D., Gardner, W. L., & Guler, C. E. (2016). A meta-analytic review of authentic and transformational leadership: A test for redundancy. *The Leadership Quarterly, 27*(4), 634-652.

Banks, G. C., O'Boyle, E., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., et al. (2016). Questions About Questionable Research Practices in the Field of Management A Guest Commentary. *Journal of Management, 42*(1), 5-20.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31*(3), 323-338.

Bansal, P., & Corley, K. (2011). The coming of age for qualitative research: Embracing the diversity of qualitative methods. *Academy of Management Journal, 54*(2), 233-237.

Barling, J., Weber, T., & Kelloway, E. K. (1996). Effects of transformational leadership training on attitudinal and financial outcomes: A field experiment. *Journal of Applied Psychology, 81*(6), 827-832.

Barnes, C. M., Dang, C., Leavitt, K., Guarana, C., & Uhlmann, E. L. (2016). Archival data in micro organizational research: A toolkit for moving to a broader set of topics. *Journal of Management, doi: 10.1177/0149206315604188*

Barrett, G. V. (1972). Research models of the future for industrial and organizational psychology. *Personnel Psychology, 25*(1), 1-17.

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization, 6*(3), 285-327.

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *Jama-Journal of the American Medical Association, 289*(4), 454-465.

Bennis, W. G., & O'Toole, J. (2005). How business schools lost their way. *Harvard Business Review, 83*(5), 96-+.

Bergh, D. D., Perry, J., & Hanke, R. (2006). Some predictors of SMJ article impact. *Strategic Management Journal, 27*(1), 81-100.

Bertamini, M., & Munafò, M. R. (2012). Bite-Size Science and Its Undesired Side Effects. *Perspectives on Psychological Science, 7*(1), 67-71.

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management. *Strategic Management Journal, 37*(2), 257-261.

Bettis, R. A., Gambardella, A., Helfat, C., & Mitchell, W. (2014). Quantitative empirical analysis in strategic management. *Strategic Management Journal, 35*(7), 949-953.

Bluhm, D. J., Harman, W., Lee, T. W., & Mitchell, T. R. (2011). Qualitative research in management: A decade of progress. *Journal of Management Studies, 48*(8), 1866-1891.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research, 36*(1), 48-86.

Bombari, D., Mast, M. S., Canadas, E., & Bachmann, M. (2015). Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges. *Frontiers in Psychology, 6*.

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*(3), 709-750.

Brown, D. J., & Lord, R. G. (1999). The utility of experimental research in the study of transformational/charismatic leadership. *The Leadership Quarterly, 10*(4), 531-539.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433-1436.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21*(10), 1363-1368.

Certo, S. T., Busenbark, J. R., Woo, H.-S., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal, 37*, 2639-2657.

Chan, H. F., Frey, B. S., Gallus, J., Schaffner, M., Torgler, B., & Whyte, S. (2014). Do the best scholars attract the highest speaking fees? An exploration of internal and external influence. *Scientometrics, 101*(1), 793-817.

Chatterji, A. K., Findley, M., Jensen, N. M., Meier, S., & Nielson, D. (2016). Field experiments in strategy research. *Strategic Management Journal, 37*(1), 116-132.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462-494.

Chetty, R. (2015). Behavioral economics and public policy: A pragmatic perspective. *American Economic Review, 105*(5), 1-33.

Clougherty, J. A., Duso, T., & Muck, J. (2016). Correcting for self-selection based endogeneity in management research: Review, recommendations and simulations. *Organizational Research Methods, 19*(2).

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics, 142*(2), 636-654.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2*(4), 447.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology, 71*(2), 179-184.

Cortina, J. M. (2015). The revolution with a solution: Culling the madness from our methods. *Presedential address: Society for Industrial and Organizational Psychology, Philadelphia, U.S.A.*

Cortina, J. M. (2016). Defining and operationalizing theory. *Journal of Organizational Behavior, 37*(8), 1142-1149.

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

Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (2016). Degrees of Freedom in SEM Are We Testing the Models That We Claim to Test? *Organizational Research Methods*, 1094428116676345.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

Day, D. V., & Dragoni, L. (2015). Leadership development: An outcome-oriented review based on time and levels of analyses. *Annual Review of Organizational Psychology and Organizational Behavior, 2*(1), 133-156.

Day, D. V., Fleenor, J. W., Atwater, L. E., Sturm, R. E., & McKee, R. A. (2014). Advances in leader and leadership development: A review of 25 years of research and theory. *The Leadership Quarterly, 25*(1), 63-82.

De Neve, J.-E., Mikhaylov, S., Dawes, C. T., Christakis, N. A., & Fowler, J. H. (2013). Born to lead? A twin design and genetic association study of leadership role occupancy. *The Leadership Quarterly, 24*(1), 45-60.

Delbridge, R., & Fiss, P. C. (2013). Editors' comments: Styles of theorizing and the social organization of knowledge. *Academy of Management Review, 38*(3), 325-331.

DeMonaco, H. J., Ali, A., & Von Hippel, E. A. (2005). The major role of clinicians in the discovery of off-label drug therapies. *Pharmacotherapy, 26*(3), 323-332.

Denrell, J. (2003). Vicarious learning, undersampling of failure, and the myths of management. *Organization Science, 14*(3), 227-243.

Denrell, J. (2005). Selection bias and the perils of benchmarking. *Harvard Business Review, 83*(4), 114-199.

Diamond, J. M., & Robinson, J. A. (2010). *Natural experiments of history*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Dixon-Woods, M., Shaw, R., Agarwal, S., & Smith, J. (2004). The problem of appraising qualitative research. *Quality & Safety in Health Care, 13*(3), 223-225.

Dubin, R. (1976). Theory building in applied areas. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology* (pp. 17-40). Chicago: Rand McNally.

Ductor, L. (2015). Does Co-authorship Lead to Higher Academic Productivity? [Article]. *Oxford Bulletin of Economics and Statistics, 77*(3), 385-407.

Duncan, G. J., Magnusson, K. A., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development, 1*, 59-80.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology, 30*(3), 257-303.

Durand, R., & Vaara, E. (2009). Causation, counterfactuals, and competitive advantage. *Strategic Management Journal, 30*(12), 1245-1264.

Eagly, A. H. (2016). When passionate advocates meet research on diversity: Does the honest broker stand a chance? *Journal of Social Issues, 72*(1), 199-222.

Eden, D. (2017). Field experiments in organizations. *Annual Review of Organizational Psychology and Organizational Behavior, http://dx.doi.org/10.1146/annurev-orgpsych-041015-062400*.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13*, 668-689.

Edwards, J. R., Berry, J. W., & Stewart, V. (2016). Bridging the great divide between theoretical and empirical management research. *Working paper. Kenan-Flagler Business School. University of North Carolina*.

Edwards, J. R., & Christian, M. S. (2014). Using accumulated knowledge to calibrate theoretical propositions. *Organizational Psychology Review*, 2041386614535131.

Ellemers, N. (2013). Connecting the dots: Mobilizing theory to reveal the big picture in social psychology (and why we should do this). *European Journal of Social Psychology, 43*(1), 1-8.

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing, 12*(1), 77-84.

Falk, A., Meier, S., & Zehnder, C. (2013). Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association, 11*(4), 839-852.

Fanelli, D., & Larivière, V. (2016). Researchers' Individual Publication Rate Has Not Increased in a Century. *PLoS ONE, 11*(3), e0149504.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories. *Perspectives on Psychological Science, 7*(6), 555-561.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*(4), 659-676.

Fischer, T., Dietz, J., & Antonakis, J. (2016). Leadership process model: A review and synthesis. *Journal of Management, http://dx.doi.org/10.1177/0149206316682830*.

Fiske, S. T. (1995). Social cognition. In A. Tesser (Ed.), *Advanced Social Psychology* (pp. 149-193). Boston: McGraw-Hill.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505.

Friedrich, T. L., Vessey, W. B., Schuelke, M. J., Mumford, M. D., Yammarino, F. J., & Ruark, G. A. (2014). Collectivistic leadership and George C. Marshall: A historiometric analysis of career events. *The Leadership Quarterly, 25*(3), 449-467.

Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *Journal of Personality and Social Psychology, 85*, 453-466.

García-Pérez, M. A. (2016). Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement*.

Geddes, B. (2003). *Paradigms and sand castles: Theory building and research design in comparative politics*. Ann Arbor: University of Michigan Press.

Gennetian, L. A., Magnuson, K., & Morris, P. A. (2008). From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology, 44*(2), 381-394.

Gephart, R. P. (2004). Qualitative research and the Academy of Management Journal. *Academy of Management Journal, 47*(4), 454-462.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research, 37*(1), 3-30.

Gerring, J. (2007). *Case study research: Principles and practices*. New York: Cambridge University Press.

Gerring, J. (2012). *Social science methodology: A unified framework* (2nd ed.). Cambridge ; New York: Cambridge University Press.

Gerring, J., & McDermott, R. (2007). An experimental template for case study research. *American Journal of Political Science, 51*(3), 688-701.

Gibbert, M., & Ruigrok, W. (2010). The ''What'' and ''How'' of Case Study Rigor: Three Strategies Based on Published Work. *Organizational Research Methods, 13*(4), 710-737.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research. *Organizational Research Methods, 16*(1), 15-31.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine, 8*(341), 341ps312-341ps312.
Grimes, D. A., & Schulz, K. F. (2005). Compared to what? Finding controls for case-control studies. *The Lancet, 365*(9468), 1429-1433.
Grote, G. (2016). There is hope for better science. *European Journal of Work and Organizational Psychology*, 1-3.
Guide Jr, V. D. R., & Ketokivi, M. (2015). Notes from the Editors: Redefining some methodological criteria for the journal. *Journal of Operations Management, 37*, v-viii.
Haeseler, G., & Leuwer, M. (2003). High-affinity block of voltage-operated rat IIA neuronal sodium channels by 2,6 di-tert-butylphenol, a propofol analogue. *European Journal of Anaesthesiology, 20*(3), 220-224.
Hagger, M. S., & Chatzisarantis, N. L. D. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.
Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*, 507-544.
Hambrick, D. C. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal, 50*(6), 1346-1352.
Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. *Strategic Organization, 1*(1), 51-78.
Hamilton, D. (1991). Research papers: Who's uncited now? *Science, 251*(4989), 25-25.
Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009-1055.
Haslam, S. A., & McGarty, C. (2001). A 100 years of certitude? Social psychology, the experimental method and the management of scientific uncertainty. *British Journal of Social Psychology, 40*(1), 1-21.
Haslam, S. A., Reicher, S. D., Millard, K., & McDonald, R. (2015). 'Happy to have been of service': The Yale archive as a window into the engaged followership of participants in Milgram's 'obedience' experiments. *British Journal of Social Psychology, 54*(1), 55-83.
Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, Calif.: Sage.
Hayduk, L. A. (2014). Shame for disrespecting evidence: The personal consequences of insufficient respect for structural equation model testing. *BMC Medical Research Methodology, 14*, 124-124.
Herbst, D., & Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science, 350*(6260), 545-549.
Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodolgoical challenge for psychologists? *Behavioral and Brain Sciences, 24*(3), 383-403.
Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior, 18*(1), 59-92.
Hoch, J. E., Bommer, W. H., Dulebohn, J. H., & Wu, D. (2016). Do Ethical, Authentic, and Servant Leadership Explain Variance Above and Beyond Transformational Leadership? A Meta-Analysis. *Journal of Management*, 0149206316665461.
Hodgkin, L., Symonds, M., & Elgar, M. (2014). Leaders benefit followers in the collective movement of a social sawfly. *Proceedings of the Royal Society of London B: Biological Sciences, 281*(1796), 20141700.
Hogarth, R. M. (2005). The challenge of representative design in psychology and economics. *Journal of Economic Methodology, 12*(2), 253-263.

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, Sharking, and Tharking. *Journal of Management, 43*(1), 5-18.

House, R. J., & Aditya, R. N. (1997). The social scientific study of leadership: Quo vadis? *Journal of Management, 23*(3), 409-473.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124.

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645-654.

Ioannidis, J. P. (2014). How to make more published research true. *PLoS Med, 11*(10), e1001747.

Ioannidis, J. P. (2016a). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly, 94*(3), 485-514.

Ioannidis, J. P. (2016b). Why most clinical research is not useful. *PLoS Med, 13*(6), e1002049.

Ioannidis, J. P., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: evaluation and improvement of research methods and practices. *PLoS Biol, 13*(10), e1002264.

Jacoby, S. (2009). *The age of American unreason* (1st Vintage Books ed.). New York: Vintage Books.

Jamison, J., Karlan, D., & Schechter, L. (2008). To deceive or not to deceive: The effect of deception on behavior in future laboratory experiments. *Journal of Economic Behavior & Organization, 68*(3–4), 477-488.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (in press). On the reproducibility of psychological science. *Journal of the American Statistical Association, doi: 10.1080/01621459.2016.1240079*.

Kacmar, K. M., & Whitfield, J. M. (2000). An additional rating method for journal articles in the field of management. *Organizational Research Methods, 3*(4), 392-406.

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS ONE, 10*(8), e0132382.

Karabag, S. F., & Berggren, C. (2016). Misconduct, Marginality and Editorial Practices in Management, Business and Economics Journals. *PLoS ONE, 11*(7), e0159492.

Kepes, S., Banks, G. C., & Oh, I.-S. (2014). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology, 29*(2), 183-203.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3 ed.). New York: Holt, Rinehart and Winston.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Klein, D. F. (2008). The loss of serendipity in psychopharmacology. *Journal of the American Medical Association, 299*(9), 1063-1065.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *The American Economic Review, 105*(5), 491-495.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Landis, R. S., James, L. R., Lance, C. E., Pierce, C. A., & Rogelberg, S. G. (2014). When is nothing something? Editorial for the null results special issue of Journal of Business and Psychology. *Journal of Business and Psychology, 29*(2), 163-167.

Lapierre, L. M., Edwards, J. R., Oswald, F. L., Shockley, K. M., & Landis, R. S. (2017). Toward integrating deductive and inductive research using quantitative methods. *Working paper. Kenan-Flagler Business School. University of North Carolina.*

Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics, 49*(3), 186-205.

Larivière, V., & Costas, R. (2016). How Many Is Too Many? On the Relationship between Research Productivity and Impact. *PLoS ONE, 11*(9), e0162709.

Leavitt, K., Mitchell, T. R., & Peterson, J. (2010). Theory Pruning: Strategies to Reduce Our Dense Theoretical Landscape. *Organizational Research Methods, 13*(4), 644-667.

Leigh, J. P. (1988). Assessing the importance of an independent variable in multiple-regression--Is stepwise unwise? *Journal of Clinical Epidemiology, 41*(7), 669-677.

Li, M. (2013). Using the propensity score method to estimate causal effects: A review and practical guide. *Organizational Research Methods, 16*(2), 188-226.

Lin, L. I. (1989). A concordance correlation-coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255-268.

Locke, E. A. (2007). The case for inductive theory building. *Journal of Management, 33*(6), 867-890.

Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., von Bastian, C. C., Carruth, N. P., et al. (2016). No evidence of the ego-depletion effect across task characteristics and individual differences: A pre-registered study. *PLoS ONE, 11*(2), e0147770.

Maccallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure-analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science, 31*, 323-326.

Maddox, J. (1988). When to believe the unbelievable. *Nature, 333*(6176), 787.

Maddox, J., Randi, J., & Stewart, W. W. (1988). "High-dilution" experiments a delusion. *Nature, 334*(6180), 287-290.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research. *Perspectives on Psychological Science, 7*(6), 537-542.

Marcus, B., Weigelt, O., Hergert, J., Gurt, J., & Gelléri, P. (2016). The use of snowball sampling for multi source organizational research: Some cause for concern. *Personnel Psychology*, n/a-n/a.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320 - 341.

Marušić, A., Bošnjak, L., & Jerončić, A. (2011). A Systematic Review of Research on the Meaning, Ethics and Practices of Authorship across Scholarly Disciplines. *PLoS ONE, 6*(9), e23477.

Mathieu, J. E. (2016). The problem with [in] management theory. *Journal of Organizational Behavior, 37*(8), 1132-1141.

Matsa, D. A., & Miller, A. R. (2013). A female style in corporate leadership? Evidence from quotas. *American Economic Journal: Applied Economics, 5*(3), 136-169.

McHugh, K. A., Yammarino, F. J., Dionne, S. D., Serban, A., Sayama, H., & Chatterjee, S. (2016). Collective decision making, leadership, and collective intelligence: Tests with agent-based simulations and a Field study. *Leadership Quarterly, 27*(2), 218-241.

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*(5), 859-867.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*.

Mellor, S., & Mark, M. M. (1998). A quasi-experimental design for studies on the impact of administrative decisions: Applications and extensions of the regression-discontinuity design. *Organizational Research Methods, 1*(3), 315-333.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economics Statistics, 13*(2), 151-161.

Miller, C. C., & Bamberger, P. (2016). Exploring emergent and poorly understood phenomena in the strangest of places: The footprint of discovery in replications, meta-analyses, and null findings. *Academy of Management Discoveries, 2*(4), 313-319.

Miller, D. (2007). Paradigm prison, or in praise of atheoretic research. *Strategic Organization, 5*(2), 177-184.

Mitchell, G. (2012). Revisiting truth or triviality. *Perspectives on Psychological Science, 7*(2), 109-117.

Mooney, C. (2005). *The Republican war on science*. New York: Basic Books.

Murtaugh, P. A. (2014). In defense of P values. *Ecology, 95*(3), 611-617.

Niccolai, L. M., Ogden, L. G., Muehlenbein, C. E., Dziura, J. D., Vázquez, M., & Shapiro, E. D. (2007). Methodological issues in design and analysis of a matched case–control study of a vaccine's effectiveness. *Journal of Clinical Epidemiology, 60*(11), 1127-1131.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Nosek, B. A., Alter, G., Banks, G., Borsboom, D., Bowman, S., Breckler, S., et al. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia. *Perspectives on Psychological Science, 7*(6), 615-631.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). [journal article]. *Behavior Research Methods, 48*(4), 1205-1226.

Nuzzo, R. (2015). Fooling ourselves. *Nature, 526*(7572), 182-185.

O'Boyle, E. H., Banks, G., & Gonzalez-Mulé, E. (2017). The Chrysalis Effect. *Journal of Management, 43*(2), 376-399.

O'Boyle, E. H., Banks, G., Walter, S., Carter, K., & Weisenberger, K. (2015). *What Moderates Moderators? A Meta-Analysis of Interactions in Management Research.* Paper presented at the Academy of Management Proceedings.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716-4711/4718.

Orne, M. T. (2009). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifacts in Behavioral Research* (pp. 110-137). New York: Oxford University Press.

Ortmann, A., & Hertwig, R. (1997). Is deception acceptable? *American Psychologist, 52*(7), 746-747.

Ortmann, A., & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics, 5*, 111-131.

Papanikolaou, G. N., Baltogianni, M. S., Contopoulos-Ioannidis, D. G., Haidich, A.-B., Giannakakis, I. A., & Ioannidis, J. P. (2001). Reporting of conflicts of interest in guidelines of preventive and therapeutic interventions. *BMC Medical Research Methodology, 1*(1), 1.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science. *Perspectives on Psychological Science, 7*(6), 528-530.

Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *The Academy of Management Review, 18*(4), 599-620.

Pfeiffer, T., Bertram, L., & Ioannidis, J. P. (2011). Quantifying selective reporting and the Proteus phenomenon for multiple datasets with similar bias. *PLoS ONE, 6*(3), e18362.

Podsakoff, N. P., Podsakoff, P. M., MacKenzie, S. B., & Klinger, R. L. (2013). Are we really measuring what we say we're measuring? Using video techniques to supplement traditional construct validation procedures. *Journal of Applied Psychology, 98*(1), 99-113.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 89*(5), 879-903.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control It. *Annual Review of Psychology, 63*(1), 539-569.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods*.

Popper, K. R. (1989). *Conjectures and refutations: The growth of scientific knowledge* (5th ed.). London ; New York: Routledge.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science, 26*(5), 653-656.

Reeb, D., Sakakibara, M., & Mahmood, I. P. (2012). From the Editors: Endogeneity in international business research. *Journal of International Business Studies, 43*(3), 211-218.

Robinson, D. H., & Toledo, A. H. (2012). Historical development of modern anesthesia. *Journal of Investigative Surgery, 25*(3), 141-149.

Rönkkö, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management, 47–48*, 9-27.

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology, 6*(1715).

Rosenthal, R., & Rosnow, R. L. (2009). *Artifacts in behavioral research : Robert Rosenthal and Ralph L. Rosnow's classic books: A re-issue of Artifact in behavioral research, Experimenter effects in behavioral research and The volunteer subject*. New York: Oxford University Press.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2*(1), 1-26.

Sandström, U., & van den Besselaar, P. (2016). Quantity and/or Quality? The Importance of Publishing Many Papers. *PLoS ONE, 11*(11), e0166149.

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement, 72*(6), 910-932.

Shadish, W. R., & Cook, T. D. (1999). Comment-design rules: More steps toward a complete theory of quasi-experimentation. *Statistical Science, 14*(3), 294-300.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
Shaver, J. M. (2005). Testing for mediating variables in management research: Concerns, implications, and alternative strategies. *Journal of Management, 31*(3), 330-353.
Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psycholgoy experiments: But are they necessary? *Personality and Social Psychology Review, 2*(3), 218-226.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science, 22*(11), 1359-1366.
Simmons, J. P., & Simonsohn, U. (in press). Power posing: P-Curving the evidence. *Psychological Science*.
Simonton, D. K. (2003). Qualitative and quantitative analyses of historical data. *Annual Review of Psychology, 54*, 617-640.
Simonton, D. K. (2009). The "other IQ": Historiometric assessments of intelligence and related constructs. *Review of General Psychology, 13*(4), 315.
Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*, 160384.
Smith, J. E., Gavrilets, S., Mulder, M. B., Hooper, P. L., Mouden, C. E., Nettle, D., et al. (2016). Leadership in mammalian societies: Emergence, distribution, power, and payoff. *Trends in Ecology & Evolution, 31*(1), 54-66.
Smith, K. M., & Apicella, C. L. (2016). Winners, losers, and posers: The effect of power poses on testosterone and risk-taking following competition. *Hormones and Behavior, http://dx.doi.org/10.1016/j.yhbeh.2016.11.003*.
Stock, J. H., & Watson, M. W. (2011). *Introduction to econometrics* (3rd ed.). Boston: Addison-Wesley.
Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., et al. (2016). Enhancing reproducibility for computational methods. *Science, 354*(6317), 1240-1241.
Sturm, R. E., & Antonakis, J. (2015). Interpersonal power: A review, critique, and research agenda. *Journal of Management, 41*(1), 136-163.
Tavris, C., & Aronson, E. (2007). *Mistakes were made (but not by me): Why we justify foolish beliefs, bad decisions, and hurtful acts*. New York: Harcourt.
Thompson, D. F. (1993). Understanding financial conflicts of interest. *New England Journal of Medicine, 329*(8), 573-576.
Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623-1626.
Tonidandel, S., King, E. B., & Cortina, J. M. (in press). Big data methods. *Organizational Research Methods*, doi:10.1177/1094428116677299.
Tullock, G. (2001). A comment on Daniel Klein's "A plea to economists who favor libert.". *Eastern Economic Journal, 27*(2), 203.
Turner, M. S. (2016). Editorial: Throwing deep. *Science, 351*(6279), 1243.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.
van der Meij, L., Schaveling, J., & van Vugt, M. (2016). Basal testosterone, leadership and dominance: A field study and meta-analysis. *Psychoneuroendocrinology, 72*, 72-79.

van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic-transformational leadership research: Back to the drawing board? *The Academy of Management Annals, 7*(1), 1-60.

Varian, H. R. (1997). How to build an economic model in your spare time. *The American Economist, 41*(2), 3-10.

Vermeulen, F. (2005). On rigor and relevance: Fostering dialectic progress in management research. *Academy of Management Journal, 48*(6), 978-982.

Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., van de Velde, B., et al. (2015). Blinded by the light: How a focus on statistical "significance" may cause p-value misreporting and an excess of p-values just below. 05 in communication science. *Communication Methods and Measures, 9*(4), 253-279.

Waldman, D. A., Wang, D., & Fenters, V. (2016). The added value of neuroscience methods in organizational research. *Organizational Research Methods*.

Weiner-Levy, N., & Popper-Giveon, A. (2013). The absent, the hidden and the obscured: Reflections on "dark matter" in qualitative research. *Quality & Quantity, 47*(4), 2177-2190.

Wicklund, R. A. (1990). Zero-variable theories in the analysis of social phenomena. *European Journal of Personality, 4*(1), 37-55.

Wilson, E. O. (1998). *Consilience: The unity of knowledge* (1st ed.). New York: Knopf : Distributed by Random House.

Woehr, D. J., & Lance, C. E. (1991). Paper people versus direct observation: An empirical examination of laboratory methodologies. *Journal of Organizational Behavior, 12*(5), 387-397.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wright, P. M. (2016). Making great theories. *Journal of Management Studies*.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science, 316*(5827), 1036-1039.

Yip, G. M. S., Chen, Z.-W., Edge, C. J., Smith, E. H., Dickinson, R., Hohenester, E., et al. (2013). A propofol binding site on mammalian GABAA receptors identified by photolabeling. *Nature Chemical Biology, 9*(11), 715-720.

Zaccaro, S. J., & Horn, Z. N. J. (2003). Leadership theory and practice: Fostering an effective symbiosis. *The Leadership Quarterly, 14*(6), 769-806.

Zehnder, C., Herz, H., & Bonardi, J.-P. (in press). A productive clash of cultures: Injecting economics into leadership research. *The Leadership Quarterly,* http://dx.doi.org/10.1016/j.leaqua.2016.10.004.

Zelditch, M. (1969). Can you really study an army in the laboratory. In A. Etzioni & E. Lehman (Eds.), *A sociological reader on complex organizations* (pp. 528-539). New York: Holt, Rinehart and Winston.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics, 13*(1), 75-98.