4

What's in a *p*? Reassessing Best Practices for Conducting and Reporting Hypothesis-Testing Research

Klaus E. Meyer, Arjen van Witteloostuijn, and Sjoerd Beugelsdijk

Introduction

> The value for which $p = 0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. (Fisher 1925: 45)

K. E. Meyer (✉)
Western University, London, ON, Canada
e-mail: kmeyer@ivey.uwo.ca; kmeyer@ceibs.edu

A. van Witteloostuijn
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: a.van.witteloostuijn@vu.nl

S. Beugelsdijk
University of Groningen, Groningen, The Netherlands
e-mail: s.beugelsdijk@rug.nl

L. Eden et al. (eds.), *Research Methods in International Business*, JIBS Special Collections, https://doi.org/10.1007/978-3-030-22113-3_4

If one were to believe all results published in business journals, scholars would be able to predict the future (almost) perfectly. In the vast majority of the papers published, almost all theoretically derived hypotheses are empirically supported. For example, of the 711 hypotheses tested in articles published in the *Journal of International Business Studies (JIBS), Strategic Management Journal (SMJ)*, and *Organization Science* in the 2016 volumes, about 89% find empirical support for the theoretical predictions. In a similar exercise in 1959, Sterling reported a statistical significance percentage of 97% (Sterling 1959). The above interpretation of scholars as futurologists assumes that published research is representative of the population of all completed studies on a specific topic. There is plenty of evidence that this is not the case (Sterling 1959; Ioannidis 2005). What is known as the "file-drawer problem" is very common: scientific studies with negative or nil-results often remain unpublished (Rosenthal 1979; Rothstein et al. 2005).

Moreover, accumulating evidence suggests that authors actively engage in pushing significance levels just below the magic threshold of $p = 0.05$, a phenomenon referred to as '*p*-hacking' or 'search for asterisks' (Bettis 2012; Brodeur et al. 2016). Similarly, some authors appear to engage in HARKing, which stands for Hypothesizing After the Results are Known (Bosco et al. 2016; Kerr 1998). The problem of both practices is that the reported significance levels are misleading because readers are given no information how many nulls and negatives ended up in the research dustbin along the way. Editorial boards, reviewers, and authors are increasingly aware of the challenge to minimize 'the search for asterisks,' In this editorial, we document recent initiatives and suggest ten concrete guidelines in order to minimize the risk of reporting false positives (i.e., type I errors), and more generally improve the quality of hypothesis-testing research and statistical reporting in our field.

Our editorial responds to the recent surge of pleas to change extant research practices, across a wide variety of disciplines, including business studies. For instance, Bettis et al. (2016a, b) in strategic management, Barley (2016) in administrative sciences, Aguinis et al. (2010) in organizational studies, and van Witteloostuijn (2016) in international business raise major concerns about the current state of affairs. These recent pleas, triggered by hot debates in disciplines such as medicine (Crosswell et al. 2009; Ioannidis 2005; Lexchin et al. 2003) and psychology (Gigerenzer 2004; John et al. 2012; Simmons et al. 2011), fit into a long tradition of work highlighting the need for the scholarly community to continuously improve its research practices (e.g., Sterling 1959; Rosenthal 1979). In this editorial, we focus in particular on calls for more transparency regarding the process of empirical research, and hence more accurate reporting and comprehensive interpretation of empirical results. Our aim is to derive from the ongoing discussions, a set of concrete and actionable treatments, which we translate into guidelines and best practices for *JIBS* authors.

Our starting point is the observation that current practices stimulate the publication of false positives. This argument is anything but new and the reasons for this problem have been extensively analysed, a particularly forceful voice being Ioannidis's (2005, 2012). The root of the problem is the publication bias, caused by journals seeking theoretical novelty with empirical confirmation, in combination with counterproductive university-level career incentives focused on publications in a limited number of journals (for a recent summary, see van Witteloostuijn 2016). However, a recent upsurge of scandals such as Stapel's data-fabricating misconduct in social psychology (*New York Times* 2011) triggered a powerful movement toward changing the ways in which the scientific community has institutionalized practices that stimulate rather than discourage such behaviour. Examples, among many, are orchestrated replication projects (e.g., *Open Science Collaboration* 2015) and journal repository requirements (e.g., the *American Economic Review*). In the business studies domain, the recent change of statistical reporting guidelines by the *Strategic Management Journal* (Bettis et al. 2016a, b), swiftly followed by *Organization Science* and other journals (see, e.g., Lewin et al. 2016), is a clear signal that research practices are currently being revised and updated.

As the leading journal in our field, *JIBS* is committed to engage in this debate, being part of this wider movement continuously (re)assessing the criteria for what counts as rigorous empirical research. We hope that our suggestions will help further improving the work published in (international) business, as well as in triggering an ongoing reflection on what best research practices entail. To do so, we propose ten guidelines that are concrete and actionable. These guidelines serve as suggestions (not as fixed rules), providing direction for authors submitting papers employing quantitative hypothesis-testing methods. These guidelines should not result in a uniform straightjacket, but help advance research practices and stimulate the search for solutions to shortcomings in contemporary practice. Research best practices are not set in stone, but experience shows that a set of benchmarks for both researchers and reviewers can be very helpful to push the quality bar of research upward.

Challenges to Current Practice

The Focus on *p*-Values and False Positives

The null hypothesis significance testing practice was introduced by Fisher (1925) to distinguish between interesting relationships and noise. Null hypothesis significance testing has quickly become the norm in social sciences, including business studies. Before computers and software packages such as SPSS and STATA became widely available, the *p*-values associated with specific test statistics related to a particular relationship were looked up in a statistical table. As *p*-values were given for a limited set of cutoff values (particularly $p = 0.10$, $p = 0.05$, and $p = 0.01$), a practice emerged to report *p*-values with respect to these benchmarks (e.g., $p = 0.05$), and to indicate the significant estimates with *, ** or ***. Fisher (1925) suggested, somewhat arbitrarily, using $p = 0.05$ as the most appropriate cutoff level. With increased computing power, however, scholars became able to calculate exact *p*-values for even the most advanced statistical models. But due to path dependency, the old asterisks habit remained in place.

Despite the importance and influence of Fisher's work, and the intuitive attractiveness of using a simple cutoff value, the focus on *p*-values is not without its negative external effects. Particularly, the focus on *p*-values leads to publication bias. It has always been the case that journals have an interest in publishing interesting results – i.e., significant estimates – and not noise (to paraphrase Fisher), but the introduction of the publish-or-perish culture appears to have increased the publication bias. It has been argued that this development is due to a counterproductive academic reward structure, arising from the combination of top-tier journals' preference for 'statistically significant results' and a highly competitive tenure-track system in many universities that relies disproportionately on top-tier journal publications (Bedeian et al. 2010; Pashler and Wagenmakers 2012).

This reward structure encourages practices inconsistent with statistical best practice (Wasserstein and Lazar 2016), specifically *ex post* writing of hypotheses supposedly *ex ante* tested, also referred to as HARKing (Kerr 1998), and of manipulating of empirical results to achieve threshold values, varyingly referred to as *p*-hacking (Head et al. 2015; Simmons et al. 2011) star wars (Brodeur et al. 2016), and searching for asterisks (Bettis 2012). The heavy focus on significant effects opens the door to a variety of questionable (and occasionally plain bad) practices, some of which we discuss below in greater detail. Most fundamentally, such approaches are inconsistent with Popper's (1959) falsification criterion, which is the philosophical foundation for conducting hypothesis tests in the first place (van Witteloostuijn 2016). As a consequence, the reliability and validity of cumulative work are not as high as they could be without biases in the publication process.

The publication bias arises from two practices. First, papers reporting significant relationships are more likely to be *selected* for publication in journals, leading to a bias towards tests rejecting the null hypothesis. Second, authors 'fine-tune' their regression analysis to turn marginally nonsignificant relations (those just above $p = 0.01$, $p = 0.05$ or $p = 0.10$) to significant relations (i.e., just below these thresholds), which causes an *inflation* of significance levels in (published and unpublished) empirical tests.[1] As said, these biases in article selection and significance inflation are anything but new (Sterling 1959), and evidence for such unbalances has been firmly established in sciences (Head et al. 2015).

The selection bias has received a great deal of attention in medical research, obviously because of the immediate medical and societal implications of prescribing medications based on possibly flawed results. In the medical field, the problem of selection bias is exacerbated by the intricate relationships between pharmaceutical companies and research (Lexchin et al. 2003). However, the problem of publication bias for social sciences is not to be underestimated either: exactly because in social sciences, reality and truth are partially socially constructed (or at least socially interpreted), public policies, managerial practices, and other practical implications derived from social science research can have great impact on many.

Selection bias can be found in economics (Brodeur et al. 2016), political science (Gerber et al. 2001), and psychology (Ferguson and Heene 2012) too. With the advent of meta-analytical techniques, it has become more and more common to explore the sensitivity of the results for a selection bias. Typically, it is found that – all else equal – the probability of finding significant regression coefficients in published articles is much higher than in working papers addressing the same topic (Gorg and Strobl 2001; Rothstein et al. 2005).[2] The variation in the severity of the selection bias across domains has been related to characteristics such as the size of the discipline and the degree of methodological consensus, as well as to the extent to which there is competition between theoretical predictions (Brodeur et al. 2016). Interestingly, the selection bias is lower when theory is contested (Doucouliagos and Stanley 2013), suggesting that academic debate remains critical even, or perhaps especially so, in the face of so-called stylized facts (i.e., established findings). Also, papers published by tenured and older researchers seem to "suffer" less from p-hacking (Brodeur et al. 2016), probably because for them career concerns are less of an issue.

In econometrics, the discussion of inflation bias goes back to the debate on pretesting in the 1970s and 1980s. It was generally acknowledged that as a result of running multiple tests, and leaving out insignificant variables, the final model typically includes focal variables with p-values that are inflated. One econometric strategy developed by Leamer in response to the discussion on the inclusion of control variables in the early 1980s has been to perform a so-called extreme-bounds analysis (Leamer 1985). The basic idea of this analysis is to analyse the consequences of changing the set of control variables for the estimated effect of x_i on a specific dependent variable. Instead of selecting a fixed set of control variables (that happen to give the lowest p-values), extreme-bounds analysis implies a series of regressions in which the coefficient of the variable of interest is estimated by changing the set of control variables (for an application, see Beugelsdijk et al. 2004). Although this is an interesting method that received follow-up especially in economics (Sala-i-Martin 1997; Angrist and Pischke 2010), but not in business studies, extreme-bounds analysis is a rather mechanical way to explore just one dimension of robustness: sensitivity of the coefficient of the variable of interest to (selective) inclusion of control variables. It remains vulnerable to "meta-level" *p*-hacking and inflated *p*-values because of its vulnerability to the selection of the set of control variables in the first place.

Moreover, *p*-hacking occurs not only by selecting control variables depending on results obtained, but takes many different forms and shapes (Bosco et al. 2016; Head et al. 2015; John et al. 2012). For example, the decision whether to drop or include influential observations may be biased if made after the initial analysis. Some even suggested that overoptimism among academic researchers is one of the reasons why too many false positives are reported (Mullane and Williams 2013). As editors, we also observed reviewers asking for changes that promote significance (and confirmation of hypotheses), a practice running counter to establishing the validity of empirical results.

The practice of *p*-hacking is problematic because it not only affects individual careers, but also erodes the reliability of scientific studies. Bettis et al. (2016a, b) illustrate the challenge as follows. Suppose three junior scholars test the same hypothesis. Scholars A and B find no significant results; they quickly move to other topics because 'not statistically significant' will not be published in top management journals. Scholar C finds a result significant at $p < 0.05$ level, which gets published in a high-impact outlet on the basis of which s/he receives tenure. The published result is treated as scientifically proven, and not challenged. Yet the actual evidence is that two out of three studies did not find a significant effect – and no one knows how many regressions scholar C ran in addition to the one with the significant effect. This problem is not unique for nonexperimental field work; experimental study designs are not immune to *p*-hacking either, as researchers may well stop their experiments once analysis yields a significant *p*-value.

The practice of *p*-hacking may have been a matter of pluralistic ignorance in the past (many may oppose these practices, but assume that others support them, leading to collective inaction and thereby sustained support). However, the increased publicity regarding these practices, which in some extreme situations of fraud have even led to legal cases (Bhattacharjee 2013), call for action. Scholars under pressure of 'publish or perish' face a slippery slope, moving from the subjectivity of 'sloppy science', to incomplete reporting that inhibits replication, to deliberate exclusion of key variables and/or observations, to manipulation of data, and to outright fabrication of data.

Globalization and the Internet facilitate the tracking of suspicious articles (*The Economist*, June 14, 2014), and amplify negative reputation effects after serious statistical fraud, for both authors and journals. Starting in 2010, the blog *RetractionWatch.com* discusses and reports on retractions of scientific papers. Excluding repeat offenders and adjusting for the growth of published medical and nonmedical literature, the number of articles retracted increased by a factor 11 between 2001 and 2010 (Grieneisen and Zhang 2012). One interpretation of this increase is that scientific inquiry is in crisis. Our interpretation is that open access, convergence of knowledge on statistics (partly thanks to the Internet), and increased awareness of publication ethics in tandem increase the pressure to adhere to proper statistical standards, enhance transparency and thereby boost the post-publication detection of poor practices.

The Biases and Misinterpretations of *p*-Logic Practices

The biases that cause inflated *p*-values (i.e., *p*-values that are lower than they "truly" are)[3] are problematic because the final result is research reporting too many false positives, which, in turn, lead to misguided advice to practice (e.g., Aguinis et al. 2010). As a simple yet powerful illustration, we took the last two years (2015–2016) of *JIBS, Organization Science*, and *SMJ*, and collected information on the *p*-values of all variables of interest in the estimated regression models. We followed the approach of Brodeur et al. (2016) and collected for all tests of a variable of interest in a hypothesis-testing paper information on the coefficient, reported *p*-values, and standard errors of the coefficient (or *t*-value when reported). The vast majority of the articles present the coefficient and the standard error; only few report *t*-values. We omit control variables.

For the three journals combined, this amounts to 313 articles and 5579 null hypothesis tests. This includes robustness tests (but excludes the ones published in online appendices). We do not round coefficients and standard errors, but use the full data as provided in the articles considered. Out of the 5579 hypothesis tests extracted from the three journals, 3897 are rejected at the $p < 0.10$ level, 3461 at the $p < 0.05$ level, and 2356 at $p < 0.01$ level. To obtain a homogenous sample, we transform the *p*-values into the equivalent *z*-statistics. A *p*-value of 0.05 becomes a *z*-statistic of 1.96. Following Brodeur et al. (2016), we simply construct the ratio of the reported coefficient and the standard error, assuming a standard normal distribution.[4]

The findings are visualized in Fig. 4.1. It shows the raw distribution of *z*-scores (*p*-values) in a histogram as well as the kernel density plot,

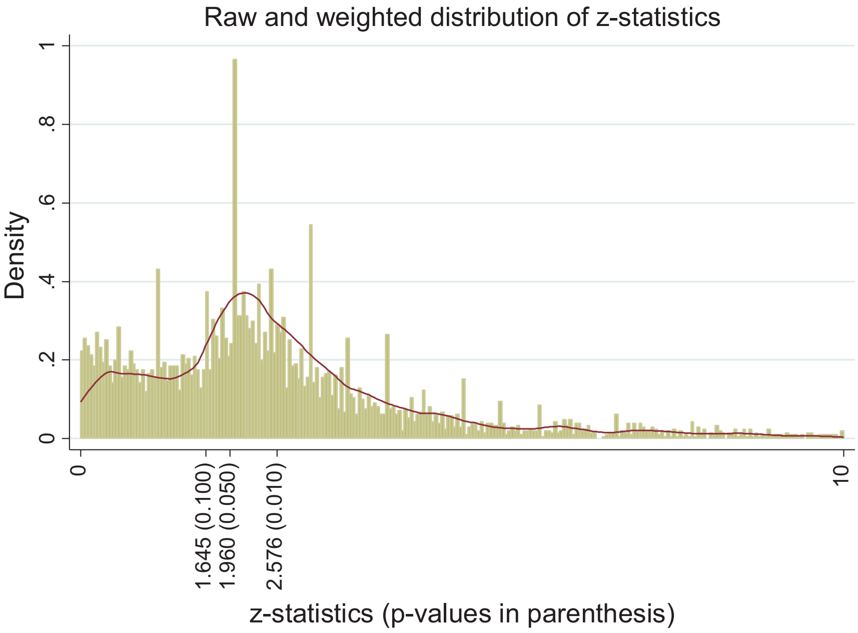

Fig. 4.1 Camel-shaped distribution of *p*-values in *JIBS, OrgScience* and *SMJ* (2015 and 2016). (Note: The graph shows the histogram as well as the kernel density plot of the weighted distribution of *z*-scores in all hypotheses testing articles published in *JIBS, Organization Science*, and *SMJ* in 2015 and 2016)

weighted for the number of hypotheses tested in an article. A kernel density plot is a nonparametric technique to visualize the underlying distribution of a continuous variable, in this case the distribution of *p*-values. It is nonparametric because it does not assume any underlying distribution such as a normal one. Intuitively, a kernel density plot can be seen as a sum of bumps. In plotting the kernel density plot, we weigh by number of hypotheses tests per article, because we want to give each article equal weight in the overall distribution. Some papers may have many hypotheses (e.g., Choi and Contractor 2016), whereas others may only have one or two (e.g., Husted et al. 2016). Separate graphs for *JIBS, SMJ* and *Organization Science* produce similar distributions and density plots (available upon request from the authors). Including or excluding robustness tests does not affect overall findings either.[5]

The shape of the figure is striking. The distribution of *p*-values in these three top management journals is not normally distributed, but has a camel-shaped distribution with a local maximum just above 1.96 (*p*-value is under 0.05), and a valley just left of 1.96. The combination of a spike just above the *p*-value of 0.05 and the valley just below in the distribution of *p*-values close to the critical value of 0.05 (critical from a reporting point of view) corresponds with similar findings in economics and psychology. Brodeur et al. (2016) also find such a camel-shaped distribution of *p*-values for top economics journals like the *American Economic Review, Quarterly Journal of Economics*, and *Journal of Political Economy.* Masicampo and Lalande (2012) report a significantly higher incidence of *p*-values just below $p = 0.05$ for the *Journal of Experimental Psychology*, *Journal of Personality and Social Psychology*, and *Psychological Science*. Hence such a skewed distribution of *p*-values in (these) business journals is no exception to the distribution of *p*-values in other disciplines. The finding is not the result of a selection bias (only significant results are published), because a similar exercise comparing conference papers and published papers in strategy research shows "an abundance of false or inflated findings" also for conference papers, suggesting that "even in early stage work, authors seem to filter results to find those that are statistically significant" (Goldfarb and King 2016: 169). Combined, this evidence is strongly suggestive of a possible inflation bias resulting from *p*-hacking.

Another reason why the focus on *p*-values leads to too many false positives is cumulative incidence of false positives. If a study was conducted in a strictly sequential manner, where first the hypotheses are developed and then a single test was conducted, then the *p*-value would indicate what we stated above. However, in practice, scholars often conduct many tests, and develop their theory *ex post* but present it as if the theory had been developed first. In part, this is driven by the review process as authors anticipate less favourable reactions of reviewers to nonsignificant results (Orlitzky 2012; Pfeffer 2007). Even within the review process, hypotheses may be added or dropped, often on request of reviewers (Bedeian et al. 2010; Pfeffer 2007). However, if a 'best' result is selected from many regressions, then the *p*-value overstates the degree of support for the theoretical argument. In the extreme case, if in truth there is no effect but a numbers cruncher runs 20 different regressions, then on average one of these 20 regressions (i.e., 5%) should be significant at the $p < 0.05$ level (Bobko 2001).

To the best of our knowledge, there is no empirical work in business studies on the cumulative incidence of false positives. However, in medical research, it appears to be very serious. In randomized trials among 150,000 men and women receiving multiple cancer screenings, the risk of a false positive test is 50% higher after the 14th test (Crosswell et al. 2009). The practical consequence of a false positive may be more severe in a medical screen than in a test of management theory, but unbiased evidence is an essential precondition for business scholars to be relevant to practicing managers and thus to make a positive real-world impact (Aguinis et al. 2010).

Paradoxically, the focus on *p*-values does not coincide with a thorough understanding of the meaning of *p*-values: *p*-values are often misinterpreted (Aguinis et al. 2010). The *p*-value generated by regression analyses is "the probability under a specified statistical model that a statistic summary of the data (e.g. the sample mean difference between two compared groups) would be equal to or more extreme than its observed value" (Wasserstein and Lazar 2016: 131). Or, in the word of classic textbook authors, the *p*-value is the probability that the sample value would be at least as large as the value observed if the null hypothesis was true (Wonnacott and Wonnacott 1990: 294). The regression result does not prove or disprove a hypothesis, it does not provide evidence regarding the reliability of the research (Branch 2014: 257), and it does not make statements about a population other than the sample.

Moreover, the *p*-value does not tell us anything about the strength of a particular association: lower *p*-values do not make relations more substantively significant, although a finding at $p < 0.01$ is often interpreted as a stronger result than one at $p < 0.05$. For example, a regression analysis of Z on X and Y may lead to a *p*-value of 0.051 for variable X and a *p*-value of 0.049 for variable Y, yet the effect size of X can be significantly larger. As noted long ago, this aspect is often overlooked (in economics, see McCloskey 1985, and McCloskey and Ziliak 1996; in psychology, see Kirk 1996). In the above example, given the selection bias, the finding on X may even never be published. However, a *p*-value of 0.05 is just a rule of thumb suggested by Fisher in 1925 in times without computers and statistical software packages, but was never meant to be interpreted as an absolute yes-or-no threshold.

In other words, *p*-values of 0.06 versus 0.04 are (almost) equally interesting. This is especially relevant for intellectually controversial and thought-provoking pieces, where we do not want a manuscript to get rejected on the basis of a *p*-value of 0.06. Moreover, statistical significance does not say anything about effect size. Although it has become more common to include effect size discussions, not all published articles discuss effect size, and many original submissions received by *JIBS* do not (yet) include an explicit discussion of the effect size. A count for all hypotheses-testing papers in the 2016 volumes of *JIBS* (54%), *Organization Science* (40%) and *SMJ* (56%) suggests substantial variation in the practice of discussing effect size for the variables of interest.

Towards Better Practice

Alternative Study Designs

Scholars may be able to enhance the rigor of their empirical evidence through their study design. First, they may conduct multiple studies to test the same hypothesis, thus providing not only evidence of validity under different conditions, but also reducing the opportunities for HARKing. In academic disciplines investigating behaviours of individuals, such as organizational psychology, organizational behaviour, and human resource management, it is established good practice to include multiple studies to test a new hypothesis (see, e.g., the *Journal of Applied Psychology).* In international business, where the validity of theory across geographic contexts is a key theme, offering evidence from two or more countries would often be a valuable contribution (Meyer 2006). However, for many of the research questions of interest for international business scholars, this is not realistically feasible, especially if the unit of analysis is firms or countries rather than individual people.

Second, experimental study designs offer interesting opportunities to advance international business knowledge that have yet to be fully exploited in the field. Specifically, experimental study designs allow varying specific variables of interests while keeping everything else constant, which is usually not feasible using field data. However, the empirical evidence of experimental studies also has been challenged due to sample selection biases (Henrich et al. 2010a, b) and endogeneity issues (Antonakis et al. 2010). Recent *JIBS* contributions by Buckley, Devinney, and Louviere (2007) and by Zellmer-Bruhn, Caligiuri, and Thomas (2016) outline opportunities to apply experimental designs in the field of international business, and offer methodological guidance (cf. van Witteloostuijn 2015).

As *JIBS* is interested in both rigor and relevance, we as editors are acutely aware that these methodological research design alternatives hold great potential, but are not always suitable to address many of the research questions of interest to the international business research community. Therefore, the challenge remains how we can improve the reliability of research findings based on testing hypotheses using regression analysis with single-sample field data.

Enhancing Reporting Practices

In a nutshell, *JIBS* expects that authors do the best feasible analysis with the available data in their line of research, do not engage in any research malpractices, report statistical results based on a full analysis of *p*-values, and provide maximum transparency to enable other scholars to build on their work (including reproduction and replication; cf. Bettis et al. 2016b; Hubbard et al. 1998). In the context of the current editorial, this translates into ten suggestions for how research and reporting practices can be enhanced.

Rigorous scholarship requires discussing the evidence for and against a hypothesis based on the full evidence, not limited to a single *p*-value of a specific test. The *American Statistical Association* (*ASA*) has recently debated this concern and issued guidelines (Wasserstein and Lazar 2016). In our view, these guidelines represent current best practice, and *JIBS* editors and reviewers can refer to these guidelines when assessing papers submitted for publication. Authors should in particular avoid over-interpreting the strength of evidence for or against a hypothesis based on levels of significance. Rather, in line with guidelines by *SMJ* (Bettis et al. 2016a, b) and others, actual *p*-values, confidence intervals, and effect sizes should be fully reported and discussed (see also Bosco et al. 2015; Hunter and Schmidt 2015). Thus, the results of hypotheses tests should normally include the following:

Guideline 1: At a basic level, all regression analyses should include, for each coefficient, standard errors (as well as mention the confidence intervals for the variable of interest) and, for each regression model, the number of observations as well as the R^2 statistics or equivalent

Guideline 2: Authors should refer to the actual *p*-value rather than the threshold *p*-value when assessing the evidence for and against their hypothesis

Guideline 3: Authors should *not* report asterisks to signal *p*-value thresholds

For guideline 1, in straightforward OLS models, the standard error of the coefficient can be calculated on the basis of the estimated coefficient and the *p*-value, but for more complicated models this is not so straightforward. We therefore expect authors to report the estimated coefficient, its standard error and *exact p*-value where relevant. Guideline 2 is in line with the call for comprehensive assessment without undue focus on the traditional threshold rules of thumb. The actual *p*-values may be included in the results table, but in many instances it may suffice to report them within the results section in the main text. In addition, Hunter and Schmidt (2015) suggest discussing not just the estimated coefficient, but also the confidence interval associated with the point estimate.

The discussion should include a reflection regarding levels of significance, given the evidence in similar studies. Sample size is critical here. For example, studies with few observations (for instance, when countries are the unit of analysis) obtain lower levels of significance than studies of independent individual decisions generated using Big Data methodologies. Guideline 3 responds to the observation that journals without "stars" have a lower probability of *p*-hacking (Brodeur et al. 2016).

Evaluating the Evidence

Good scientific practice requires that authors assess hypotheses based on a comprehensive assessment using all available evidence, rather than a singular focus on a single test statistic in a specific regression analysis. When interpreting the results, it is good practice to offer reflections and supplementary analyses that enable readers to comprehensively assess the empirical evidence. Specifically, we recommend authors to follow the following guidelines from 4 to 6 when writing their methods and results sections:

Guideline 4: Reflections on effect sizes are included, reporting and discussing whether the effects (the coefficients and, if appropriate, marginal effects) are substantive in terms of the research question at hand

Guideline 5: Outlier observations are discussed carefully, especially when they have been eliminated from the sample (e.g., through technical practices such as 'winzorizing')

Guideline 6: Null and negative findings are equally interesting as are positives, and hence are honestly reported, including a discussion of what this implies for theory

In other disciplines, such as psychology, the discussion of effect sizes has already become standard (see, e.g., Zedeck 2003: 4) and is required, for example, at the *Journal of Applied Psychology.* Ideally, effect size is a standardized, scale-free measure of the relative size of the effect. Although not without criticism, Cohen's d is an example of such a measure (Cohen 1969).

Reflections regarding effect sizes are especially important when dealing with large datasets where it is easy to obtain statistical significance even for small effects, an increasing challenge in the era of Big Data, reflecting the fact that significance is a function of sample size, as well as the alpha and effect size. As effect-size reporting is not so straightforward, we provide further suggestions and more fine-grained guidelines. The appropriate methods to generate effect sizes vary across empirical methods, and may require additional analyses. We briefly discuss several methods that are common in IB.

First, for OLS and GLS types of models, effect sizes should be calculated and reported in the usual way using the standard error of the estimated coefficient. Standardized coefficients help for interpretative reasons. Moreover, explicit comparisons can make the interpretation much more informative. For example, authors may use wording such as "Ceteris paribus, a one standard deviation increase of cultural distance (which is comparable with a change in distance from, say, US–UK to, e.g., US–Italy) reduces the longevity of joint ventures with two to four years. For comparability, the effect of a similar increase of one standard deviation of geographic distance results in a reduction of joint longevity by eight years."

Second, for logit and hazard models, we expect a discussion of effect sizes that are readily interpretable (cf. Hoetker 2007; Zelner 2009). One way is to provide odds ratios, but most readers find these hard to interpret without additional explanation. We therefore suggest authors to provide at a minimum a clear intuitive explanation of their findings. The following example can illustrate this. Assume one is interested in exploring the relation between the institutional environment in which an R&D subsidiary operates, and the probability that this subsidiary generates new product innovations (measured as a 0–1). After having established a statistically significant relation between institutional setting and subsidiary innovative performance, the effect size discussion should explore the probability that a subsidiary generates a product innovation. For example: "The probability of a subsidiary reporting an innovation is 5% higher when they are located in a country with a favorable institutional setting (e.g., country X) compared to the likelihood of a subsidiary developing product innovations in a country with a less favorable institutional regime (e.g., country Y) for which we find a probability of 1% – all else equal."

Third, for multilevel analysis (Peterson et al. 2012), the common practice is to calculate intra-class correlations, which provide scores for the explained variance at each level of analysis. In business studies, the lowest level of aggregation typically concerns a team-or subsidiary-level-dependent variable, firm and industry next and country last. Intra-class correlations often suggest that most of the variation is at lower levels. For example, intra-class correlations in cross-cultural psychology research suggest that most variation in values is supposedly at the individual level, and not at the country level (typically about 90% versus 10%) (Fisher and Schwartz 2011). However, it is critical to keep in mind here that the measurement error is also at the individual or firm level; the larger the measurement error at this level, the higher the explained variance at this level, and the lower the relative variance explained at the other levels. Correcting for measurement error in multilevel models is therefore critical (Fox and Glas 2002). Authors using multilevel methods should take this into account when reporting their findings.

Fourth, in interaction models (with 'moderating effects'), the common practice is to select one standard deviation above the mean and one standard deviation below the mean, and then draw two lines as if these coefficients reflect the full range of possible scores of a moderating variable, implicitly assuming they do not have a confidence interval – i.e., that uncertainty regarding the interaction coefficient is absent. This approach – though common – is incomplete because it ignores the margin of error with which the interaction effect is estimated, and it does not show the marginal effect for the whole range of scores on the moderating variable (the one standard deviation below and above the mean may not necessarily be representative values).

Ideally, authors should report confidence intervals for interaction effects over the relevant range of the explanatory variable. For linear models, Brambor et al. (2006) and Kingsley et al. (2017) nicely explain how to do this; for nonlinear models, we refer to Haans et al. (2016). There are various ways to provide more information on the nature and magnitude of the interaction effect. Here, using an otherwise standard STATA do-file (see Appendix 1 for details), in Fig. 4.2, we provide an example of a graph visualizing how to discuss interaction effects. This is just one (simple) example of how to unpack the nature of the interaction effect (see also Williams (2012), and Greene (2010), for alternative approaches, and Hoetker (2007), Wiersema and Bowen (2009), Zelner (2009) for a discussion of logit and probit models, also including STATA do files).

In this example with continuous variables, the two outer lines give the 95% confidence range for the interaction line, which shows the marginal effect of variable X on dependent variable Y for the full range of possible scores of the moderator variable M.[6] The small dots represent all observations for M in the sample (and not just a score one standard deviation below and above the mean of M). Only if the two lines reflecting the confidence interval are *both* below and above the horizontal zeroline, the interaction effect is significant. In Fig. 4.2, this is the case for values of M left of A and values of M right of B. This graph shows that, although the

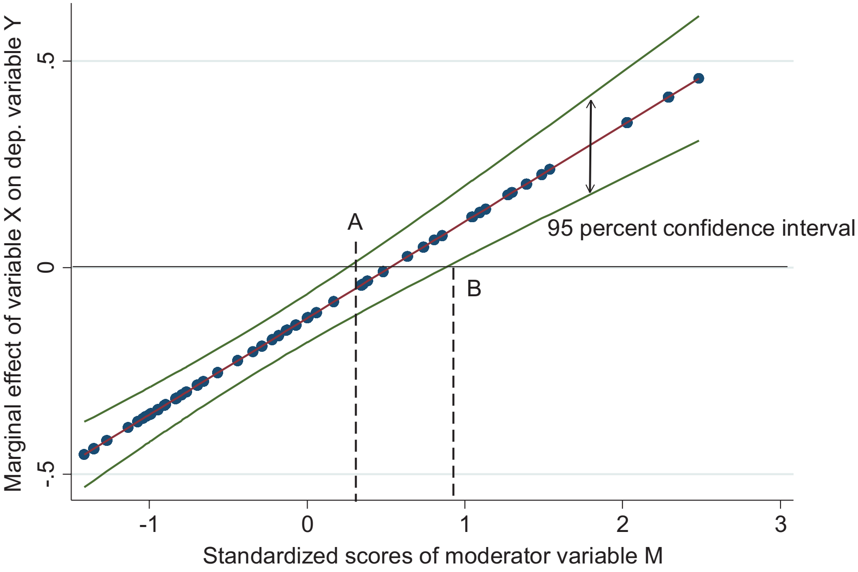

Fig. 4.2 Illustrating the effect size in interaction models

average interaction effect may turn out to be significant in a regression model, there is a range for the moderator variable M for which the effect is insignificant (between A and B), and the effect of X on Y conditional on M is negative for low values of M and positive for high values of M. Such a graph is (much) more informative than the standard practice of just showing the point estimate of X on Y for just two values of M. This observation is important, because by its very nature many international business studies are based on the starting point that key relations are moderated by contextual variation and contingent upon the external environment.

To summarize, we suggest the following more fine-grained guidelines with respect to effect-size reporting, further specifying guideline 4:

Guideline 4a: When discussing effect size, authors should take the confidence interval associated with the estimated coefficient into account as well as the minimum and maximum effect (not just one standard deviation above and below the mean), thus providing a range of the strength of a particular relationship. This may be done graphically for more complex models

Guideline 4b: When discussing effect sizes, where possible and relevant, authors should compare the range of the effect size of the variable of interest with other variables included in the regression model

The handling of outliers can significantly influence the result of regression analyses, especially if the underlying hypothesis test calls for a nonlinear relationship. Thus, as stated in guideline 5, authors should explain not only how they handled outliers, but also what the outliers tell with respect to their underlying theory. A popular but problematic approach is the use of winsorized datasets – i.e., datasets that have been transformed by eliminating extreme values (e.g., the highest and lowest 5% of the data) to reduce the effect of possibly spurious outliers. If this practice is used, non-winsorized datasets and datasets with different threshold levels for the winsorizing should be included in the robustness analysis, and discrepancies must be explained.

The results of nonsignificant or negative results can be of substantive interest for the creation of cumulative scientific knowledge, as indicated in guideline 6. This applies in particular when results fail to fully confirm received knowledge from earlier research, and when the analysis is based on high-quality data, rigorous methods and sufficient statistical power.

Causality and Endogeneity

Another critical issue involves inference of causality and endogeneity from empirical analyses. In economics, for instance, it has become common to stop using terminology related to making any causal inference unless there is a solid identification strategy in place, which is a major challenge with any cross-sectional study design. Notwithstanding that theory normally suggests a causal direction that may be illustrated by directional arrows, many commonly used empirical techniques for cross-section data do not test for the direction of causality. Solid identification strategies are often not easy to find in nonexperimental social sciences. As a result, scholars have started to adapt their language more and more, using words like "association" and "relation" instead of "determinant" and "effect" or "affect" (cf. Reeb et al. 2012).

Apart from careful language, *JIBS* expects authors to deal with the issue of endogeneity to the extent possible (Rosnow and Rosenthal 1984; Shadish et al. 2002). Probably the best-known technical solutions are lagged explanatory variables and the instrumental variables method (Angrist and Krueger 2001). Such technical solutions can often support a causal interpretation, but cannot prove causality. After all, many of these solutions are still based on correlations, while lagged variables may be subject to inertia. Moreover, the instrument may fulfil the statistical criterion of "independence," but not at a more substantive level.

The above suggests two additional guidelines:

Guideline 7: In the absence of a clear strategy designed explicitly to identify causes and effects, authors should be careful in using terminology suggesting causal relationships between variables of interest, and accordingly adjust their language in the wording of the hypotheses and in the discussion of the empirical results

Guideline 8: To the extent feasible, authors should address issues of causality and endogeneity, either by offering technical solutions or by adopting an appropriate research design

Note that guideline 8 provides direction, and should not be interpreted as "must do." Given the difficulty of finding truly exogenous instruments, we would not want otherwise excellent papers to not make it to *JIBS* because the manuscript does not have a section on endogeneity.

Robustness Tests

In view of the challenges that may arise from using a single test statistic from a single regression equation to test a hypothesis, it is important that authors assess the evidence comprehensively. In particular, by conducting a variety of robustness tests, authors can show that a significant finding is not due to an idiosyncrasy of the chosen empirical model and/or estimation strategy. Evidence from empirical tests becomes more convincing when it is supported by appropriate robustness tests.

The discussion section of a paper provides opportunities for a comprehensive assessment of the evidence, beyond the statistical properties of the specific tests used in the focal regression analysis. What tests are appropriate varies with the design of the study and the nature of the data. It is a normal part of the reviewing process that reviewers suggest some additional robustness tests, and authors are expected to seriously engage with such suggestions.[7] If this additional work were to result in an excessive number of tables, an additional file with these tables and short explanation of them can be included in a supplement to the paper that will be made available on the *JIBS* website. Robustness tests may include, for example, additional analyses with

- alternative proxies of focal constructs (i.e., variables mentioned in the hypotheses as independent or explanatory variables), especially for those that involve abstract concepts that cannot be measured directly;
- alternative sets of control variables, especially when correlation is present in the dataset between a focal explanatory variable and a control variable; and/or

- alternative functional forms of the regression models, especially for the hypotheses that suggest nonlinear effects (Haans et al. 2016; Meyer 2009), or moderating or mediating effects (Andersson et al. 2014; Cortina et al. 2015).

A guiding principle to perform certain robustness tests is the importance to rule out alternative explanations for the same finding. In the discussion section, an informative reflection on the outcomes of the robustness analyses, in relation to the hypotheses and alternative theories, can be included to clearly identify the study's findings vis-à-vis the extant literature. While such robustness tests are common practice, we suggest that more can be done to effectively use such tests. This gives the next guideline:

Guideline 9: Authors are expected to conduct a variety of robustness tests to show that the significant finding is not due to an idiosyncrasy of the selected empirical measures, model specifications and/or estimation strategy

From HARKing to Developing Theory

Hypothesizing After the Results are Known (HARKing) in search of hypotheses for already known positive results is causing great harm to scientific progress (Bosco et al. 2016). We would like to note that HARKing is not the same as "playing with your data" to explore the nature of relationships and get better feeling for possibly interesting patterns in a dataset. HARKing refers to the practice of datamining and, after significant results are established, developing or adjusting theoretical arguments *ex post*, but presenting the theory as if already in place *ex ante*. The issue with HARKing is that we have no knowledge of the many nulls and negatives that were found but not reported along the way, and therefore readers cannot be sure as to the true power of the statistical evidence. While papers in business studies journals appear to confirm groundbreaking hypotheses, we rarely see reports about falsification outcomes.[8] As indicated in our opening paragraph, about 89% of all hypotheses in *JIBS* (82%), *SMJ* (90%) and *Organization Science* (92%) were confirmed in the 2016 volumes.[9] Yet no information is provided about the many "interventions" applied to produce this abundance of positive results.

To tackle this problem, no journal can operate a policing force to monitor and sanction what is happening behind the closed doors of our authors' offices. Eliminating HARKing requires an orchestrated effort to seriously change deeply embedded practices in the scholarly community (Ioannidis 2005, 2012). What we can do, for now, is firstly to reduce the focus on single test statistics when assessing results in favour of comprehensive assessments, and thereby to reduce the incentives to engage in HARKing (hence our guidelines 1–9), and secondly to mentor and train a new generation of scholars to intrinsically dislike HARKing practices. Here, key is that established scholars lead by example. Of course, this also requires broader institutional change to remove some of the incentives that disproportionately reward scholars finding statistically significant results (Ioannidis 2012; van Witteloostuijn 2016). What journals can do boils down to, basically, two alternatives (or a combination of both).

For one, some journals are introducing the option to submit the theory first, and the empirical tests and results later (see, e.g., *Comprehensive Results in Social Psychology*, and *Management and Organization Review;* cf. Lewin et al. 2016). If the theory is accepted after a thorough review, the final manuscript will be published (of course, conditional on appropriate data and the state-of-the-art empirical analyses). This approach is nascent and it is still an open question how successful this two-step approach will be. For now, we therefore suggest that *JIBS* take the alternative route.

This second alternative is to encourage and recognize theorizing from empirical findings – i.e., the inductive leg of the development of theory. We expect papers (both submitted and published ones) to report the initial hypotheses honestly (that is, the ones drafted *before* running analyses). Developing theory *after* running analyses (to have a better explanation of the findings) is perfectly legitimate, but this could be done in a post hoc section, explicitly discussing this change of theory in relation to the results. Similarly, removing hypotheses because the evidence is weak can be problematic.

Empirical phenomena or relationships in conflict with established theories can be a powerful driver of new theoretical developments (Doh 2015). In (international) business studies, such building theory from data is more common in qualitative research, yet it is a valid methodology also with respect to quantitative data.[10] Thus, for example, a theoretical model may be motivated by connecting a surprising theoretical finding with a relevant stream of theoretical literature different than what motivated the study at the outset. Theory developed in this way *ex post* should be tested on another dataset, be it within the same paper or in a new study, similar to theory development in grounded theory research using qualitative data. The critical methodological issue here is that authors do not pretend to have a higher level of empirical support for their new theoretical ideas than what their empirical analysis provides. Thus, as editors, we encourage development or post hoc revisions of theory on the basis of empirical findings in the discussion section. This gives our final guideline:

Guideline 10: HARKing is a research malpractice. Theory developed by interpreting empirical phenomena or results should be reported as such (for example, in the discussion section)

The Role of Reviewers

In advancing international business research towards the standards reflected in the ten guidelines, we need the constructive engagement of reviewers. Firstly, this implies that we have to prevent reviewers from pushing authors towards practices that we critiqued above. This includes practices that we as editors occasionally see, such as demanding a different theoretical post hoc framing for the results already present in the paper, elimination of single hypotheses on the grounds of weak empirical support, and/or because they have been tested in prior research. Sharpening hypotheses or adding hypotheses is fine, but not around results already present in the original version. Offering post hoc alternative hypotheses to better align with findings is a natural step in the scientific research cycle, if done in the open.

Secondly, beyond avoiding negative practices, reviewers should look for positive contributions to enhance the rigor of a given study. For example, reviewers may suggest additional ways to illustrate empirical findings, or robustness tests that enhance the credibility of the results. At the same time, reviewers should avoid being perfectionists and, e.g., ask for tests that require nonexistent data, but use best practice in the given line of research as their benchmark when assessing how to evaluate the rigor of a paper under review.

Concluding Remarks

Empirical research is also the art of the feasible. In the theoretical world of an econometrics or statistics classroom, datasets have statistical properties that real-world datasets can rarely or never meet. While scholars should aspire to collect and work with high-quality datasets, as editors and reviewers we are realistic in setting our expectations. However, given these limitations, we as *JIBS* editors strongly believe that improvements are feasible, and are necessary to advance international business research to the next level and to address frequently voiced concerns regarding the validity of scholarly knowledge. This editorial has outlined what we consider good practices for conducting hypothesis testing research, and reporting and discussing the associated empirical results. We expect *JIBS's* editors, reviewers and authors to aspire to these standards. These guidelines are not written in stone, but offer benchmarks for both researchers and reviewers to enhance the quality of published international business work.

Standards are not set in stone also because they will be subject to continuous reassessment. This editorial is a clear sign of this. Debates among editors of (international) business studies journals are ongoing, and many journals are revising their editorial policies in view of these debates. This editorial has outlined concrete and actionable steps that we can take at *JIBS.* We are convinced that we, as a scholarly community, need to – and will be able to – change established research and publication practices to improve upon the current state of the art. We will all benefit from that, and will be ready to produce new and cumulative knowledge in international business that will be impactful, from both academic and societal perspectives.

Acknowledgements We gratefully acknowledge the constructive comments from Editor-in-Chief Alain Verbeke, eleven editors of *JIBS*, as well as from Bas Bosma, Lin Cui, Rian Drogendijk, Saul Estrin, Anne-Wil Harzing, Jing Li, Richard Walker, and Tom Wansbeek. We also thank Divina Alexiou, Richard Haans, Johannes Kleinhempel, Sjanne van Oeveren, Britt van Veen, and Takumin Wang for their excellent research assistance. Sjoerd Beugelsdijk thanks the Netherlands Organization for Scientific Research (NWO grant VIDI 452-011-10). All three authors contributed equally to this editorial.

Appendix 1: Stata Do File to Create Fig. 4.2

Model:

Dependent variable = Y
Independent variable = X
Moderator variable = M
Interaction variable = X∗M

To generate Fig. 4.2:

```
predictnl me = _b[X] + _b[X∗M]∗M if e(sample),
se(seme)
gen pw1 = me–1.96∗seme
gen pw2 = me + 1.96∗seme
scatter me M if e(sample) || line me pw1 pw2 M if e(sample), pstyle(p2
   p3 p3) sort legend(off) ytitle (“Marginal effect of X on Y”).
```

Notes

1. In many disciplines contributing to international business research, conventional Type 1 error probabilities are $p < 0.05$ or 0.01. There are situations where a higher Type 1 error probability, such as $p < 0.10$, might

be justified (Cascio and Zedeck 1983; Aguinis et al. 2010), for example, when the dataset is small and a larger dataset is unrealistic to obtain.

2. Note that according to Dalton et al. (2012), the selection bias (or file-drawer problem) does not appear to affect *correlation tables* in published versus unpublished papers.
3. A "true" *p*-value would be the *p*-value observed in a regression analysis that was designed based on all available theoretical knowledge (e.g., regarding the measurement of variables and the inclusion of controls), and not changed after seeing the first regression results.
4. Brodeur et al. (2016) extensively test whether this assumption holds, as well as the sensitivity of the overall distribution to issues like rounding, the number of tests performed in each article, number of tables included, and many more. Similar to Brodeur et al. (2016), we explored the sensitivity of the shape of the distribution to such issues, and we have no reason to assume that the final result in Figure 4.1 is sensitive to these issues.
5. The spikes at *z*-scores of 3, 4, and 5 are the result of rounding and are an artefact of the data. As coefficients and standard errors reported in tables are rounded – often at 2 or 3 digits – very small coefficients and standard errors automatically imply ratios of rounded numbers, and as a consequence, result in a relatively large number of *z*-scores with the integer value of 3, 4, or 5. This observation is in line with the findings reported for Economics journals by Brodeur et al. (2016).
6. The data on which the graph is based are taken from Beugelsdijk et al. (2014).
7. If authors believe that certain suggested additional tests are not reasonable or not feasible (for example, because certain data do not exist), then they should communicate that in their reply. The editor then has to evaluate the merits of the arguments of authors and reviewers, if necessary bringing in an expert on a particular methodology at hand. If the latter is required, this can be indicated in the Manuscript Central submission process.
8. A laudable exception is the recent special issue of *Strategic Management Journal* on replication (Bettis et al. 2016b).
9. The grand total is heavily influenced by *SMJ* with 362 tested hypotheses, vis-à vis 164 in *JIBS* and 185 in *Organization Science*.
10. An interesting alternative may be abduction. For example, see Dikova, Parker, and van Witteloostuijn (2017), who define abduction as "as a

form of logical inference that begins with an observation and concludes with a hypothesis that accounts for the observation, ideally seeking to find the simplest and most likely explanation." See also, e.g., Misangyi and Acharya (2014).

References

Aguinis, H., S. Werner, J.L. Abbott, C. Angert, J.H. Park, and D. Kohlhausen. 2010. Customer-centric research: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods* 13 (3): 515–539.

Andersson, U., A. Cuervo-Cazurra, and B.B. Nielsen. 2014. Explaining interaction effects within and across levels of analysis. *Journal of International Business Studies* 45 (9): 1063–1071.

Angrist, J.D., and A. Krueger. 2001. Instrumental variables and the search for identification: Form supply and demand to natural experiments. *Journal of Economic Perspectives* 15 (4): 69–85.

Angrist, J.D., and J.S. Pischke. 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24 (2): 3–30.

Antonakis, J., S. Bendahan, P. Jacquart, and R. Lalive. 2010. On making causal claims: A review and recommendations. *Leadership Quarterly* 21 (6): 1086–1120.

Barley, S.R. 2016. 60th anniversary essay: Ruminations on how we became a mystery house and how we might get out. *Administrative Science Quarterly* 61 (1): 1–8.

Bedeian, A.G., S.G. Taylor, and A. Miller. 2010. Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education* 9 (4): 715–725.

Bettis, R.A. 2012. The search for asterisks: Compromised statistical tests and flawed theory. *Strategic Management Journal* 33 (1): 108–113.

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

Bettis, R.A., C.E. Helfat, and M.J. Shaver. 2016b. Special issue: Replication in strategic management. *Strategic Management Journal* 37 (11): 2191–2388.

Beugelsdijk, S., H.L.F. de Groot, and A.B.T.M. van Schaik. 2004. Trust and economic growth: A robustness analysis. *Oxford Economic Papers* 56 (1): 118–134.

Beugelsdijk, S., A. Slangen, M. Onrust, A. van Hoorn, and R. Maseland. 2014. The impact of home-host cultural distance on foreign affiliate sales: The moderating role of cultural variation within host countries. *Journal of Business Research* 67 (8): 1638–1646.

Bhattacharjee, Y. 2013. The mind of a con man. *New York Times Magazine*, April 26.

Bobko, P. 2001. *Correlation and regression: Applications for industrial organizational psychology and management*. 2nd ed. Thousand Oaks: Sage.

Bosco, F.A., H. Aguinis, K. Singh, J.G. Field, and C.A. Pierce. 2015. Correlational effect size benchmarks. *Journal of Applied Psychology* 100 (2): 431–449.

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

Brambor, T., W.R. Clark, and M. Golder. 2006. Understanding interaction models: Improving empirical analyses. *Political Analysis* 14 (1): 63–82.

Branch, M. 2014. Malignant side-effects of null-hypothesis testing. *Theory and Psychology* 24 (2): 256–277.

Brodeur, A., M. Le, M. Sangnier, and Y. Zylberberg. 2016. Star wars: The empirics strike back. *American Economic Journal: Applied Economics* 8 (1): 1–32.

Buckley, P., T. Devinney, and J.J. Louviere. 2007. Do managers behave the way theory suggests? A choice-theoretic examination of foreign direct investment location decision-making. *Journal of International Business Studies* 38 (7): 1069–1094.

Cascio, W.F., and S. Zedeck. 1983. Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology* 36 (3): 517–526.

Choi, J., and F. Contractor. 2016. Choosing an appropriate alliance governance mode: The role of institutional, cultural and geographic distance in international research & development (R&D) collaborations. *Journal of International Business Studies* 47 (2): 210–232.

Cohen, J. 1969. *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cortina, J.M., T. Kohler, and B.B. Nielsen. 2015. Restriction of variance interaction effects and their importance for international business. *Journal of International Business Studies* 46 (8): 879–885.

Crosswell, J.M., et al. 2009. Cumulative incidence of false positive results in repeated, multimodal cancer screening. *Annals of Family Medicine* 7 (3): 212–222.

Dalton, D.R., H. Aguinis, C.A. Dalton, F.A. Bosco, and C.A. Pierce. 2012. Revisiting the file drawer problem in meta-analysis: An empirical assessment of published and non-published correlation matrices. *Personnel Psychology* 65 (2): 221–249.

Dikova, D., S.C. Parker, and A. van Witteloostuijn. 2017. Capability, environment and internationalization fit, and financial and marketing performance of MNEs' foreign subsidiaries: An abductive contingency approach. *Cross-Cultural and Strategic Management* 24 (3): 405–435.

Doh, J. 2015. Why we need phenomenon-based research in international business. *Journal of World Business* 50 (4): 609–611.

Doucouliagos, C., and T.D. Stanley. 2013. Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys* 27 (2): 316–339.

Economist. 2014. *When science gets it wrong: Let the light shine in*. June 14. http://www.economist.com/news/science–and–technology/21604089-two-big-recent-scientific-results-are-looking-shakyand-it-open-peer-review. Accessed 23 Mar 2017.

Ferguson, C.J., and M. Heene. 2012. A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science* 7 (6): 555–561.

Fisher, R.A. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R., and S. Schwartz. 2011. Whence differences in value priorities? Individual, cultural, and artefactual sources. *Journal of Cross-Cultural Psychology* 42 (7): 1127–1144.

Fox, P.J., and C.A.W. Glas. 2002. Modeling measurement error in a structural multilevel model. In *Latent variable and latent structure models*, ed. G.A. Marcoulides and I. Moustaki. London: Lawrence Erlbaum Associates.

Gerber, A.S., D.P. Green, and D. Nickerson. 2001. Testing for publication bias in political science. *Political Analysis* 9 (4): 385–392.

Gigerenzer, G. 2004. Mindless statistics. *Journal of Socio-Economics* 33 (5): 587–606.

Goldfarb, B., and A. King. 2016. Scientific Apophenia in strategic management research: Significance tests & mistaken inference. *Strategic Management Journal* 37 (1): 167–176.

Gorg, H., and E. Strobl. 2001. Multinational companies and productivity spillovers: A meta-analysis with a test for publication bias. *Economic Journal* 111: F723–F739.

Greene, W. 2010. Testing hypotheses about interaction terms in nonlinear models. *Economics Letters* 107: 291–296.

Grieneisen, M.L., and M. Zhang. 2012. A comprehensive survey of retracted articles from the scholarly literature. *PLoS One* 7 (10): e44118. https://doi.org/10.1371/journal.pone.0044118.

Haans, R.F.P., C. Pieters, and Z.L. He. 2016. Thinking about U: Theorizing and testing U-and inverted U-shaped relationships in strategy research. *Strategic Management Journal* 37 (7): 1177–1196.

Head, M.L., L. Holman, R. Lanfear, A.T. Kahn, and M.D. Jennions. 2015. The extent and consequences of p–hacking in science. *PLoS Biology* 13 (3): e1002106. https://doi.org/10.1371/journal.pbio.1002106.

Henrich, J., S.J. Heine, and A. Norenzayan. 2010a. The weirdest people in the world? *Behavioral and Brain Sciences* 33 (2–3): 61–83.

———. 2010b. Most people are not WEIRD. *Nature* 466: 29.

Hoetker, G. 2007. The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal* 28 (4): 331–343.

Hubbard, R., D.E. Vetter, and E.L. Little. 1998. Replication in strategic management: Scientific testing for validity, generalizability, and usefulness. *Strategic Management Journal* 19 (3): 243–254.

Hunter, J.E., and F.L. Schmidt. 2015. *Methods of meta-analysis: Correcting error and bias in research findings*. 2nd ed. Thousand Oaks: Sage.

Husted, B.W., I. Montiel, and P. Christmann. 2016. Effects of local legitimacy on certification decision to global and national CSR standards by multinational subsidiaries and domestic firms. *Journal of International Business Studies* 47 (3): 382–397.

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

———. 2012. Why science is not necessarily self-correcting. *Perspectives on Psychological Science* 7 (6): 645–654.

John, L.K., G. Loewenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science* 23 (5): 524–532.

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

Kingsley, A.F., T.G. Noordewier, and R.G. Vanden Bergh. 2017. Overstating and understating interaction results in international business research. *Journal of World Business* 52 (2): 286–295.

Kirk, R.E. 1996. Practical significance: A concept whose time has come. *Educational and Psychological Measurement* 56 (5): 746–759.

Leamer, E.E. 1985. Sensitivity analyses would help. *American Economic Review* 75 (3): 308–313.

Lewin, A.Y., C.Y. Chiu, C.F. Fey, S.S. Levine, G. McDermott, J.P. Murmann, and E. Tsang. 2016. The critique of empirical social science: New policies at *Management and Organization Review*. *Management and Organization Review* 12 (4): 649–658.

Lexchin, J., L.A. Bero, B. Djulbegovic, and O. Clark. 2003. Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *British Medical Journal* 326 (7400): 1167–1170.

Masicampo, E.J., and D.R. Lalande. 2012. A peculiar prevalence of *p*-values just below 0.05. *Quarterly Journal of Experimental Psychology* 65 (11): 2271–2279.

McCloskey, D.N. 1985. The loss function has been mislaid: The rhetoric of significance tests. *American Economic Review* 75 (2): 201–205.

McCloskey, D.N., and S.T. Ziliak. 1996. The standard error of regressions. *Journal of Economic Literature* 34: 97–114.

Meyer, K.E. 2006. Asian management research needs more self-confidence. *Asia Pacific Journal of Management* 23 (2): 119–137.

———. 2009. Motivating, testing, and publishing curvilinear effects in management research. *Asia Pacific Journal of Management* 26 (2): 187–193.

Misangyi, V.F., and A.G. Acharya. 2014. Substitutes or complements? A configurational examination of corporate governance mechanisms. *Academy of Management Journal* 57 (6): 1681–1705.

Mullane, K., and M. Williams. 2013. Bias in research: the rule rather than the exception? *Elsevier Journal.* http://editorsupdate.elsevier.com/issue-40-september-2013/bias-in-research-the-rule-rather-than-the-exception. Accessed 23 Mar 2017.

New York Times. 2011. *Fraud case seen as a red flag for psychology research.* November 2. http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html?-r=1&ref=research. Accessed 15 Jan 2017.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*. https://doi.org/10.1126/science.aac4716.

Orlitzky, M. 2012. How can significance tests be deinstitutionalized? *Organizational Research Methods* 15 (2): 199–228.

Pashler, H., and E.-J. Wagenmakers. 2012. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* 7 (6): 528–530.

Peterson, M., J.L. Arregle, and X. Martin. 2012. Multi-level models in international business research. *Journal of International Business Studies* 43 (5): 451–457.

Pfeffer, J. 2007. A modest proposal: How we might change the process and product of managerial research. *Academy of Management Journal* 50 (6): 1334–1345.

Popper, K. 1959. *The logic of scientific discovery*. London: Hutchinson.

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

Rosenthal, R. 1979. The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86 (3): 638–641.

Rosnow, R.L., and R. Rosenthal. 1984. *Understanding behavioral science: Research methods for customers*. New York: McGraw-Hill.

Rothstein, H.R., A.J. Sutton, and M. Borenstein. 2005. *Publication bias in meta-analysis, prevention, assessment and adjustment*. New York: Wiley.

Sala-i-Martin, X. 1997. I just ran two million regressions. *American Economic Review* 87 (2): 178–183.

Shadish, W.R., T.D. Cook, and D. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Simmons, J.P., L.D. Nelson, and U. Simonsohn. 2011. Falsepositive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22 (11): 1359–1366.

Sterling, T.D. 1959. Publication decision and their possible effects on inferences drawn from tests of significance – Vice versa. *Journal of the American Statistical Association* 54 (285): 30–34.

van Witteloostuijn, A. 2015. Toward experimental international business: Unraveling fundamental causal linkages. *Cross Cultural & Strategic Management* 22 (4): 530–544.

———. 2016. What happened to Popperian falsification? Publishing neutral and negative findings. *Cross Cultural & Strategic Management* 23 (3): 481–508.

Wasserstein, R. L., and N. A. Lazar. 2016. The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician,* 70(2): 129–133. http://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108. (ASA = American Statistical Association).

Wiersema, M.F., and H.P. Bowen. 2009. The use of limited dependent variable techniques in strategy research: Issues and methods. *Strategic Management Journal* 30 (6): 679–692.

Williams, R. 2012. Using the margins command to estimate and interpret adjusted predictions and marginal effects. *Stata Journal* 12 (2): 308.

Wonnacott, T.H., and R.J. Wonnacott. 1990. *Introductory statistics for business and economics*. New York: Wiley.

Zedeck, S. 2003. Editorial. *Journal of Applied Psychology* 88 (1): 3–5.

Zellmer-Bruhn, M., P. Caligiuri, and D. Thomas. 2016. From the editors: Experimental designs in international business research. *Journal of International Business Studies* 47 (4): 399–407.

Zelner, B. 2009. Using simulation to interpret results from logit, probit, and other nonlinear models. *Strategic Management Journal* 30 (12): 1335–1348.