Journal of International Business Studies (2017) 48, 653–663
© 2017 Academy of International Business All rights reserved 0047-2506/17
www.jibs.net

EDITORIAL

Science's reproducibility and replicability crisis: International business is not immune

Herman Aguinis[1], Wayne F. Cascio[2] and Ravi S. Ramani[1]

[1] *Department of Management, School of Business, George Washington University, 2201 G St. NW, Washington, DC 20052, USA;* [2] *The Business School, University of Colorado Denver, 1475 Lawrence Street, Denver, CO 80202, USA*

Correspondence:
H Aguinis, Department of Management, School of Business, George Washington University, 2201 G St. NW, Washington, DC 20052, USA.
Tel: (+1) 202-994-6976;
e-mail: haguinis@gwu.edu

Abstract
International business is not immune to science's reproducibility and replicability crisis. We argue that this crisis is not entirely surprising given the methodological practices that enhance systematic capitalization on chance. This occurs when researchers search for a maximally predictive statistical model based on a particular dataset and engage in several trial-and-error steps that are rarely disclosed in published articles. We describe systematic capitalization on chance, distinguish it from unsystematic capitalization on chance, address five common practices that capitalize on chance, and offer actionable strategies to minimize the capitalization on chance and improve the reproducibility and replicability of future IB research.

Journal of International Business Studies (2017) **48**, 653–663.
doi:10.1057/s41267-017-0081-0

Keywords: capitalization on chance; quantitative methods; reproducibility; replicability; credibility

Received: 17 March 2017
Accepted: 5 April 2017
Online publication date: 12 May 2017

INTRODUCTION

International business (IB) and many other management and organization studies' disciplines are currently immersed in an important debate regarding the credibility and usefulness of the scholarly knowledge that is produced (Cuervo-Cazurra, Andersson, Brannen, Nielsen, & Reuber, 2016; Davis, 2015; George, 2014; Meyer, van Witteloostuijn, & Beugelsdijk, 2017). A critical issue in this debate is the lack of ability to reproduce and replicate results described in published articles (Bakker, van Dijk, & Wicherts, 2012; Bergh, Sharp, & Li, 2017; Cuervo-Cazurra et al., 2016; Ioannidis, 2005; Open Science Collaboration, 2015). *Reproducibility* means that someone other than a published study's authors is able to obtain the same results using the authors' own data, whereas *replicability* means that someone other than a published study's authors is able to obtain substantially similar results by applying the same steps in a different context and with different data. Clearly, it is difficult to make the case that research results are credible and useful if they are irreproducible and not replicable. Unfortunately, there is a proliferation of evidence indicating that lack of reproducibility and replicability are quite pervasive (e.g., Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016; Cortina, Green, Keeler, & Vandenberg, 2016; Cuervo-Cazurra et al., 2016; Ioannidis, 2005; Open Science Collaboration, 2015; Schwab & Starbuck,

2017). Accordingly, as noted by Verbeke, Von Glinow, and Luo, "… the IB discipline faces the challenges of remaining at par with the methodological standards in adjacent fields for validity, reliability, replicability and generalizability" (Verbeke et al., 2017: 6). In short, IB is not immune to science's reproducibility and replicability crisis.

We argue that concerns about lack of reproducibility and replicability are actually not entirely surprising because of current methodological practices that enhance *systematic capitalization on chance*. Systematic capitalization on chance occurs when a researcher searches for a maximally predictive statistical model based on a particular dataset, and it typically involves several trial-and-error steps that are rarely disclosed in published articles. Currently, there is tremendous pressure to publish in the so-called top journals because the number of such publications has an important impact on faculty performance evaluations and rewards, including promotion and tenure decisions (Aguinis, Shapiro, Antonacopoulou, & Cummings, 2014; Butler, Delaney, & Spoelstra, 2017; Nosek, Spies, & Motyl, 2012). Thus researchers are strongly motivated to produce manuscripts that are more likely to be accepted for publication. This means submitting manuscripts that report tests of hypotheses that are statistically significant and "more highly" significant, models that fit the data as well as possible, and effect sizes that are as large as possible (Meyer et al., 2017). To paraphrase Friedman and Sunder (1994: 85), many researchers "torture the data until they confess" that effects are statistically significant, large, and supportive of favored hypotheses and models. Each of these outcomes – which together are more likely to produce the desired result of a successful publication – can be reached more easily by systematically capitalizing on chance.

Researchers today have more "degrees of freedom" regarding methodological choices than ever (Freese, 2007). Many of these degrees of freedom involve practices that enhance capitalization on chance and improve the probability of successful publication. For example, researchers may include or delete outliers from a manuscript depending on which course of action results in a larger effect-size estimate (Aguinis, Gottfredson, & Joo, 2013). As a second illustration, researchers may capitalize on chance by selecting a particular configuration of control variables after analyzing the impact of several groups of control variables on results and selecting the final set based on which configuration results in better fit indices for a favored model (Bernerth & Aguinis, 2016).

We emphasize that our focus on systematic capitalization on chance is different from *unsystematic capitalization of chance*, which is due to random fluctuations in any given sample drawn from a population. Unsystematic capitalization on chance is a known phenomenon and part of all inferential statistical tests. Specifically, the goal of inferential statistics is to maximize the predictive power of a model based on the data available by minimizing errors in prediction using sample scores (Cascio & Aguinis, 2005). For example, ordinary least squares (OLS) regression, which is one of the most frequently used data-analytic approaches in IB and other fields (e.g., Aguinis, Pierce, Bosco, & Muslin, 2009; Boellis, Mariotti, Minichilli, & Piscitello, 2016; Fitzsimmons, Liao, & Thomas, 2017; Fung, Zhou, & Zhu, 2016), minimizes the sum of the squared differences between fitted values and observed values. Unsystematic capitalization on chance is addressed by conducting inferential tests of the parameter estimates that include their standard errors, thereby providing information about the precision in the estimation process (i.e., larger sample sizes are associated with greater precision and smaller standard errors). Most articles in IB research include information on sample size and standard errors, which allows consumers of research to independently evaluate the accuracy of the estimation process and the meaning of results for theory and practice, thereby accounting for unsystematic capitalization on chance.[1]

Next, we describe several common practices that enhance systematic capitalization on chance and illustrate these practices using articles published in *Journal of International Business Studies* (JIBS). Because each of the issues we discuss is so pervasive, we do not "name names." We do not believe it would be helpful or constructive to point fingers at particular authors. However, we mention variable names and the overall substantive context of each study so that the methodological issues we discuss are directly and specifically relevant for an IB readership. Then, we offer best-practice recommendations on how to minimize capitalization on chance in future IB research. Similar to previously published JIBS guest editorials (e.g., Andersson, Cuervo-Cazurra, & Nielsen, 2014; Chang, van Witteloostuijn, & Eden, 2010; Meyer et al., 2017; Reeb, Sakakibara, & Mahmood, 2012), these recommendations serve as resources for researchers, including doctoral students and their training, as well as for

journal editors and reviewers evaluating manuscript submissions.

COMMON METHODOLOGICAL PRACTICES THAT ENHANCE SYSTEMATIC CAPITALIZATION ON CHANCE

In this section, we discuss five common methodological practices that enhance systematic capitalization on chance: (1) selection of variables to include in a model, (2) use of control variables, (3) handling of outliers, (4) reporting of *p* values, and (5) hypothesizing after results are known (HARKing). We describe each of these issues and elaborate on how they lead to lack of reproducibility and replicability.

Selection of Variables to Include in a Model

The selection of variables to include in a model encompasses both the choice of variables to include, as well as the specification of the nature of the relations among these variables. Rapid advances in computational methodologies have allowed researchers to analyze increasingly larger amounts of data without much additional effort or cost (Simmons, Nelson, & Simonsohn, 2011). Within the field of IB in particular, researchers routinely deal with "Big Data," that is, large amounts of information stored in archival datasets (Harlow & Oswald, 2016). Because these datasets were not collected directly in response to a particular research question, they contain many variables that can be restructured to produce "favorable" results (i.e., better fit estimates, larger effect-size estimates) (Chen & Wojcik, 2016). For example, consider the case of firm performance, which is one of the most frequently measured constructs in IB. As Richard, Devinney, Yip, and Johnson (2009) noted, firm performance can be defined and assessed in terms of objective measures (e.g., shareholder returns, Tobin's *q*), and subjective measures (e.g., reputation, comparative ranking of firms). The choice of which firm-performance measure is examined should be driven by theory, and there should be a clear justification for why a particular measure was used, given the aims of the study (Richard et al., 2009).

Three recent articles published in JIBS have used the following measures of firm performance: (Study 1) increased reputation, overall performance, increased number of new products and customers, and enhanced product quality; (Study 2) return on assets; and (Study 3) return on equity, market-to-book ratio of assets, sales efficiency, and corporate risk-taking. Of these three studies, two did not provide any explanation or rationale for why they used those specific measures of firm performance, and the third cited "prior research" without providing any references or arguments in support of this particular choice. A healthily skeptical readership cannot judge if the firm-performance measures used in these studies were chosen because they aligned with the theories the researchers were testing, or because these measures produced outcomes that supported the favored hypotheses. Moreover, it is not possible to ascertain if, initially, several measures of firm performance were considered, but only those that produced the most favorable results were retained in the published article.

Systematic capitalization on chance in terms of which variables are included in a predictive model, and how this final set of variables is chosen, has a direct detrimental impact of future efforts to reproduce and replicate substantive results. Almost 25 years ago, MacCallum, Roznowski, and Necowitz (1992) reported that researchers were making post-hoc modifications to improve the fit of models by utilizing results provided by the data-analytical software. MacCallum et al. (1992: 491) noted that this process of re-specifying models based on the data was "inherently susceptible to capitalization on chance" because the modifications were driven not by substantive reasons, but by the peculiarities of the dataset itself. Despite calls for a more thoughtful approach to the use and reporting of these modifications (e.g., Bentler, 2007; Hurley et al., 1997), recent reviews show that they are still widely used, but rarely reported (Banks, O'Boyle et al., 2016; Cortina et al., 2016; Sijtsma, 2016). For example, a recently published article in JIBS reported "relaxing" 35 of 486 constraints, including those associated with measurement error terms, until the model reached an acceptable fit. The article does not include any information on which specific paths were changed or any theory or measurement rationale for each of these "improvements" other than the goal of achieving a superior model fit. Given the popularity of data-analytical approaches such as structural equation modeling in research reported in JIBS (e.g., Funk, Arthurs, Treviño, & Joireman, 2010; Lisak, Erez, Sui, & Lee, 2016), we suspect that there are many other instances where researchers systematically capitalize on chance by making such modifications until an optimally fitting model is found – without

necessarily reporting which paths were added or deleted from the original model, and why.

Use of Control Variables

Statistical controls are variables considered to be extraneous (i.e., non-central) to the hypotheses being tested but that could provide alternative explanations for results. Control variables are used very frequently in management and organization studies (Becker, 2005; Carlson & Wu, 2012; Spector & Brannick, 2011). For example, control variables are used by entering them in a hierarchical manner when conducting multiple regression analyses, under the presumption that they eliminate contamination between the predictor and outcome variables (Bernerth & Aguinis, 2016). However, the assumptions and theoretical rationale underlying the use of control variables, namely, that including them provides a "truer" test of relations and that the controls used are measured reliably, are seldom tested (Bernerth & Aguinis, 2016). Researchers rarely make explicit the reasons why certain variables (and not others) were chosen as controls (Becker, 2005; Spector & Brannick, 2011). Finally, control variables reduce the statistical power of the test and the variance associated with the criterion that can potentially be explained by substantive variables (Breaugh, 2008), thereby increasing the chance that the results obtained are an artifact of the choice of control variables used (Bernerth & Aguinis, 2016). Control variables therefore increase systematic capitalization on chance as researchers test several models, including and excluding controls piecemeal until they obtain a desired result (Banks, O'Boyle et al., 2016; Bernerth & Aguinis, 2016; Simmons et al., 2011). As noted by Cuervo-Cazurra et al. (2016: 894), "without specific knowledge about which controls were included, how they were measured and where they come from, replication is impossible".

Systematic capitalization on chance regarding the use of control variables seems pervasive in IB research. For example, four recent studies published in JIBS included the following sets of control variables: (Study 1) retained earnings scaled by the book value of assets, the ratio of shareholders' equity to the book value of assets, the natural logarithm of the ratio of current year sales revenue to prior year sales, and an indicator variable denoting the incidence of share repurchases; (Study 2) the natural log of a firm's book value of tangible assets per employee and the log number of employees; (Study 3) gender, age, job rank, exposure to female managers, and organizational sector; and (Study 4) organizational tenure, tenure with supervisor, group size, and country affiliation. Of these four studies, two did not provide any explanation or rationale for the authors' choice to include those specific control variables or information on any control variables that were initially included but later excluded. The authors of the other two studies justified their choices by citing "past research" examining the impact of the same control variables. But, readers have no way of knowing whether the control variables had a conceptual justification, or whether they were added in a post-hoc manner after much trial and error involving several potential controls, and the final set was chosen because it improved model fit or provided better results in support of the favored hypotheses.

Handling of Outliers

Outliers are "data points that deviate markedly from others" (Aguinis et al., 2013: 270), and are commonly found in management and organization studies (Hunter & Schmidt, 2015; Rousseeuw & Leroy, 2003). Outliers are a challenge because they can substantially affect results obtained when testing hypotheses (Bobko, 2001; Orr, Sackett, & DuBois, 1991). Because of their outsized influence, the management of outliers presents an opportunity for researchers to systematically capitalize on chance when analyzing data, often in the direction of supporting their hypotheses (Cortina, 2002). However, many researchers routinely fail to disclose whether they tested for outliers within their datasets, whether any outliers were identified, the type of outliers found, and the rationale behind choosing to include or exclude outliers from analyses (Aguinis et al., 2013).

Recently published articles in JIBS suggest the presence of systematic capitalization on chance regarding the management of outliers. For example, reported practices include winsorizing firm-level variables at the 5% level to account for outliers, trimming the sample by excluding observations at the top and bottom one percentile of variables, and removing an outlier based on studentized residuals and Cook's D.[2] In none of these cases did the authors define the type of outlier they were addressing. Specifically, error outliers (i.e., data points that lie at a distance from other data points), interesting outliers (i.e., non-error data points that lie at a distance from other data points and may contain valuable or unexpected knowledge), or influential outliers (i.e., non-error data

points that lie at a distance from other data points, are not error or interesting outliers, and also affect substantive conclusions). In addition, in none of these published articles did the authors take appropriate steps such as correcting the data for error outliers and reporting the results with and without outliers (Aguinis et al., 2013). Therefore by not providing clear and detailed reporting of the manner in which they addressed the issue of outliers, it is virtually impossible to reproduce and replicate substantive conclusions.

Reporting of *p* values

Another issue that involves systematic capitalization on chance refers to the reporting of p values associated with tests of significance. Despite its many flaws, null hypothesis significance testing (NHST) continues to be the choice of researchers in management and organization studies (Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016; Meyer et al., 2017). In NHST, the tenability of a null hypothesis (i.e., no effect or relation) is primarily judged based on the observed p value associated with the test of the hypothesis, and values smaller than 0.05 are often judged as providing sufficient evidence to reject it (Bettis et al., 2016; Goldfarb & King, 2016). Of the many problems associated with this interpretation of p values, the most pernicious is that it motivates researchers to engage in a practice called "p-hacking" and to report "crippled" p values (see below) (Aguinis, Werner, Abbott, Angert, Park, & Kohlhausen, 2010; Banks, Rogelberg et al., 2016). For example, consider a researcher who interprets $p = 0.0499$ as sufficient evidence for rejecting the null hypothesis, and $p = 0.0510$ as evidence that the null hypothesis should be retained, and believes that journals are more likely to look favorably on rejected null hypotheses. This researcher will be highly motivated to "p-hack," that is, find some way, such as using control variables or eliminating outliers, to reduce the p value below the 0.05 threshold (Aguinis et al., 2010, Goldfarb & King, 2016; Starbuck, 2016; Waldman & Lilienfeld, 2016). Similarly, this researcher will be motivated to report p values using cutoffs (e.g., $p < 0.05$), rather that report the actual p value (0.0510). Using this cutoff not only "cripples" the amount of information conveyed by the statistic (Aguinis, Pierce, & Culpepper, 2009), but also allows the researcher to claim that his or her hypothesis was supported (Aguinis et al., 2010).

Many of the aforementioned practices regarding the reporting of p values are commonly found in articles published in JIBS. For example, recent studies in JIBS reported p values by using cutoffs instead of reporting actual p values, using multiple p value cutoffs within the same article, and using the term "marginally significant" to indicate $p < 0.10$. In classical hypothesis testing, 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 & Zedeck, 1983), but it is the responsibility of the researcher to provide such justification explicitly (Aguinis et al., 2010). In classical hypothesis testing, results either are or are not significant; there is no such thing as "marginally significant" results. The examples regarding the use of control variables and outliers provided above, along with evidence from other fields, such as strategic management (Bettis et al., 2016; Goldfarb & King, 2016) and psychology (Bakker & Wicherts, 2011; Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2015) suggest the existence of published articles in which researchers exercised their "degrees of freedom" to systematically manipulate the data to obtain a significant (i.e., $p < 0.05$) result. Engaging in these practices increases systematic capitalization on chance and diminishes the probability that results will be reproducible and replicable.

Hypothesizing After Results are Known (HARKing)

Hypothesizing after results are known (HARKing) occurs when researchers retroactively include or exclude hypotheses from their study after analyzing the data, that is, post-hoc hypotheses presented as a-priori hypotheses, *without acknowledging having done so* (Kerr, 1998). A key issue regarding HARKing is lack of transparency. Specifically, epistemological approaches other than the pervasive positivistic model, which has become dominant in management and related fields since before World War II (Cortina, Aguinis, & DeShon, 2017), are indeed useful and even necessary. For example, inductive and abductive approaches can lead to important theory advancements and discoveries (Bamberger & Ang, 2016; Fisher & Aguinis, 2017; Hollenbeck & Wright, 2016). So, we are not advocating a rigid adherence to a positivistic approach, but rather, methodological plurality that is fully transparent so that results can be reproduced and replicated.

While primary-level and meta-analysis estimates based on self-reports indicate that 30–40% of researchers engage in HARKing, the number is likely higher because only a minority of researchers are likely to admit openly that they engaged in this

practice (Banks, O'Boyle et al., 2016; Bedeian, Taylor, & Miller, 2010; Fanelli, 2009). Consider the study by John, Loewenstein, and Prelec (2012), who surveyed 2,155 academic psychologists regarding nine questionable research practices, including "reporting an unexpected finding as having been predicted from the start." John et al. (2012) asked these researchers (a) whether they had engaged in those practices (self-admission rate), (b) the percentage of other psychologists who had engaged in those practices (prevalence estimate), and (c) among those psychologists who had, the percentage that would admit to having done so (admission estimate). For this particular question addressing HARKing, the self-admission rate was about 30%, but the prevalence rate was about 50%, and the admission estimate was about 90%. More recently, O'Boyle, Banks, and Gonzalez-Mulé (2017) examined doctoral dissertations and the subsequent academic journal articles that they spawned. Their results revealed that the ratio of supported versus non-supported hypotheses was roughly 2 to 1. That is, somewhere between dissertation defense and published journal article, authors chose, altered, or introduced hypotheses after examining their data, likely to enhance the probability of publication (Bedeian et al., 2010; Edwards & Berry, 2010; Starbuck, 2016).

Even more worrisome is that many instances of HARKing are driven and even encouraged by reviewers and editors as part of the peer-review process (Banks, O'Boyle et al., 2016; Bedeian et al., 2010). In fact, Bosco, Aguinis, Field, Pierce, and Dalton (2016) conducted a survey of authors who had published in *Journal of Applied Psychology* and *Personnel Psychology* and found that 21% reported that at least one hypothesis change had occurred as a result of the review process. Because HARKing involves researchers fabricating or altering hypotheses based on the specific peculiarities of their datasets and not openly and honestly reporting so, it represents a particularly blatant instance of systematic capitalization on chance.

To illustrate the aforementioned discussion, we reviewed all articles published in JIBS in 2016 that proposed and quantitatively tested hypotheses. Let us be clear: our intentions are not to disparage any of the researchers or studies we examined, but simply to highlight trends. Across 30 studies published in JIBS in 2016 that met our criteria, researchers proposed 137 hypotheses, of which 115 (84%) received complete or partial support, and only 22 (16%) were not supported. Based on these results, it seems that researchers are almost five-times more likely to find support for their favored hypotheses than they are to reject them. While not definitive, these results, combined with known self-reports of researchers admitting to HARKing, are a "smoking gun" (Bosco et al., 2016) that hints at the existence of HARKing in IB research.

STRATEGIES TO MINIMIZE CAPITALIZATION ON CHANCE

Meta-analysis seems to be a possible solution to understand whether a particular body of work has been subjected to capitalization on chance because it allows researchers to account for variables that create fluctuations in the observed estimates of effect sizes (Hunter & Schmidt, 2015). Because meta-analysis can correct for the effects of methodological and statistical artifacts, such as sampling error and measurement error, it has become a popular methodological approach in IB research (e.g., Fischer & Mansell, 2009; Stahl, Maznevski, Voigt, & Jonsen, 2010; van Essen, Heugens, Otten, & van Oosterhout, 2012). However, meta-analysis only corrects for unsystematic capitalization on chance and not for systematic capitalization on chance. As noted by Eysenck almost 40 years ago: "garbage in, garbage out is a well-known axiom of computer specialists; it applies here [for meta-analysis] with equal force" (Eysenck, 1978: 517). In other words, if effect-size estimates in primary-level studies are upwardly biased due to systematic capitalization on chance, accumulating all of those estimates will lead to a meta-analytic summary effect that will be similarly biased. Thus even if the estimated parameters are used to create distributions (i.e., funnel plots) (Dalton, Aguinis, Dalton, Bosco, & Pierce, 2012; Macaskill, Walter, & Irwig, 2001), systematic capitalization on chance biases the entire distribution. In short, meta-analysis is not a solution to address systematic capitalization on chance and its biasing effects on results and substantive conclusions.

Cross-validation is another strategy that could potentially be used to minimize the effects of capitalization on chance, but it also addresses its unsystematic and not its systematic variety. Specifically, ρ_c, an estimate of cross-validity in the population, refers to whether parameter estimates (usually regression coefficients) derived from one sample can predict outcomes to the same degree in the population as a whole or in other samples drawn from the same population. If cross-validity is

low, the use of assessment tools and prediction systems derived from one sample may not be appropriate in other samples from the same population. Cascio and Aguinis (2005) provided a detailed discussion of various approaches to cross-validation and recommended estimating the cross-validity in the population (i.e., ρ_c) by adjusting the sample-based multiple correlation coefficient (R) by a function of sample size (N) and the number of predictors (k). It is important to note that what most computer outputs label "adjusted R^2" is only an intermediate step in computing the cross-validity in the population. Adjusted R^2 does not address the issue of prediction optimization based on the capitalization on chance factors in the original sample and, therefore, underestimates the shrinkage (i.e., amount by which observed values were overestimated). Based on a careful review of the relevant literature, Cascio and Aguinis (2005) suggested appropriate formulas for estimating cross-validity in the population. Next, we offer suggestions for how to minimize systematic capitalization on chance specifically regarding each of the first five issues we mentioned earlier.

Issue #1 is the selection of variables in models. To improve reproducibility and replicability, researchers must clearly report the rationale behind the decision rules used in determining the sample-size and data-collection procedures, and report all the variables that they have considered (Simmons et al., 2011). If a construct can be assessed using several measures available (e.g., firm performance), researchers should justify their choice in light of theoretical considerations and the aims of their study (Richard et al., 2009). When making modifications to models, researchers should consider sample size, as modifications made to models drawing on small samples are likely to yield larger and more idiosyncratic results (MacCallum et al., 1992). Because each modification made to a model increases the fit of the model to the data in hand and decreases replicability (MacCallum et al., 1992), researchers should explicitly report all modifications made to their models, the theoretical rationale for the modifications, and the fit statistics for each model tested (Credé & Harms, 2015; MacCallum et al., 1992).

Issues #2 and #3 relate to the use of control variables and the handling of outliers. Choosing which variables to use as controls or which data points to include or exclude from the analyses offers researchers an opportunity to systematically capitalize on chance. To minimize this, researchers should provide a theoretical justification for the choice of each control variable, along with evidence of prior empirical work showing a relationship between the proposed control and the focal variable. They should explain why the control variable is integral to the model they propose to test, and offer evidence regarding the reliability of the control variable (Bernerth & Aguinis, 2016). When reporting results, researchers should provide descriptive statistics for all control variables, as well as reporting results with and without control variables (Aguinis & Vandenberg, 2014; Becker, 2005; Bernerth & Aguinis, 2016). Regarding outliers, researchers should provide evidence showing that they tested for outliers in their datasets. They should specify the rules used to identify and classify outliers as error, interesting, or influential, and disclose whether influential outliers affect model fit or prediction. Finally, they should test their models using robust approaches (e.g., absolute deviation) and report results with and without outliers (Aguinis et al., 2013).

Issue #4 is the reporting of p values. Relying on arbitrary p values (such as 0.05) to guide decisions motivates researchers to engage in "p-hacking," report "crippled" results, and conflate statistical and practical significance (Aguinis et al., 2010). To counter these deleterious effects, researchers should formally state the α level used to evaluate their hypotheses given the relative seriousness of making a Type I (probability of wrongly rejecting the null hypothesis) versus Type II (probability of mistakenly retaining the null hypothesis) error; justify the use of multiple cutoffs within the same paper; report complete p values to the second decimal place; not use terms such as "marginally significant" or "very significant" when referring to p values; and discuss the practical significance of their results in terms of the context of their study (Aguinis et al., 2010).

Lastly, we examined how researchers might systematically capitalize on chance through HARKing by creating and reporting hypotheses after analyzing their data, either of their own volition, or as directed to by reviewers, and not describing hypotheses as being post hoc in an open and honest manner. To counter this practice, researchers should conduct more studies using strong inference testing and report results of post-hoc hypotheses in a separate section from a-priori hypotheses (Banks, O'Boyle et al., 2016; Bosco et al., 2016; Hollenbeck & Wright, 2016). In addition, influential and highly visible journals like JIBS can play a prominent role in countering this practice by encouraging more

replication studies, promoting inductive and abductive research (Fisher & Aguinis, 2017), and using study registries where authors post the details of their proposed research before collecting and analyzing the data (Aguinis & Vandenberg, 2014; Bosco et al., 2016; Kepes & McDaniel, 2013).

CONCLUSIONS

A manuscript is more likely to be accepted for publication if results are statistically significant, effect-size estimates are large, and hypotheses and models are supported. So, consciously or not, it is in the best interests of researchers to achieve these outcomes, and this is facilitated by engaging in methodological practices that systematically capitalize on chance, which, in turn, lead to lack of reproducibility and replicability. Irreproducible and non-replicable research results threaten the credibility, usefulness, and very foundation of all scientific fields; IB is certainly not immune.

Our intention in this editorial is not to point fingers at authors, journal editors, or reviewers. Rather, we believe that there are systemic issues that we must tackle collectively because they are the result of multiple causes operating at different levels of analysis. They include, among other factors, author motivation, methodological training (or lack thereof) of authors and reviewers, the rapid progress of methodological advancements, the availability of large, archival datasets, the low cost of computing tools, increased competition for journal space, pressures on universities to produce increasingly high levels of research output, and university promotion and tenure systems that encourage publishing as many articles as possible in the so-called top journals.

We addressed five admittedly selective issues that are particularly prone to being affected by systematic capitalization on chance. Some of the issues we discussed are not new and have already been noted in the methodological literature and also in the substantive literature in IB (e.g., Cascio, 2012). We also offered suggestions on how to minimize the detrimental effects of capitalization on chance. But, realistically, even if researchers are aware of how to do things right, the issue of context (i.e., reward systems, manuscript-review processes) will remain as powerful hurdles. Thus we believe that a critical and necessary step is to enforce good methodological practices through the review process and also journal policies – because these are actions within the purview of journals. For example, Verbeke et al. offered guidelines for reviewers, including being "promoters of good methods" (Verbeke et al., 2017: 6) and the *Journal of Management* has recently included the following item on its reviewer-evaluation form: "To ensure that all papers have at least one reviewer with deep knowledge of the methods used, given your expertise in the statistical methods used in this paper, please indicate your comfort/confidence in your ability to rigorously evaluate the results reported: (Very uncomfortable, some discomfort, comfortable, confident, very confident, not applicable)" (Wright, 2016).

As an actionable implication of our discussion, we offer the following modest proposal. Our recommendation is to include additional items on the manuscript-submission form such that authors acknowledge, for example, whether hypotheses were created retroactively after examining the results. Similar items can be included on the manuscript-submission form regarding the selection of variables in a model, handling of control variables and outliers, and other methodological choices and judgment calls that capitalize on chance systematically. Clearly, not all methodological details can be included in a manuscript itself due to page limitations, and this is why some journals have chosen to reduce the font size of the Method section (Cortina et al., 2017). However, given that many journals allow authors to submit a supplemental file to be posted online, together with any published article, page limitations as a reason for not including sufficient detail about methodological procedures are no longer a valid constraint.

In closing, we believe that the motivation *not* to engage in systematic capitalization on chance needs to be greater than the motivation to engage in such practices. Hopefully, our article will provide a small step in this direction. One thing is clear, however: Lack of reproducibility and replicability, retractions, and negative effects on the credibility and usefulness of our research are unlikely to improve if we do not take proactive and tangible actions to implement a change in course.

ACKNOWLEDGEMENTS

We thank Alain Verbeke and two *Journal of International Business Studies* anonymous reviewers for their highly constructive feedback that allowed us to improve our manuscript.

NOTES

[1]As noted by an anonymous reviewer, multilevel modeling is as susceptible to capitalization on chance as other methods, including OLS regression. Although the existence of a dependent data structure allows multilevel modeling to produce more accurate standard errors compared to OLS regression (Aguinis & Culpepper, 2015), this is an improvement regarding unsystematic but not systematic capitalization on chance.

[2]These are different ways to "manage outliers." Winsorization involves transforming extreme values to a specified percentile of the data (e.g., a 90th percentile Winsorization would transform all the data below the 5th percentile to the 5th percentile, and all the data above the 95th percentile would be set at the 95th percentile). Studentized residuals are computed by dividing a residual by an estimate of its standard deviation, and Cook's D measures the effect of deleting a given observation.

REFERENCES

Aguinis, H., & Culpepper, S. A. 2015. An expanded decision making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods*, 18(2): 155–176.

Aguinis, H., Gottfredson, R. K., & Joo, H. 2013. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2): 270–301.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. 2009. First decade of *Organizational Research Methods*: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, 12(1): 69–112.

Aguinis, H., Pierce, C. A., & Culpepper, S. A. 2009. Scale coarseness as a methodological artifact: Correcting correlation coefficients attenuated from using coarse scales. *Organizational Research Methods*, 12(4): 623–652.

Aguinis, H., Shapiro, D. L., Antonacopoulou, E., & Cummings, T. G. 2014. Scholarly impact: A pluralist conceptualization. *Academy of Management Learning and Education*, 13(4): 623–639.

Aguinis, H., & Vandenberg, R. J. 2014. An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior*, 1(1): 569–595.

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

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

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.

Bakker, M., & Wicherts, J. M. 2011. The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, 43(3): 666–678.

Bamberger, P., & Ang, S. 2016. The quantitative discovery: What is it and how to get it published. *Academy of Management Discoveries*, 2(1): 1–6.

Banks, G. C., O'Boyle, Jr., E. H. *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.

Becker, T. E. 2005. Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, 8(3): 274–289.

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

Bentler, P. M. 2007. On tests and indices for evaluating structural models. *Personality and Individual Differences*, 42(5): 825–829.

Bergh, D. D., Sharp, B., & Li, M. 2017. Tests for identifying "red flags" in empirical findings: Demonstration and recommendations for authors, reviewers and editors. *Academy of Management Learning & Education*, 16(1): 110–124.

Bernerth, J. & Aguinis, H. 2016. A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, 69(1): 229–283.

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.

Bobko, P. 2001. *Correlation and regression* (2nd edn). Thousand Oaks, CA: Sage.

Boellis, A., Mariotti, S., Minichilli, A., & Piscitello, L. 2016. Family involvement and firms' establishment mode choice in foreign markets. *Journal of International Business Studies*, 47(8): 929–950.

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.

Breaugh, J. A. 2008. Important considerations in using statistical procedures to control for nuisance variables in non-experimental studies. *Human Resource Management Review*, 18(4): 282–293.

Butler, N., Delaney, H., & Spoelstra, S. 2017. The grey zone: Questionable research practices in the business school. *Academy of Management Learning & Education*, 16(1): 94–109.

Carlson, K. D., & Wu, J. 2012. The illusion of statistical control: Control variable practice in management research. *Organizational Research Methods*, 15(3): 413–435.

Cascio, W. F. 2012. Methodological issues in international HR management research. *International Journal of Human Resource Management*, 23(12): 2532–2545.

Cascio, W. F., & Aguinis, H. 2005. Test development and use: New twists on old questions. *Human Resource Management*, 44(3): 219–235.

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

Chang, S. J., van Wittleloostuijn, A., & Eden, L. 2010. From the editors: Common method variance in international business research. *Journal of International Business Studies*, 41(2): 178–184.

Chen, E. E., & Wojcik, S. P. 2016. A practical guide to big data research in psychology. *Psychological Methods*, 21(4): 458–474.
Cortina, J. M. 2002. Big things have small beginnings: An assortment of "minor" methodological misunderstandings. *Journal of Management*, 28(3): 339–362.
Cortina, J. M., Aguinis, H., & DeShon, R. P. 2017. Twilight of dawn or of evening? A century of research methods in the *Journal of Applied Psychology*. *Journal of Applied Psychology*, 102(3): 274–290.
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*. doi: 10.1177/1094428116676345.
Credé, M., & Harms, P. D. 2015. 25 years of higher-order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. *Journal of Organizational Behavior*, 36(6): 845–872.
Cuervo-Cazurra, A., Andersson, U., Brannen, M.Y., Nielsen, B., & Reuber, A. R. 2016. From the editors: Can I trust your findings? Ruling out alternative explanations in international business research. *Journal of International Business Studies*, 47(8): 881–997.
Dalton, D. R., Aguinis, H., Dalton, C. A., Bosco, F. A., & Pierce, C. A. 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.
Davis, G. F. 2015. What is organizational research for? *Administrative Science Quarterly*, 60(2): 179–188.
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(4): 668–689.
Eysenck, H. J. 1978. An exercise in mega-silliness. *American Psychologist*, 33(5), 517.
Fanelli, D. 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, 4, e5738. doi: 10.1371/journal.pone.0005738.
Fischer, R., & Mansell, A. 2009. Commitment across cultures: A meta-analytical approach. *Journal of International Business Studies*, 40(8): 1339–1358.
Fisher, G., & Aguinis, H. 2017. Using theory elaboration to make theoretical advancements. *Organizational Research Methods*. doi: 10.1177/1094428116689707.
Fitzsimmons, S., Liao, Y., & Thomas, D. 2017. From crossing cultures to straddling them: An empirical examination of outcomes for multicultural employees. *Journal of International Business Studies*, 48(1): 63–89.
Friedman, D., & Sunder, S. 1994. *Experimental methods: A primer for economists*. New York, NY: Cambridge University Press.
Freese, J. 2007. Replication standards for quantitative social science: Why not sociology. *Sociological Methods & Research*, 36(2): 153–172.
Fung, S. K., Zhou, G., & Zhu, X. J. 2016. Monitor objectivity with important clients: Evidence from auditor opinions around the world. *Journal of International Business Studies*, 47(3): 263–294.
Funk, C. A., Arthurs, J. D., Treviño, L. J., & Joireman, J. 2010. Consumer animosity in the global value chain: The effect of international production shifts on willingness to purchase hybrid products. *Journal of International Business Studies*, 41(4): 639–651.
George, G. 2014. Rethinking management scholarship. *Academy of Management Journal*, 57(1): 1–6.
Goldfarb, B., & King, A. A. 2016. Scientific apophenia in strategic management research: Significance tests & mistaken inference. *Strategic Management Journal*, 37(1): 167–176.
Harlow, L. L., & Oswald, F. L. 2016. Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21(4): 447–457.
Hollenbeck, J. H. & Wright, P. M. 2016. Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management*, 43(1): 5–18.
Hunter, J. E., & Schmidt, F. L. 2015. *Methods of meta-analysis: Correcting error and bias in research findings* (3rd edn). Thousand Oaks, CA: Sage.
Hurley, A. E. *et al.* 1997. Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior*, 18(6): 667–683.
Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Med*, 2(8): e124. doi: 10.1371/journal.pmed.0020124.
John, L. K, Loewenstein, G., & Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23(5), 524–532.
Kepes, S., & McDaniel, M. A. 2013. How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology*, 6(3): 252–268.
Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3): 196–217.
Lisak, A., Erez, M., Sui, Y., & Lee, C. 2016. The positive role of global leaders in enhancing multicultural team innovation. *Journal of International Business Studies*, 47(6): 655–673.
Macaskill, P., Walter, S., & Irwig, L. 2001. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20(4), 641–654.
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.
Meyer, K. E., van Witteloostuijn, A., & Beugelsdijk, S. 2017. What's in a *p*? Reassessing best practices for conducting and reporting hypothesis-testing research. *Journal of International Business Studies*. doi:10.1057/s41267-017-0078-8.
Nosek, B. A., Spies, J. R., & Motyl, M. 2012. Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6): 615–631.
Nuijten, M. B., Hartgerink, C. H., Assen, M. A., Epskamp, S., & Wicherts, J. M. 2015. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4): 1–22.
O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. 2017. The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, 43(2): 376–399.
Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251): aac4716. doi: 10.1126/science.aac4716.
Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. 1991. Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology*, 44(3): 473–486.
Reeb, D., Sakakibara, M., & Mahmood, I. P. 2012. Endogeneity in international business research. *Journal of International Business Studies*, 43(3): 211–218.
Richard, P. J., Devinney, T. M., Yip, G. S., & Johnson, G. 2009. Measuring organizational performance: Towards methodological best practice. *Journal of Management*, 35(3): 718–804.
Rousseeuw, P. J., & Leroy, A. M. 2003. *Robust regression and outlier detection.* Hoboken, NJ: Wiley.
Schwab, A., & Starbuck, W. H. 2017. A call for openness in research reporting: How to turn covert practices into helpful tools. *Academy of Management Learning & Education*, 16(1): 125–141.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11): 1359–1366.
Sijtsma, K. 2016. Playing with data – Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, 81(1): 1–15.

Spector, P. E., & Brannick, M. T. 2011. Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, 14(2): 287–305.
Stahl, G. K., Maznevski, M. L., Voigt, A., & Jonsen, K. 2010. Unraveling the effects of cultural diversity in teams: A meta-analysis of research on multicultural work groups. *Journal of International Business Studies*, 41(4): 690–709.
Starbuck, W. H. 2016. 60th anniversary essay: How journals could improve research practices in social science. *Administrative Science Quarterly*, 61(2): 165–183.
van Essen, M., Heugens, P. P., Otten, J., & van Oosterhout, J. 2012. An institution-based view of executive compensation: A multilevel meta-analytic test. *Journal of International Business Studies*, 43(4): 396–423.
Verbeke, A., Von Glinow, M. Y., & Luo, Y. 2017. Becoming a great reviewer: Four actionable guidelines. *Journal of International Business Studies*, 48(1): 1–9.
Waldman, I. D., & Lilienfeld, S. O. 2016. Thinking about data, research methods, and statistical analyses: Commentary on Sijtsma's (2014) "Playing with Data". *Psychometrika*, 81(1): 16–26.
Wright, P. M. 2016. Ensuring research integrity: An editor's perspective. *Journal of Management*, 42(5): 1037–1043.

ABOUT THE AUTHORS

Herman Aguinis (PhD, University at Albany, State University of New York) is the Avram Tucker Distinguished Scholar and Professor of Management at the George Washington University School of Business. His research focuses on global talent management and organizational research methods. He has published more than 140 journal articles, is a Fellow of the Academy of Management (AOM), and has received the AOM Research Methods Division Distinguished Career Award for lifetime contributions.

Wayne F. Cascio (PhD, University of Rochester) holds the Robert H. Reynolds Distinguished Chair in Global Leadership at the University of Colorado Denver. He has published 28 books and more than 185 articles and book chapters. An editor of the *Journal of International Business Studies* (JIBS), his research focuses on global talent management. In 2016, he received the World Federation of People Management Associations' Lifetime Achievement Award.

Ravi S. Ramani (MBA, Johnson & Wales University) is a PhD Candidate in the Department of Management at the George Washington University School of Business. His research examines employees' emotional experiences at work; the changing paradigms of leadership and talent development; research methods; and practically relevant scholarship. His work received the Best Paper award from the Academy of Management Entrepreneurship Division and has been presented at several conferences.

Accepted by Alain Verbeke, Editor-in-Chief, 5 April 2017. This paper was single-blind reviewed.