Journal of International Business Studies (2021) 52, 23–44
© 2021 The Author(s) All rights reserved 0047-2506/21
www.jibs.net

Open

COMMENTARY

The anatomy of an award-winning meta-analysis: Recommendations for authors, reviewers, and readers of meta-analytic reviews

Piers Steel[1],
Sjoerd Beugelsdijk[2] and
Herman Aguinis[3]

[1] Haskayne Business School, University of Calgary, 2500 University Dr NW, Calgary, AB T2N 1N4, Canada; [2] Faculty of Economics and Business, University of Groningen, Nettelbosje 2, 9700 AV Groningen, Netherlands; [3] Department of Management, School of Business, The George Washington University, 2201 G St. NW, Washington, DC 20052, USA

Correspondence:
P Steel, Haskayne Business School, University of Calgary, 2500 University Dr NW, Calgary, AB T2N 1N4, Canada
e-mail: piers.steel@haskayne.ucalgary.ca

Abstract
Meta-analyses summarize a field's research base and are therefore highly influential. Despite their value, the standards for an excellent meta-analysis, one that is potentially award-winning, have changed in the last decade. Each step of a meta-analysis is now more formalized, from the identification of relevant articles to coding, moderator analysis, and reporting of results. What was exemplary a decade ago can be somewhat dated today. Using the award-winning meta-analysis by Stahl et al. (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, 2010) as an exemplar, we adopted a multi-disciplinary approach (e.g., management, psychology, health sciences) to summarize the anatomy (i.e., fundamental components) of a modern meta-analysis, focusing on: (1) data collection (i.e., literature search and screening, coding), (2) data preparation (i.e., treatment of multiple effect sizes, outlier identification and management, publication bias), (3) data analysis (i.e., average effect sizes, heterogeneity of effect sizes, moderator search), and (4) reporting (i.e., transparency and reproducibility, future research directions). In addition, we provide guidelines and a decision-making tree for when even foundational and highly cited meta-analyses should be updated. Based on the latest evidence, we summarize what journal editors and reviewers should expect, authors should provide, and readers (i.e., other researchers, practitioners, and policymakers) should consider about meta-analytic reviews.

Journal of International Business Studies (2021) **52**, 23–44.
https://doi.org/10.1057/s41267-020-00385-z

Keywords: meta-analysis; literature review; quantitative review; synthesis; research methodology

The online version of this article is available Open Access

Received: 28 July 2020
Accepted: 20 October 2020
Online publication date: 5 January 2021

INTRODUCTION

Scientific knowledge is the result of a multi-generational collaboration where we cumulatively generate and connect findings gleaned from individual studies (Beugelsdijk, van Witteloostuijn,

& Meyer, 2020). Meta-analysis is critical to this process, being the methodology of choice to quantitatively synthesize existing empirical evidence and draw evidence-based recommendations for practice and policymaking (Aguinis, Pierce, Bosco, Dalton, & Dalton, 2011; Davies, Nutley, & Smith, 1999). Although meta-analyses were first formally conducted in the 1970s, it was not until the following decade that they began to be promoted (e.g., Hedges, 1982; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Rosenthal & Rubin, 1982), which subsequently spread across almost all quantitative fields, including business and management (Cortina, Aguinis, & DeShon, 2017). Aguinis, Pierce, et al. (2011) reported a staggering increase from 55 business and management-related articles using meta-analysis for the 1970–1985 period to 6918 articles for the 1994–2009 period.

Although there are several notable examples of meta-analysis, there are many more that are of suspect quality (Ionnadis, 2016). Consequently, we take the opportunity to discuss components of a modern meta-analysis, noting how the methodology has continued to advance considerably (e.g., Havránek et al., 2020). To illustrate the evolution of meta-analysis, we use the award-winning contribution by Stahl, Maznevski, Voigt and Jonsen (2010) who effectively summarized and made sense of the voluminous correlational literature on team diversity and cultural differences.

It is difficult to overstate how relevant Stahl et al.'s (2010) topic of diversity has become. Having a diverse workforce that reflects the larger society has only grown as a social justice issue over the last decade (Fujimoto, Härtel, & Azmat, 2013; Tasheva & Hillman, 2019). Furthermore, team diversity also has potential organizational benefits, the "value-in-diversity" thesis (Fine, Sojo, & Lawford-Smith, 2020). Consequently, their meta-analysis speaks to the innumerable institutional efforts to increase diversity as well as those who question these efforts' effectiveness (e.g., the "Google's Ideological Echo Chamber" memo that challenged whether increasing gender diversity in the programming field would increase performance; Fortune, 2017).

The focus of our article is on meta-analytic methodology. Stahl et al. make a useful contrast as, although its methodology was advanced for its time, the field has evolved rapidly. We draw upon recently established developments to contrast traditional versus modern meta-analytic methodology, summarizing our recommendations in Table 1. Our goal is to assist authors planning to carry out a meta-analytical study, journal editors and reviewers asked to evaluate their resulting work, and consumers of the knowledge produced (i.e., other researchers, practitioners, and policymakers) highlighting common areas of concern. Accordingly, we offer recommendations and, perhaps more importantly, specific implementation guidelines that make our recommendations concrete, tangible, and realistic.

MODERN METHODOLOGY

Using Stahl et al. as an exemplar, we summarize the anatomy (i.e., fundamental components) of a modern meta-analysis, focusing on: (1) data collection (i.e., literature search and screening, coding), (2) data preparation (i.e., treatment of multiple effect sizes, outlier identification and management, publication bias), (3) data analysis (i.e., average effect sizes, heterogeneity of effect sizes, moderator search), and (4) reporting (i.e., transparency and reproducibility, future research directions). Stahl et al. graciously shared their database with us, which we re-analyzed using more recently developed procedures.

Stage 1: Data Collection

Data collection is the creation of the database that enables a meta-analysis. Inherently, there is tension between making a meta-analysis manageable, that is small enough that it can be finished, and making it comprehensive and broad to make a meaningful contribution. With the research base growing exponentially but research time and efficiency remaining relatively constant, the temptation is to limit the topic arbitrarily by journals, by language, by publication year, or by the way constructs are measured (e.g., specific measure of cultural distance). The risk is that the meta-analysis is so narrowly conceived that, as Bem (1995: 172) puts it, "Nobody will give a damn." One solution is to acknowledge that meta-analysis is increasingly becoming a "Big Science" project, requiring larger groups of collaborators. Although well-funded meta-analytic laboratories do exist, they are almost exclusively in the medical field. In business, it is likely that influential reviews will increasingly become the purview of well-managed academic crowdsourcing projects (i.e., Massive Peer Production) whose leaders can tackle larger topics (i.e., community augmented meta-analyses; Tsuji, Bergmann, & Cristia, 2014), such as exemplified by Many Labs (e.g., Klein et al., 2018).

Table 1 Summary of recommendations and implementation guidelines for authors, reviewers, and readers of meta-analytic reviews

Recommendations	Implementation guidelines
Stage 1: Data collection Organize and implement the search process and data extraction from primary-level studies	**Literature search and screening** • Acknowledge that meta-analysis is increasingly becoming a "Big Science" project, requiring larger groups of collaborators • Conduct a pre-meta-analysis scoping study to ensure that the research question is small enough to be manageable, large enough to be meaningful, there is sufficient research base for analysis, and that recent reviews have not already addressed the same topic • Ensure authors' prolonged interest and deep knowledge of the topic to be meta-analyzed • Avoid the construct identity fallacy: different measures used for the same underlying construct (i.e., jingle) and the same construct is referred to using different labels (i.e., jangle) • Avoid biases in the search process: availability bias by searching the "grey literature," cost bias by accessing pay-walled journals, familiarity bias by consulting databases in other disciplines, language bias by searching non-English journals, and The Matthew Effect by not excluding low-citation sources • Implement a variety of search strategies, including "snowballing" (aka ancestry searching or "pearl growing") • To manage and document the search process, as per PRISMA, use recent software developments, such as www.covidence.org, www.hubmeta.com, or https://revtools.net/ • Engage an information specialist (e.g., a librarian) in the search process
	Coding of the primary studies • Implement procedures such as psychometric corrections and conversion of statistics to effect size estimates (e.g., *r*s, *d*s) using available and standardized tools such *psychmeta* • Consider trade-offs between increased measurement variance and using a larger meta-analytic database by teasing apart broad constructs into component dimensions or by merging selected measures • Archive the data perpetually through an Open Science repository rather than "making data available from the authors" • Establish commensurability among measures, drawing on convergent and content validity as well as previous taxonomic work and expert opinion • Reserve kappa for checking agreement on qualitative decisions • Use a battery of measurement equivalence indexes to gather evidence that the different measures used assess the same underlying construct • Include a transparent description of the search process and taxonomy of key constructs
Stage 2: Data preparation Clean the data to perform the meta-analysis	**Treatment of multiple effect sizes** • Keep multiple correlations of the same relationship from the same sample statistically separate, preferably by using composite scores if intercorrelations between measures are available • Consider alternative techniques to group measures such as the Robust Error Variance (RVE) approach and a multilevel meta-analytic approach
	Outlier identification and management • Do not use arbitrary cutoffs to identify and eliminate outliers • Conduct analyses to determine whether outlying observations are error, influential, or interesting outliers • Consider the possibility that some outliers may be legitimate observations • Report results with and without outliers
	Publication bias • Complement or replace the fail-safe N procedure to detect publication bias with a selection-based method, such as published versus unpublished studies, symmetry methods such as Egger's regression, Trim-and-Fill technique, and the precision-effect test and estimate with standard errors (PET-PEESE)

Table 1 (*Continued*)

Recommendations	Implementation guidelines
Stage 3: Data analysis Assess heterogeneity of effect sizes	**Average effect sizes** • Report the average association between variables as the *initial* stage of theory testing • Report not only the average size but also its meaning and importance by placing it within a particular context and domain • Use contemporary effect-size benchmarks such as small = 0.10, medium = 0.18, and large = 0.32 for correlations • Adopt a random-effects and, if using psychometric corrections, Morris weights rather than a fixed-effects approach to calculating effect sizes • Go beyond average effect sizes by using them as input for subsequent meta-analytic structural equation modeling (MASEM) • Extend or fill out the MASEM matrix with results derived from Individual Participant Data (IPD) • Address nonsensical meta-analytically-derived correlation matrices by excluding problematic cells, collapsing highly correlated variables into factors to avoid multicollinearity. **Heterogeneity of effect sizes** • Assess the degree of dispersion of effect sizes around the average • Report heterogeneity of effect sizes, providing at a minimum credibility intervals, T^2 (i.e., SD_r or the random-effects variance component), and I^2 (i.e., percentage of total variance attributable to T^2) • Employ a Bayesian approach that corrects for artificial homogeneity created by small samples • Use asymmetric distributions in the case of skewed credibility intervals **Moderator search** • Organize the search for moderators using Cattell's Data Cube: (a) sample, (b) variables, and (c) occasions • Implement meta-regression (MARA) instead of subgrouping analysis when assessing continuous moderators
Stage 4: Reporting Ensure transparency and that meta-analytic progress continues	**Transparency and reproducibility** • Describe all procedures in sufficient detail so that others will be able to reproduce all data collection and analysis steps • Make the meta-analytic database available in an Open Science archive • If practical, turn your meta-analysis into a "living systematic review" that can be updated in real time **Future research directions** • Write future research directions as if you were in charge of the field and needed to direct subsequent studies, highlighting important understudied relationships • Consider future meta-analyses focused on alternative construct definitions and measures • Direct future projects towards understudied elements and away from relationships that have been overly emphasized, perhaps to the point of recommending a moratorium • Describe what moderators need to be considered in future research (e.g., sample characteristics, variables, contextual variation) • Determine the need to update a meta-analysis by using the decision framework summarized in Figure 1

With a large team or a smaller but more dedicated group, researchers have a freer hand in determining how to define the topic and the edges that define the literature. To this end, Tranfield, Denyer and Smart (2003) discussed that the identification of a topic, described as Phase 0, "may be an iterative process of definition, clarification, and refinement" (Tranfield et al., 2003: 214). Relatedly, Siddaway, Wood and Hedges (2019) highlighted scoping and planning as key stages that precede the literature search and screening procedures. Indeed, it is useful to conduct a pre-meta-analysis scoping study, ensuring that the research question is small enough to be manageable, large enough to be meaningful, there is sufficient research base for analysis, and that other recent or carried out reviews have not already addressed the same topic. Denyer and Tranfield (2008) stressed how an author's prior and prolonged interest in the topic is immensely helpful, exemplified by a history of publishing in a particular domain. In fact, deep familiarity with the nuances of a field assists in every step of a meta-analytic review. Consistent with this point, Stahl et al.'s *References* section shows this familiarity, containing multiple publications by the first two authors. Gunter Stahl has emphasized cultural values while Martha Maznevski has focused on team development, with enough overlap between the two that Maznevski published in a handbook edited by Stahl (Maznevski, Davison, & Jonsen, 2006).

Once a worthy topic within one's capabilities has been established, the most arduous part of meta-analysis begins. First is the literature search and screening (i.e., locating and obtaining relevant studies) and second is coding (i.e., extracting the data contained within the primary studies).

Literature search and screening

Bosco, Steel, Oswald, Uggerslev and Field (2015) alluded to academia's "Tower of Babel" or what Larsen and Bong (2016) more formally labeled as the construct identity fallacy. These terms convey the idea that there can be dozens of terms and scores of measures for the same construct (i.e., jingle) and different constructs can go by the same name (i.e., jangle), such as cultural distance versus the Kogut and Singh index (Beugelsdijk, Ambos, & Nell, 2018; Maseland, Dow, & Steel, 2018). Furthermore, many research fields have exploded in size, almost exponentially (Bornmann & Mutz, 2015), making a literature search massively harder. Then there are the numerous databases within which the targeted articles may be hidden due to their often flawed or archaic organization (Gusenbauer & Haddaway, 2020), especially their keyword search functions. As per Spellman's (2015) appraisal, "Our keyword system has become worthless, and we now rely too much on literal word searches that do not find similar (or analogous) research if the same terms are not used to describe it" (Spellman, 2015: 894).

Given this difficulty and that literature searches often occur in an iterative manner, where researchers are learning the parameters of the search as they conduct them (i.e., "Realist Search"; Booth, Briscoe, & Wright, 2020), there is an incentive to filter or simplify the procedure and to not properly document such a fundamentally flawed process so as to not leave it open to critique from reviewers' potentially idealistic standards (Aguinis, Ramani, & Alabduljader, 2018). The result can be an implicit selection bias, where the body of articles is a subset of what is of interest (Lee, Bosco, Steel, & Uggerslev, 2017). Rothstein, Sutton and Borenstein (2005) described four types of bias: availability bias (selective inclusion of studies that are easily accessible to the researcher), cost bias (selective inclusion of studies that are available free or at low costs), familiarity bias (selective inclusion of studies only from one's own field or discipline), and language bias (selective inclusion of studies published in English). The last of these is particularly common as well as particularly ironic in international business (IB) research. To this list, we would like to add citation bias due to *The Matthew Effect* (Merton, 1968). With increased public information on citation structures thanks to software such as Google Scholar, there is the risk of a selective inclusion of those studies that are heavily cited, at the expense of studies that have not been picked up (yet). Each of these biases can be addressed, respectively, by searching the grey literature, finding access to paywalled scientific journals, including databases outside one's discipline, engaging in translation (at least those languages used in multiple sources), and not using a low citation rate as an exclusion criterion.

How was Stahl et al.'s literature search process? Adept for its time. They drew from multiple databases, which is recommended (Harari, Parola, Hartwell, & Riegelman, 2020), and they supplemented with a variety of other techniques, including manual searches. They provided a sensible set of keywords but also contacted researchers operating in the team field to acquire the "grey literature" of obscure or unpublished works. Some other techniques could be added, such Ones, Viswesvaran and Schmidt's (2017) suggestion that "snowballing" (aka "ancestry searching" or "pearl growing"; Booth, 2008) should be *de rigueur*. In other words, "by working from the more contemporary references for meta-analysis, tracking these references for the prior meta-analytic work on which they relied, and iteratively continuing this process, it is possible to identify a set of common early references with no published predecessors" (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011: 9). At present, however, some of Stahl et al.'s efforts would likely be critiqued in terms of replicability or reproducibility and transparency (Aguinis et al., 2018; Beugelsdijk et al., 2020). For example, if the keywords "team" and "diversity" are entered as search terms, Google Scholar alone yields close to two million hits. Other screening processes must have occurred, though are not reported, reflected in that Stahl et al. provided a *sampling* of techniques designed to reassure reviewers that they made a concerted effort (e.g., "searches were performed on several different databases, including…. search strategies included…", Stahl et al., 2010: 697).

Presently, in efforts to increase transparency and replicability, the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) method is often recommended, which requires being extremely explicit about the exact databases, the exact search terms, and the exact results, including duplicates and filtering criteria (Moher, Liberati, Tetzlaff, & Altman, 2009). Although more onerous, the PRISMA-P version goes even further in terms of transparency, advocating pre-registering of the entire systematic review protocol encapsulated in a 17-item checklist (Moher et al., 2015). And, at present, the 2020 version of PRISMA recommends a 27-item checklist, not including numerous subitems, again with the goal of improving the trustworthiness of systematic reviews (Page et al., 2020). Given the attempt to minimize decisions in situ, proper adherence to the PRISMA protocols can be difficult when searches occur in an iterative manner, as researchers find new terms or measures as promising leads for relevant papers. When this happens, especially during the later stages of data preparation, researchers face the dilemma of either re-conducting the entire search process with the added criteria (substantively increasing the workload) or ignoring the new terms or measures (leading to a less than exhaustive search). New software has been developed to help address that search processes can be informed simultaneously with implementation, such as www.covidence.org, www.hubmeta.com, or https://revtools.net/ (with many more options curated at http://systematicreviewtools.com/, The Systematic Review Tool Box). They provide a computer-assisted walk-through of the search as well as a screening process, which starts with deduplication, and filtering on abstract or title, followed by full text filtering (with annotated decisions). Reviewers should expect that this information be reported in a supplemental file, along with the final list of all articles coded and details regarding effect sizes, sample sizes, measures, moderators, and other specific details that would enable readers to readily reproduce the creation of the meta-analytic database.

It is a challenge to determine that a search approach has been thorough and exhaustive, given that reviewers may have an incomplete understanding of the search criteria or of how many articles can be expected. In other words, although the authors may have reported detailed inclusion and exclusion criteria, as per MARS (Kepes, McDaniel, Brannick, & Banks, 2013), how can reviewers evaluate their adequacy? We anticipate that in the future this need for construct intimacy may be emphasized and a meta-analysis would require first drawing upon or even publishing a deep review of the construct. For example, prior to publishing their own award-winning monograph on Hofstede's cultural value dimensions (Taras, Kirkman, & Steel, 2010), two of the authors published a review of how culture itself was assessed (Taras, Rowney, & Steel, 2009), as well as a critique of the strengths and challenges of Hofstede's measure (Taras & Steel, 2009). Another example is a pre-meta-analytic review of institutional distance (Kostova, Beugelsdijk, Scott, Kunst, Chua, & van Essen, 2020), where several of the authors previously published on the topic (e.g., Beugelsdijk, Kostova, Kunst, Spadafora, & van Essen, 2018; Kostova, Roth, & Dacin, 2008; Scott, 2014). Once authors have demonstrated prolonged and even affectionate familiarity with the topic ("immersion in the literature"; DeSimone, Köhler, & Schoen, 2019: 883), reviewers may be further reassured that the technical aspects of the search were adequately carried out if a librarian (i.e., an information specialist) was reported to be involved (Johnson & Hennessy, 2019).

Coding of the primary studies

Extracting all the information from a primary study can be a lengthy procedure, as a myriad of material is typically needed beyond the basics of sample size and the estimated size of a relationship between variables (i.e., correlation coefficient). This includes details required for psychometric corrections, conversion from different statistical outputs to a common effect size (e.g., *r* or *d*), and study conditions and context that permit later moderator analysis (i.e., conditions under which a relationship between variables is weaker or stronger). Properly implementing procedures such as applying psychometric corrections for measurement error and range restriction is not always straightforward (Aguinis, Hill, & Bailey, 2021; Hunter, Schmidt, & Le, 2006; Schmidt & Hunter, 2015; Yuan, Morgeson, & LeBreton, 2020). However, while this used to be a manual process requiring intimate statistical knowledge (e.g., including knowledge of how to correct for various methodological and statistical artifacts), fortunately, this process is increasingly semi-automated. For example, the meta-analytic program *psychmeta* (the psychometric meta-analysis toolkit) provides conversion to correlations for "Cohen's d, independent samples t values (or their *p* values), two-group one-way ANOVA F values (or their *p* values), $1df\,\chi^2$ values (or their *p* values), odds ratios, log odds ratios, Fisher *z*, and the common language effect size (CLES, A, AUC)" (Dahlke & Wiernik, 2019).

However, a pernicious coding challenge is related to the literature search and screening process described earlier. For initial forays into a topic, a certain degree of conceptual "clumping" is necessary to permit sufficient studies for meta-analytic summary, in which we trade increased measurement variance for a larger database. As more studies become available, it is possible to make more refined choices and to tease apart broad constructs into component dimensions or adeptly merge selected measures to minimize mono-method bias (Podsakoff, MacKenzie, & Podsakoff, 2012). For example, Richard, Devinney, Yip and Johnson's (2009) study on organizational performance found not all measures to be commensurable, such as return on total assets often being radically different from return on sales. As a result, only a subset of the obtained literature actually represents the target construct, and this subset can be difficult to determine.

Stahl et al. methodically reported how they coded cultural diversity as well as each of their dependent variables. This is an essential start, but, reflecting the previous problem of construct proliferation, more information regarding how each dependent variable was operationalized in each study would be a welcome addition. Although some information regarding the exact measures used is available directly from the authors, which was readily provided upon request, today many journals require these data to be perpetually archived and available through an Open Science repository. The issue of commensurability applies here, as one of Stahl et al.'s dependent variables was creativity. Ma's (2009) meta-analysis on creativity divided the concept into three groups, with many separating problem-solving from artistic creativity. With only five studies on creativity available, mingling of different varieties of creativity is necessary. Still, it is important to note that Stahl et al. chose to treat studies that focused on the quality of ideas generated (e.g., Cady & Valentine, 1999) as an indicator of creativity along with more explicit measures, such as creativity of story endings (Paletz, Peng, Erez, & Maslach, 2004), leaving room for this to be re-explored as the corpus of results expanded.

To help alleviate concerns of commensurability, it is commendable that Stahl et al. used two independent raters to code the articles, documenting agreement using Cohen's kappa. Notably, kappa is used to quantify interrater reliability for *qualitative* decisions, where there is a lack of an irrefutable gold standard or "the 'correctness' of ratings cannot be determined in a typical situation" (Sun, 2011: 147). Too often, kappa is used indiscriminately to include what should be indisputable decisions, such as sample size, and when there is disagreement, coders can simply reference the original document. Qualitative judgements, where there are no factual sources to adjudicate, reflect kappa's intended purpose. Consequently, kappa can be inflated simply by including prosaic data entry decisions that reflect transcription (where it may suffice to mention double-coding with errors rectified by referencing the original document), and, with Stahl et al. reporting kappa "between .81 and .95" (Stahl et al., 2010: 699), it is unclear how it was used in this case.

Consequently, reviewers should expect authors to provide additional reassurance beyond kappa that they grouped measures appropriately. This is not simply a case of using different indices of interrater agreement (LeBreton & Senter, 2008), which often prove interchangeable themselves, but using a battery of options to show measurement equivalence and that these measures are tapping into approximately the same construct. Although few measures will be completely identical (i.e., parallel forms), there are the traditional choices of showing different types of validity evidence (Wasserman & Bracken, 2003). For example, Taras et al. (2010) were faced with over 100 different measures of culture in their meta-analysis of Hofstede's Values Survey Module. Their solution, which they document over several pages, was to begin with the available *convergent validity evidence*, that is factor or correlational studies. Given that the available associations were incomplete, they then proceeded to *content* *validity evidence*, examining not just the definitions but also the survey items for consistency with the target constructs. Finally, for more contentious decisions, they drew on 14 raters to gather further evidence regarding content validity.

As can be seen, demonstrating that different measures tap into the same construct can be laborious, and preferably future meta-analyses should be able to draw on previously established ontologies or taxonomic structures. As mentioned, there are some sources to rely on, such as Richard et al.'s (2009) work on organizational performance, Versteeg and Ginsburg's (2017) assessment of rule of law indices, or Stanek and Ones' (2018) taxonomy of personality and cognitive ability. Unfortunately, this work is still insufficient for many meta-analyses, and such a void is proving a major obstacle to the advancement of science. The multiplicity of overlapping terms and measures creates a knowledge management problem that is increasingly intractable for the individual researcher to solve. Larsen, Hekler, Paul and Gibson (2020) argued that a solution is manageable, but we need a sustained "collaborative research program between information systems, information science, and computer science researchers and social and behavioral science researchers to develop information system artifacts to address the problem" (Larsen et al., 2020: 1). Once we have an organized system of knowledge, they concluded: "it would enable scholars to more easily conduct (possibly in a fully automated manner) literature reviews, meta-analyses, and syntheses across studies and scientific domains to advance our understanding about complex systems in the social and behavioral sciences" (Larsen et al., 2020: 9).

Stage 2: Data Preparation

Literature search, screening, and coding provide the sample of primary studies and the preliminary meta-analytic database. Next, there are three aspects of the data preparation stage that leave quite a bit of discretionary room for the researcher, thus requiring explicit discussion useful not only for meta-analysts but also for reviewers as well as research consumers. First, there is the treatment of multiple effect sizes reported in a given primary-level study. Second, there is the identification and treatment of outliers. And third, the issue of publication bias.

Treatment of multiple effects sizes

A single study may choose to measure a construct in a variety of ways, each producing its own effect size estimate. In other words, effect sizes are calculated using the same sample and reported separately for each measure. Separately counting each result violates the principle of statistical independence, as all are based on the same sample. Stahl et al. chose to average effect sizes within articles, which addresses this issue; however, more effective options are now available (López-López, Page, Lipsey, & Higgins, 2018).

Typically, the goal is to focus on the key construct, and so Schmidt and Hunter (2015) recommended the calculation of composite scores, drawing on the correlations between the different measures. Unless the measures are unrelated (which suggests that they assess different constructs and therefore should not be grouped), the resulting composite score will have better coverage of the underlying construct as well as higher reliability. Other techniques include the Robust Error Variance (RVE) approach (Tanner-Smith & Tipton, 2014), which considers the dependencies (i.e., covariation) between correlated effect sizes (i.e., from the same sample). Another option is adopting a multilevel meta-analytic approach, where Level 1 includes the effects sizes, Level 2 is the within-study variation, and Level 3 is the between-study variation (Pastor & Lazowski, 2018; Weisz et al., 2017). A potential practical limitation is that these alternatives to composite scores pose large data demands, as they typically require 40–80 studies per analysis to provide acceptable estimates (Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, 2015).

Outlier identification and management

Although rarely carried out (Aguinis, Dalton, et al., 2011), outlier analysis is strongly recommended for meta-analysis. Some choices include doing nothing, reducing the weight given to the outlier, or eliminating the outlier altogether (Tabachnik & Fidell, 2014). However, whatever the choice, it should be transparent, with the option of reporting results both with and without outliers. To detect outliers, the statistical package *metafor* provides a variety of influential case diagnostics, ranging from externally standardized residuals to leave-one-out estimates (Viechtbauer, 2010). There are multiple outliers in Stahl et al.'s dataset, such as Polzer, Crisp, Jarvenpaa and Kim (2006) for Relationship Conflict, Maznevski (1995) for Process Conflict, and Gibson and Gibbs (2006) for Communication. In particular, Cady and Valentine (1999), which is the largest study for the outcome measure of Creativity and reports the sole negative correlation of − 0.14, almost triples the residual heterogeneity (Tau^2), increasing it from 0.025 to 0.065. As is the nature of outliers, and as will be shown later, their undue influence can substantially tilt results by their inclusion or exclusion.

Like the Black Swan effect, an outlier may be a legitimate effect size drawn by chance from the ends of a distribution, which would relinquish its outlier status as more effects reduce or balance its impact. Aguinis, Gottfredson and Joo (2013) offered a decision tree involving a sequence of steps to first identify outliers (i.e., whether a particular observation is far from the rest) and then decide whether specific outliers are errors, interesting, or influential. Based on the answer, a researcher can decide to eliminate it (i.e., if it is an error), retain it as is or decrease its influence, and then, regardless of the choices, it is recommended to report results with and without the outliers. Stahl et al. retained outliers, which is certainly preferable to using arbitrary cutoffs such as two standard deviations below or above the mean to omit observations from the analysis (a regrettable practice that artificially creates homogeneity; Aguinis et al., 2013). However, we do not have information on whether these outliers could have been errors.

Publication bias

Publication bias refers to a focus on statistically significant or strong effect sizes rather than a representative sample of results. This can happen for a wide of variety of reasons, including underpowered studies and questionable research practices such as *p*-hacking (Meyer, van Witteloostuijn & Beugelsdijk, 2017; Munafò et al., 2017), and it occurs frequently in a variety of fields (Ferguson & Brannick, 2012; Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014), although not all (Dalton, Aguinis, Dalton, Bosco, & Pierce, 2012). When it does occur, it has the potential to severely distort findings (Friese & Frankenbach, 2020). It is notable that Stahl et al. tested for publication bias, while only 3–30% of meta-analyses include this step (Aguinis, Dalton, et al., 2011; Kepes, Banks, McDaniel, & Whetzel, 2012). To test for publication bias, Stahl et al. used the fail-safe *N*, devised by Rosenthal (1979) for experimental research. Although Rosenthal focused on the common "file drawer" problem, his statistic is more of a general indicator of the stability of meta-analytic results (Carson, Schriesheim, & Kinicki, 1990; Dalton et al., 2012). In particular, the fail-safe *N* estimates the number of null studies that would be needed to change the average effect size a group of studies to a specified statistical significance level, especially non-significance (e.g., $p > .05$).

While at one time the fail-safe *N* was a recommended component of a state-of-the-science meta-analysis, this time has now passed. It has a variety of problems. For example if the published literature indicates a lack of relationship, that is the null itself, the equation becomes unworkable. For example, Stahl et al. were unable to give a fail-safe *N* precisely for the variables which were not significant in the first place. Consequently, for decades, researchers have recommended its disuse (Begg, 1994; Johnson & Hennessy, 2019; Scargle, 2000). Sutton (2009: 442) described it as "nothing more than a crude guide", and Becker (2005: 111) recommended that "the fail-safe *N* should be abandoned in favor of other more informative analyses." At the very least, the fail-safe *N* should be supplemented.

Although there are no perfect methods to detect or correct for publication bias, there are a wide variety of better options (Kepes et al., 2012). We can use selection-based methods and compare study sources, typically published versus unpublished, with the expectation there should be little difference between the two (Dalton et al., 2012). Also, there are a variety of symmetry-based methods, essentially where the expectation is that sample sizes or standard errors should be unrelated to effect sizes. One of most popular of these symmetry techniques is Egger's regression test, which we applied to Stahl et al. Confirming Stahl et al.'s findings, there was no detectable publication bias.

Henmi and Compas (2010) developed a simple method for reducing the effect of publication bias, which uses fixed-effect model weighting to reduce the impact of errant heterogeneity. Alternatively, the classic Trim-and-Fill technique (Duval, 2005) can also be employed, which will impute the "missing" correlations. For a more sophisticated option, there is the precision-effect test and a precision-effect estimate with standard errors (PET-PEESE), which can detect as well as correct for publication bias (see Stanley and Doucouliagos 2014 for illustrative examples and code for Stata and SPSS). Stanley (2017) identified when PET-PEESE becomes unreliable, typically when there are few studies, excessive heterogeneity, or small sample sizes, which are often the same conditions that weaken the effectiveness of meta-analytic techniques in general.

Stage 3: Data Analysis

Meta-analyses are overwhelmingly used to understand what is the overall (i.e., average) size of the relationship between variables across primary-level studies (DeSimone et al., 2019; Carlson & Ji, 2011). However, meta-analysis is just as useful, if not more so, to understand when and where a relationship is likely to be stronger or weaker (Aguinis, Pierce, et al., 2011). Consequently, we discuss the three basic elements of the data analysis stage – average effect sizes, heterogeneity, and moderators – and we emphasize theory implications.

Reflecting that many meta-analytic methodologies were under debate at that time, Stahl et al. used a combination of techniques, including psychometric meta-analysis, both a fixed-effect and random-effect approach, as well as converting correlations to Fisher's *z* after psychometric adjustments. The motivation for this blend of techniques is clear: each has its advantages (Wiernik & Dahlke, 2020). However, procedures have been refined and, consequently, we contrast Stahl et al.'s results with a modern technique that better accomplishes their aim: Morris estimators (Brannick, Potter, Benitez, & Morris, 2019).

Average effect sizes

During the early years of meta-analysis, the main question of interest was: "Is there a consistent relationship between two variables when examined across a number of primary-level studies that seemingly report contradictory results?" As Gonzalez-Mulé and Aguinis (2018) reviewed, for many meta-analyses, this is all they provided. Showing association and connection represents the *initial* stages of theory testing, and most meta-analyses have some hypotheses attached to these estimates. Given that this is the lower-hanging empirical and theoretical fruit, much of it has already been plucked and, today, unlikely by itself to satisfy demands for a novel contribution. An improved test of theory at this stage is not just positing that a relationship exists, and that it is unlikely to be zero, but how big it is (Meehl, 1990); in other words, "Instead of treating meta-analytic results similarly to NHST (i.e., limiting the focus to the presence or absence of an overall relationship), reference and interpret the MAES (meta-analytic effect sizes) alongside any relevant qualifying information" (DeSimone et al., 2019: 884). To this end, researchers have typically drawn on Cohen (1962), who made very rough benchmark estimates based on his review of articles published in the 1960 volume of *Journal of Abnormal and Social Psychology*. Contemporary effect-size estimates have been compiled by Bosco, Aguinis, Singh, Field and Pierce (2015), who drew on 147,328 correlations reported in 1660 articles, and by Paterson, Harms, Steel and Credé (2016), who summarized results from more than 250 meta-analyses. Both ascertained that Cohen's categorizations of small, medium, and large effects do not accurately reflect today's research in management and related fields. Averaging Bosco et al.'s Table 2 and Paterson et al.'s Table 3, a better generic distribution remains as per Cohen 0.10 for small (i.e., 25th percentile), but 0.18 for medium (i.e., 50th percentile) and 0.32 for large (i.e., 75th percentile). Using these distributions of effect sizes, or those compiled from other analogous meta-analyses, meta-analysts can go beyond the simple conclusion that a relationship is different from zero and, instead, critically evaluate the size of the effect within the context of a specific domain.

Stahl et al. adopted a hybrid approach to calculate average effect sizes. Initially, they reported estimates using Schmidt and Hunter's (2015) psychometric meta-analysis, correcting for dichotomization (i.e., uneven splits) and attenuation due to measurement error. They then departed from Schmidt's and Hunter's approach by transforming correlations to Fisher's *z*s (Borenstein, Hedges, Higgins, & Rothstein, 2009) and weighting by $N - 3$, the inverse of sampling error after Fisher's transformation. As Stahl et al. clearly acknowledged, this is a fixed-effects approach that assumes the existence of a single population effect. In contrast, a random-effects model assumes that there are multiple population effects that motivate the search for moderator (i.e., factors that account for substantive variability of observed effects).

Where does this leave Stahl et al., who corrected for attenuation but used a Fisher's *z* transformation with an underlying fixed-effect approach? If correlations are between $\pm$ 0.30, Fisher's *z* transformed versus untransformed correlations are almost identical. For Stahl et al.'s data, 81% of their effect sizes fell within this special case of near equivalence, making the matter almost moot. Similarly, Schmidt and Hunter (2015) used an attenuation factor, which can change weights drastically, but here the average absolute difference between raw and corrected correlation is less than 0.02, minimizing this concern. Consequently, although we do not recommend Stahl et al.'s fixed-effects approach, results should be close to equivalent to other methods, as noted by Aguinis, Gottfredson and Wright (2011).

As mentioned earlier, we re-analyzed Stahl et al.'s data using Morris weights. To calculate variance of effect sizes across primary-level studies, we used $N - 1$ in the formula rather than N, as the effect sizes are estimates and not population values. To calculate residual heterogeneity (i.e., whether variation of effect sizes is due to substantive rather than artifactual reasons), Morris estimators rely on restricted maximum likelihood. We conducted all analyses using the *metafor* (2.0-0) statistical package (Viechtbauer, 2010) in R (version 3.5.3). We found that the average effect size for creativity, for example, increased from Stahl et al.'s 0.16 to 0.18, although it was non-significant ($p = .20$). Moreover, using the random-effects model, which increased the size of confidence intervals due the inclusion of the random-effects variance component (REVC), none of the effects were significant, with a caveat due to the consideration of outliers. If we exclude Cady and Valentine (1999), the effect size of creativity increases to 0.29 and became significant ($p = 0.02$). In sum, Stahl et al. provided an excellent example that methodological choices, here regarding outliers and the model, are influential enough that a meta-analysis' major conclusions can hinge upon them.

Stahl et al. presented a single column of effect sizes, which is now insufficient for modern meta-analyses. What is preferred is a grid of them. For example, meta-analytic structural equation modeling (MASEM) is based on expanding the scope of a meta-analysis from bivariate correlations to creating a full meta-analytic correlation matrix (Bergh et al., 2016; Cheung, 2018; Oh, 2020). Given that this allows for additional theory testing options enabled by standard structural equation modeling, the publication of a meta-analysis can pivot on its use of MASEM. Options range from factor analysis to path analysis, such as determining the total variance provided by predictors or if a predictor is particularly important (e.g., dominance or relative weights analysis). It also allows for mediation tests, that is, the "how" of theory or "reasons for connections." It is even possible to use MASEM to test for interaction effects. Traditionally, the correlation between the interaction term and other variables is not reported and often must be requested directly from the original authors. Doing so is a high-risk endeavor given researchers' traditionally low response rate (Aguinis, Beaty, Boik, & Pierce, 2005; Polanin et al., 2020a), but the rise of Open Science and the concomitant Individual Participant Data (IPD) means that this information is increasingly available. Amalgamating IPD across multiple studies is usually referred to as a mega-analysis, and, as suggested here, can be used to supplement a standard meta-analysis (Boedhoe et al., 2019; Kaufmann, Reips, & Merki, 2016).

Reviewers will note that, as researchers move from simply an average of bivariate relationships towards MASEM, they can encounter incomplete and nonsensical matrices. For incomplete matrices, Landis (2013) and Bergh et al. (2016) provided sensible recommendations for filling blank cells in a matrix, such as drawing on previously published meta-analytic values or expanding the meta-analysis to target missing correlations. Nonsensical matrices (that occur increasingly as correlation matrices expand) create a non-positive definite "Frankenstein" matrix, stitched together from incompatible moderator patches. Landis (2013), as well as Sheng, Kong, Cortina and Hou (2016), provided remedies, such as excluding problematic cells or collapsing highly correlated variables into factors to avoid multicollinearity. In addition, we can employ more advanced methods that incorporate random effects and dovetail meta-regression with MASEM (e.g., Jak & Cheung, 2020). The benefit is a mature science that can adjust a matrix so that the resulting regression equations represent specific contexts. For example, synthetic validity is a MASEM application in which validity coefficients are predicted based on a meta-regression of job characteristics, meaning that we can create customized personnel selection platforms orders of magnitude less costly, faster, and more accurately (Steel, Johnson, Jeanneret, Scherbaum, Hoffman, & Foster, 2010).

Heterogeneity of effect sizes

A supplement to our previous discussion of average effect sizes is the degree of dispersion around the average effect. As noted by Borenstein et al. (2009), "the goal of a meta-analysis should be to synthesize the effect sizes, and not simply (or necessarily) to report a summary effect. If the effects are consistent, then the analysis shows that the effect is robust across the range of included studies. If there is modest dispersion, then this dispersion should serve to place the mean effect in context. If there is substantial dispersion, then the focus should shift from the summary effect to the dispersion itself. Researchers who report a summary effect are indeed missing the point of the synthesis" (Borenstein et al., 2009: 378).

Stahl et al. examined whether the homogeneity Q statistic was significant, meaning that sufficient variability of effects around the mean exists, as a precursor to moderator examination. A modern meta-analysis should complement the Q statistic with other ways of assessing heterogeneity, because Q often leads to Type II errors (i.e., incorrect conclusions that heterogeneity is not present; Gonzalez-Mulé & Aguinis, 2018), especially when there is publication bias (Augusteijn, van Aert, & van Assen, 2019). Further reporting of heterogeneity by Stahl et al. is somewhat unclear. They provided in Table 2 "Variance explained by S.E. (%)" and "Range of effect sizes," which were not otherwise explained. This oversight is, as Gonzalez-Mulé and Aguinis (2018) documented, regrettably common. In fact, they found that 16% of meta-analyses from major management journals fail to report heterogeneity at all. Stahl et al. reported the range of effect sizes for creativity was – .14 to .48. However, the actual credibility intervals, after removing the outlier, was .03 to .55, indicating that the result typically generalizes and can be strong. As per Gonzalez-Mulé and Aguinis, we recommend providing at a minimum: credibility intervals, T^2 (i.e., SD_r or the REVC), and I^2 (i.e., percentage of total variance attributable to T^2). The ability to further assess heterogeneity is facilitated by recent methodological advances, such as the use of a Bayesian approach that corrects for artificial homogeneity created by small samples (Steel, Kammeyer-Mueller, & Paterson, 2015), and by the use of asymmetric distributions in cases of skewed credibility intervals (Baker & Jackson, 2016; Jackson, Turner, Rhodes, & Viechtbauer, 2014; Possolo, Merkatas, & Bodnar, 2019).

Moderator search

Moderating effects, which account for substantive heterogeneity, can be organized around Cattell's Data Cube or the Data Box (Revelle & Wilt, 2019): (1) sample (e.g., firm or people characteristics), (2) variables (e.g., measurements), and (3) occasions (e.g., administration or setting). Typical moderator variables include country (e.g., developing vs developed), time period (e.g., decade) and published vs unpublished status (where comparison between the two can indicate the presence of publication bias). Particularly important from an IB perspective is the language and culture of survey administration, which has been shown to influence response styles (Harzing, 2006; Smith & Fischer, 2008) and response rates (Lyness & Brumit Kropf, 2007). Theory is often addressed as part of the moderator search, as per Cortina's (2016) review, "a theory is a set of clearly identified variables and their connections, the reasons for those connections, and the primary boundary conditions for those connections" (Cortina, 2016: 1142). Moderator search usually establishes the last of these – *boundary conditions* – although not exclusively. For example, Bowen, Rostami and Steel (2010) used the temporal sequence as a moderator to clarify the causal relationship between innovation and firm performance. Of note, untheorized moderators (e.g., control variables) are still a staple of meta-analyses but should be clearly delineated as robustness tests or sensitivity analyses (Bernerth & Aguinis, 2016).

After establishing average effect sizes (i.e., connections), Stahl et al. grappled deeply with the type of diversity, a boundary condition inquiry that determines how specific contexts affect these connections or effect sizes. Stahl et al. differentiated between the role of surface level (e.g., racio-ethnicity) vs deep level (e.g., cultural values) diversity and noted trade-offs. They expected diversity to be associated with higher levels of creativity, but at the potential cost of lower satisfaction and greater conflict, negative outcomes that likely diminish as team tenure increases. Note how well these moderators match up to core theoretical elements. Page's (2008) book on diversity, *The Difference*, covers in detail the four conditions that lead to diversity creating superior performance. This includes that the task should be difficult enough that it needs more than a single brilliant problem solver (i.e., task complexity), that those in the group should have skills relevant to the problem (i.e., type of diversity), that there is synergy and sharing among the group members (i.e., team dispersion), and that the group should be large and genuinely diverse (i.e., team size). A clear connection between theory, data, and analysis is a hallmark of a great paper, reflected in that the more a meta-analysis attempts to test an existing theory, the larger the number of citations it receives (Aguinis, Dalton, et al., 2011).

However, the techniques that Stahl et al. used to assess moderators have evolved considerably. Stahl et al. used subgrouping methodology, which comes in two different forms: comparison of mean effect sizes and analysis of variance (Borenstein et al., 2009). The use of such subgrouping approach has come into debate. To begin with, subgrouping should be reserved for categorical variables as otherwise it requires dichotomizing continuous moderators, usually using a median split, which reduces statistical power (Cohen, 1983; Steel & Kammeyer-Mueller, 2002). Also, it appears that

Stahl et al. used a fixed-effects model although meta-analytic comparisons are typically based on a random-effects model (Aguinis, Sturman, & Pierce, 2008), with some exceptions, such as when the subgroups are considered exhaustive (e.g., before and after a publication year; Borenstein & Higgins, 2013) or whether the research question focuses on dependent correlates differing within the same situation (Cheung & Chan, 2004). Furthermore, standard Wald-type comparisons result in massive increases in Type I errors (Gonzalez-Mulé & Aguinis, 2018), and, although useful to contrast two sets of correlations to determine whether they differ, they have limited application in determining moderators' explanatory power (Lubinski & Humphreys, 1996). A superior alternative to subgrouping is meta-regression analysis or MARA (Aguinis, Gottfredson, & Wright, 2011; Gonzalez-Mulé & Aguinis, 2018; Viechtbauer et al., 2015). Essentially, MARA is a regression model in which effect sizes are the dependent variable and the moderators are the predictors (e.g., Steel & Kammeyer-Mueller, 2002). MARA tests whether the size of the effects can be predicted by fluctuations in the values of the hypothesized moderators, which therefore are conceptualized as boundary conditions for the size of the effect. If there are enough studies, MARA enables simultaneous testing of several moderators. Evaluating the weighting options for the predictors, Viechtbauer et al. settled on the Hartung–Knapp as the best alternative. Other recommendations for MARA are given by Gonzalez-Mulé and Aguinis (2018), such as making the sensible observation that we should use R^2_{Meta}, which adjusts R^2 to reflect I^2, the known variance after excluding sampling error. Gonzalez-Mulé and Aguinis (2018) also included the R code to conduct all analyses as well as an illustrative study. Some analysis programs, such as *metafor*, provide R^2_{Meta} by default.

Stage 4: Reporting

A modern meta-analysis must be transparent and reproducible – meaning that all steps and procedures need to be described in such a way that a different team of researchers would obtain similar results with the same data. At present, this is among our greatest challenges. In psychology, half of 500 effect sizes sampled from 33 meta-analyses were not reproducible based on the available information (Maassen, van Assen, Nuijten, Olsson-Collentine, & Wicherts, 2020). Also, a modern meta-analysis not only provides more than a summary of past findings but also points towards the next steps. Consequently, it should consider future research directions, not just in terms of what studies should be conducted but when subsequent meta-analyses could be beneficial and what they should address.

Transparency and reproducibility

As Hohn, Slaney and Tafreshi (2020: 207) concluded: "It is vitally important that meta-analytic work be reproducible, transparent, and able to be subjected to rigorous scrutiny so as to ensure that the validity of conclusions of any given question may be corroborated when necessary." Stahl et al. provided their database to assist with our review, allowing the assessment of reproducibility because both of our analyses relied on the same meta-analytic data (Jasny, Chin, Chong, & Vignieri, 2011). Such responsiveness is commendable but also highlights the problem of using researchers' personal computers as archives. The data are often difficult to obtain, lost, or incomplete, and even authors of recent meta-analyses, who claim that the references or data are available upon request, and such availability is an explicit requirement for many journals, are sporadically responsive (Wood, Müller, & Brown, 2018). Hence the call for Open Science, Open Data, Open Access, and Open Archive, and the increasing number of journals that have adopted this standard of transparency (Aguinis, Banks, Rogelberg, & Cascio, 2020; Vicente-Sáez & Martínez-Fuentes, 2018). Along with the complete database, if the statistical process deviates from standard practice, ideally a copy of the analysis script should be made available in an Open Science archive. The advantages of such heightened transparency and reproducibility are several (Aguinis et al., 2018; Polanin, Hennessy, & Tsuji, 2020b), but it does introduce considerable challenges (Beugelsdijk et al., 2020).

To begin with, journal articles are an abridged version of the available data and the analysis process. By themselves, they can hide a multitude of virtues and vices. As per Stahl et al., we were unable to completely recreate some steps (though we did approximate them) as they were not sufficiently specified. Adopting an Open Science framework, choices can be examined and updated, improving the research quality, as it encourages increased vigilance by the source authors.

As Marshall and Wallace (2019: 1) concluded, "Clearly, existing processes are not sustainable: reviews of current evidence cannot be produced efficiently and, in any case, often go out of date quickly once they are published. The fundamental problem is that current EBM [evidence-based medicine] methods, while rigorous, simply do not scale to meet the demands imposed by the voluminous scale of the (unstructured) evidence base." Although originating from the medical field, this critique equally applies to management and IB (Rousseau, 2020). Our traditional methods of reporting, which Stahl et al. adopted, are flagging the extracted studies with an asterisk in the reference section or upon request from the authors. This is at present insufficient. Science is a social endeavor, and we need to be able to build on past meta-analyses to enable future ones; by making meta-analyses reproducible, that is, in having access to the coding database we are also making the process cumulative (Polanin et al., 2020b). In fact, Open Science can be considered as a stepping stone towards living systematic reviews (LSRs; Elliot et al., 2017), essentially reviews that are continuously updated in real time. Having found traction in medicine, LSRs are based around critical topics that can enable broad collaborations (along with advances in technological innovations, such as online platforms and machine learning), although not without their own challenges (Millard, Synnot, Elliott, Green, McDonald, & Turner, 2019).

Such data sharing is not without its perils, exacerbating the moral hazards associated with a common pool resource, that is, the publication base (Alter & Gonzalez, 2018; Hess & Ostrom, 2003). Traditionally, in a meta-analysis, the information becomes "consumed" once published or "extracted" in a meta-analysis, and the research base needs time to "regenerate," that is grow sufficiently that a new summary is justified. Since there is no definitive point when regeneration occurs, we encounter a tragedy of the commons, where one instrumental strategy is to rush marginal meta-analyses to the academic market, shopping them to multiple venues in search of acceptance (i.e., science's first mover advantage; Newman 2009). Open Science is likely to exacerbate this practice, as the cost of updating meta-analyses would be substantially reduced and, as Beugelsdijk et al. (2020: 897) discussed, "There would be nothing to stop others from using the fruits of their labor to write a competing article". For example, in the field of ecology, the authors of a meta-analysis on marine habitats admirably provided their complete database, which was rapidly re-analyzed by a subsequent group with a slightly different taxonomy (Kinlock et al., 2019). In a charitable reply, they viewed this as an endorsement of Open Science, concluding "Without transparent methods, explicitly defined models, and fully transparent data and code, this advancement in scientific knowledge would have been delayed if not unobtainable" (Kinlock et al., 2019: 1533). However, as they noted, it took a team of ten authors over two years to create the original database, and posting it allowed others to supersede them with relatively minimal effort. If the original authors adopt an Open Science philosophy for their meta-analytic database (which we strongly recommend), subsequent free-riding or predatory authors could take advantage and, by adding marginal updates, publish. Reviewers should be sensitive to whether a new meta-analysis provides a substantive threshold of contribution, preferably with the involvement of the previous lead authors upon whose work they are building (especially if recent). To help guide decisions, we further address this issue in our subsequent section, "The next generation of meta-analyses." In addition, journals can help to mitigate the moral hazard associated with meta-analysis' common pool resource by allowing pre-registration and conditional pre-approval of large meta-analyses.

Future research directions

A good section on future research directions," based on a close study of the entire field's findings, although perhaps sporadically used (Carlson & Ji, 2011), can be as invaluable as the core results themselves. This information allows meta-analysts to steer the field itself. We can expect meta-analysts to expound on the gap between what is already known and what is required to move forward. The components of a good Future Research section touch on many of the very stages we previously emphasized here, especially Data Collection, Data Analysis, and Reporting.

During Data Collection, researchers have had to be sensitive to inclusion and exclusion criteria and how constructs were defined and measured. This provides several insights. To begin with, the development of inclusion and exclusion criteria, along with addressing issues of commensurability, allow researchers to consider construct definition and its measurement. Was the construct well defined? Often, there are as many definitions as there are researchers, so this is an opportunity to provide some clarity. With an enhanced understanding, an evaluation of the measures can proceed, especially where they could be improved. How well do they assess the construct? Should some be favored and others abandoned?

During Data Analysis, researchers likely attempted to assemble a correlation matrix to conduct meta-analytic structural equation modeling and meta-regression. One of the more frustrating aspects of this endeavor is when the matrix is almost complete, but some cells are missing. Here is where the researcher can direct future projects towards understudied elements, as well as highlight that other relationships have been overly emphasized, perhaps to the point of recommending a moratorium. Similarly, the issues of heterogeneity and moderators come up. The results may generalize, but this may be due to overly homogenous samples or settings. Also, there was likely a need by some moderators to address theory, but the field simply did not report or contain them. Additionally, informing reviewers that the field is not yet able to address such ambitions often helps curtail a critique of their absence. This is where Stahl et al. primarily dedicated their own Future Research Agenda: process moderators should be considered (alone and in combination) and different cultural settings should be explored. In short, the researcher should stress how every future study should contextualize or describe itself (i.e., based on the likely major moderators).

Finally, we emphasized during Reporting the need for an Open Science framework. For a meta-analyst, often the greatest challenge is not the choice of statistical technique but getting enough foundational studies, especially those that fully report and are of high quality. The methodological techniques tend to converge at higher *k*, and statistical legerdemain can mitigate but not overcome an inherent lack of data. Fortunately, the Open Science movement and the increased availability of a study's underlying data (i.e., IPD) opens possibilities. Contextual and other detailed information may not be reported in a study, often due to journal space limitation, but are needed for meta-analytic moderator analyses. With Open Science, this information will be increasingly available, allowing for the improved application of many sophisticated techniques. For example, Jak and Cheung's (2020) one-stage MASEM incorporates continuous moderators for MARA but requires a minimum of 30 studies. Consequently, researchers should consider what new findings would be possible with a growing research base. In short, journal editors and reviewers should expect a synopsis of *when* a follow up meta-analysis would be appropriate and *what* the next update could accomplish with a greater and more varied database to rely on.

THE NEXT GENERATION OF META-ANALYSES

Is Stahl et al. the last word on diversity? Of course not. The entire point of Stahl et al.'s future research direction section was that it should be acted upon. Since Stahl et al., there have been a variety of advances in diversity research, such as the greater adoption of Blau's index used to calculate the actual proportion of diversity (Blau, 1977; Harrison & Klein, 2007), and Shemla et al.'s (2016) conclusions that perceived levels of diversity can be more revealing than the objective measures on which Stahl et al. focused. Furthermore, not only do research bases refine and grow, at times exponentially, but meta-analytical methodology continues to evolve. With the increased popularity of meta-analysis, we can expect continued technical refinements and advances, some of which we touched upon in our article. We have shown that some of the newer techniques affected Stahl et al. findings, which proved sensitive to outliers and whether a fixed- or random-effects model was used. As for the near future, Marshall and Wallace (2019), as well as Johnson, Bauer and Niederman (2017), argued that we will see increased adoption of machine-learning systems in literature search and screening, which already exist but tend to be in the domain of well-funded health topics such as immunization (Begert, Granek, Irwin, & Brogly, 2020). Machine learning is a response to the "torrential volume of unstructured published evidence has rendered existing (rigorous, but manual) approaches to evidence synthesis increasingly costly and impractical" (Johnson et al., 2017: 8). The typical machine-learning strategy is to constantly sort the remaining articles based on researchers' previous choices, until these researchers reject (screen out) a substantive number of articles in a row, whereupon screening stops. Since the system cannot predict perfectly, there is a tradeoff between false negatives and positives, meaning that adopters will sacrifice missing approximately 4–5% of relevant articles to reduce screening time by 30–78% (Créquit, Boutron, Meerpohl, Williams, Craig, & Ravaud, 2020). Complementing these efforts, meta-analyses may draw on a variant of the "Mark–Recapture" method commonly used in ecology to determine a population's size. Essentially, such as determining the number of fish in a pond, some are captured, marked, and released. The number of these marked fish re-captured during a subsequent effort provides the total population through the Lincoln–Petersen method. As this applies to meta-analysis, when one has a variety of terms and databases to search for a construct, subsequent searches showing an ever-increasing number of duplicate articles (i.e., articles previously "marked" and "recaptured") provides a strong indicator of thoroughness. This combination of research base growth and improved search and analysis means that meta-analyses *should* have a half-life and perhaps a short one (Shojania, Sampson, Ansari Ji, Doucette, & Moher, 2007).

Despite these ongoing advances, it is not uncommon for IB, management, and related fields to rely on meta-analyses not just one decade old but two, three or four, which can be contrasted with the Cochrane Database of Systematic Reviews where the median time for an update is approximately 3 years (Bashir, Surian, & Dunn, 2018; Bastian, Doust, Clarke, & Glasziou, 2019). For example, the classic meta-analysis on job satisfaction by Judge, Heller and Mount (2002) is still considered foundational and cited hundreds of times each year, although it relies on an unpublished personality matrix from the early 1980s, a choice of matrix that, as Park et al. (2020: 25) noted, "can substantively alter their conclusions". Because of this reliance on very early and very rough estimates, newer meta-analyses indicate that many of Judge et al.'s core findings do not replicate (Steel, Schmidt, Bosco, & Uggerslev, 2019). Exactly because techniques evolve and research bases continue to grow, it is critical to update meta-analyses, even those, or perhaps especially those, that have become classics in a field.

This issue of meta-analytic currency has been intensely debated, culminating in a two-day international workshop by the Cochrane Collaboration's Panel for Updating Guidance for Systematic Review (Garner et al., 2016). Drawing on this panel's work, as well as similar recommendations by Mendes, Wohlin, Felizardo and Kalinowski (2020), we provide a revised set of guidelines, summarized in Figure 1. Next, we apply this sequence of steps to Stahl et al. *Step one* is the consideration of currency. Does the review still address a relevant question? In the case of Stahl et al., its topic has increased in relevance, as reflected by its frequent citations and the widespread concern with diversity. *Step two* is to the

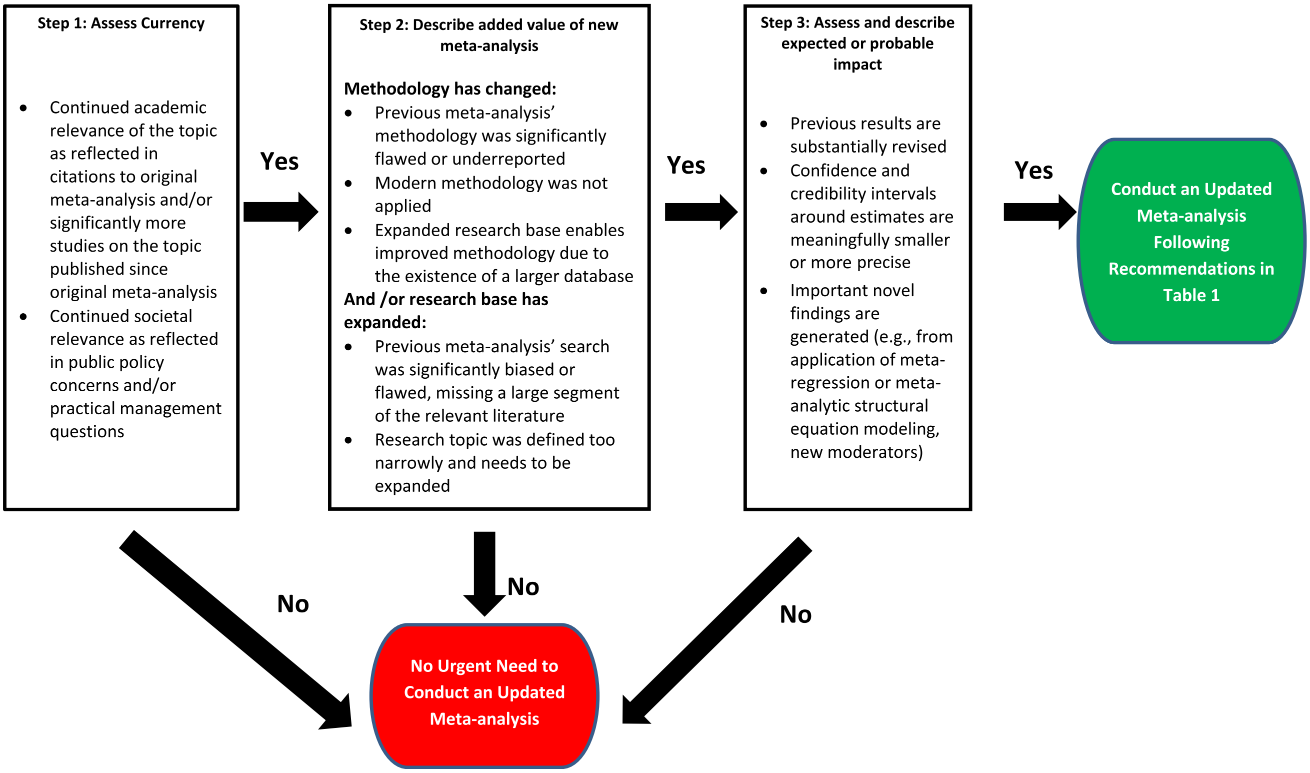
Step 1: Assess Currency
• Continued academic relevance of the topic as reflected in citations to original meta-analysis and/or significantly more studies on the topic published since original meta-analysis
• Continued societal relevance as reflected in public policy concerns and/or practical management questions
Yes
Step 2: Describe added value of new meta-analysis
Methodology has changed:
• Previous meta-analysis' methodology was significantly flawed or underreported
• Modern methodology was not applied
• Expanded research base enables improved methodology due to the existence of a larger database
And /or research base has expanded:
• Previous meta-analysis' search was significantly biased or flawed, missing a large segment of the relevant literature
• Research topic was defined too narrowly and needs to be expanded
Yes
Step 3: Assess and describe expected or probable impact
• Previous results are substantially revised
• Confidence and credibility intervals around estimates are meaningfully smaller or more precise
• Important novel findings are generated (e.g., from application of meta-regression or meta-analytic structural equation modeling, new moderators)
Yes
Conduct an Updated Meta-analysis Following Recommendations in Table 1
No
No
No
No Urgent Need to Conduct an Updated Meta-analysis

Figure 1 Decision framework to determine the need to update a meta-analysis.

consideration of methodology and/or the research base. Have any new relevant methods been developed? Did Stahl et al. miss any appropriate applications? Meta-analysis has indeed rapidly developed, and, as we review here, there are numerous refinements that could be applied, from outlier analysis to MASEM. Alternatively, have any new relevant studies been published, or new information? This is related to currency, *Step one*, as, within the thousand studies alone that cited Stahl et al., the meta-analytic database would likely double or triple. Also, an expanded research base enables the application of more sophisticated analysis techniques. *Step three* is the probable impact of the new methodology and/or studies. Can they be expected to change the findings or reduce uncertainty? This has already been shown here, that taking a random-effects approach has changed statistical significance. Of note, there merely needs to be a likelihood of impact, not an inevitability. For example, narrowing extremely wide confidence intervals without changing the average effect size is still a valuable contribution simply because it reduces uncertainty. Similarly, providing a previously unavailable complete meta-analytic database in an Open Science archive (enabling cumulative growth) can still be considered impactful (especially as it motivates all researchers to data-share or risk their meta-analysis being rapidly superseded).

By all standards, Stahl et al.'s meta-analysis is now worthy of updating, but, as mentioned, it is certainly not alone. Given that our Table 1 focuses on modern meta-analytic practices, it makes a useful litmus test in conjunction with Figure 1's decision framework for determining whether newer meta-analyses should be pursued or whether existing ones provide a sufficiently novel contribution. The more of the elements expounded in Table 1 that the more recent meta-analysis has compared to its predecessor, the more it deserves favorable treatment.

CONCLUSIONS

We have discussed key methodological junctures in the design and execution of a modern meta-analytic study. We have shown that each stage in a meta-analytical study requires a series of critical decisions. These decisions are critical because they have an impact on the results obtained and substantive conclusions for theory as well as implications for practice and policymaking. We have discussed Stahl et al.'s meta-analysis as an exemplar to explain why their article was selected as the 2020 *JIBS* decade award, but also to show how the field of meta-analysis has progressed since. Table 1 summarizes recommendations and their implementation guidelines for a modern meta-analysis. By following the different steps described in Table 1, we make explicit the anatomy of a successful meta-analysis, and we summarize what authors can be expected to do, what reviewers can be expected to ask for, and what consumers of meta-analytic reviews (i.e., other researchers, practitioners, and policymakers) can be expected to look for. Like any research method, meta-analysis is nuanced, and this is not an exhaustive list of all technical aspects or possible contributions or permutations. We can, though, summarize its spirit. When a phenomenon has been researched from a wide variety of perspectives, pulling these studies together and effectively exploring and explaining the shifting effect sizes and signs is invariably enriching.

ACKNOWLEDGEMENT

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

REFERENCES

Aguinis, H., Banks, G. C., Rogelberg, S., & Cascio, W. F. 2020. Actionable recommendations for narrowing the science-practice gap in open science. *Organizational Behavior and Human Decision Processes,* 158: 27–35.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. 2005. Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology,* 90(1): 94–107.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. 2011a. Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management,* 37(1): 5–38.

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., Gottfredson, R. K., & Wright, T. A. 2011b. Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior,* 32(8): 1033–1043.

Aguinis, H., Hill, N. S., & Bailey, J. R. 2021. Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods*. https://doi.org/10.1177/1094428119836485.

Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., & Dalton, C. M. 2011c. Debunking myths and urban legends about meta-analysis. *Organizational Research Methods,* 14(2): 306–331.

Aguinis, H., Ramani, R. S., & Alabduljader, N. 2018. What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals,* 12: 83–110.

Aguinis, H., Sturman, M. C., & Pierce, C. A. 2008. Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods,* 11(1): 9–34.

Alter, G., & Gonzalez, R. 2018. Responsible practices for data sharing. *American Psychologist,* 73(2): 146–156.

Augusteijn, H. E. M., van Aert, R. C. M., & van Assen, M. A. L. M. 2019. The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods,* 24(1): 116–134.

Baker, R., & Jackson, D. 2016. New models for describing outliers in meta-analysis. *Research Synthesis Methods,* 7(3): 314–328.

Bashir, R., Surian, D., & Dunn, A. G. 2018. Time-to-update of systematic reviews relative to the availability of new evidence. *Systematic Reviews,* 7(1): 195.

Bastian, H., Doust, J., Clarke, M., & Glasziou, P. 2019. *The epidemiology of systematic review updates: A longitudinal study of updating of Cochrane reviews, 2003 to 2018*. medRxiv: 19014134.

Becker, B. J. 2005. Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*: 111–125. West Sussex: Wiley.

Begert, D., Granek, J., Irwin, B., Brogly, C., & Xtract, A. I. 2020. Using automation for repetitive work involved in a systematic review. *CCDR,* 46(6): 174–179.

Begg, C. B. 1994. Publication bias. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis*: 399–409. New York: Russell Sage.

Bem, D. J. 1995. Writing a review article for psychological bulletin. *Psychological Bulletin,* 118(2): 172–177.

Bergh, D. D., Aguinis, H., Heavey, C., Ketchen, D. J., Boyd, B. K., Su, P., et al. 2016. Using meta-analytic structural equation modeling to advance strategic management research: Guidelines and an empirical illustration via the strategic leadership-performance relationship. *Strategic Management Journal,* 37(3): 477–497.

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

Beugelsdijk, S., Ambos, B., & Nell, P. 2018a. Conceptualizing and measuring distance in international business research: Recurring questions and best practice guidelines. *Journal of International Business Studies,* 49(9): 1113–1137.

Beugelsdijk, S., Kostova, T., Kunst, V. E., Spadafora, E., & van Essen, M. 2018b. Cultural distance and firm internationalization: A meta-analytical review and theoretical implications. *Journal of Management,* 44(1): 89–130.

Beugelsdijk, S., van Witteloostuijn, A., & Meyer, K. 2020. A new approach to data access and research transparency (DART). *Journal of International Business Studies,* 51(6): 887–905.

Blau, P. 1977. *Inequality and heterogeneity. A primitive theory of social structure*. New York: Free Press.

Boedhoe, P. S., Heymans, M. W., Schmaal, L., Abe, Y., Alonso, P., Ameis, S. H., et al. 2019. An empirical comparison of meta and mega-analysis with data from the ENIGMA obsessive-compulsive disorder working group. *Frontiers in Neuroinformatics,* 12: 102.

Booth, A. 2008. Unpacking your literature search toolbox: On search styles and tactics. *Health Information and Libraries Journal,* 25(4): 313–317.

Booth, A., Briscoe, S., & Wright, J. M. 2020. The "realist search": A systematic scoping review of current practice and reporting. *Research Synthesis Methods,* 11(1): 14–35.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. 2009. *Introduction to meta-analysis*. Chichester: Wiley.

Borenstein, M., & Higgins, J. P. T. 2013. Meta-analysis and subgroups. *Prevention Science,* 14(2): 134–143.

Bornmann, L., & Mutz, R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology,* 66(11): 2215–2222.

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

Bosco, F. A., Steel, P., Oswald, F. L., Uggerslev, K., & Field, J. G. 2015b. Cloud-based meta-analysis to bridge science and practice: Welcome to metaBUS. *Personnel Assessment and Decisions,* 1(1): 3–17.

Bowen, F. E., Rostami, M., & Steel, P. 2010. Timing is everything: A meta-analysis of the relationships between organizational performance and innovation. *Journal of Business Research,* 63(11): 1179–1185.

Brannick, M. T., Potter, S. M., Benitez, B., & Morris, S. B. 2019. Bias and precision of alternate estimators in meta-analysis: Benefits of blending Schmidt-Hunter and Hedges approaches. *Organizational Research Methods,* 22(2): 490–514.

Cady, S. H., & Valentine, J. 1999. Team innovation and perceptions of consideration: What difference does diversity make? *Small Group Research,* 30(6): 730–750.

Carlson, K. D., & Ji, F. X. 2011. Citing and building on meta-analytic findings: A review and recommendations. *Organizational Research Methods,* 14(4): 696–717.

Carson, K. P., Schriesheim, C. A., & Kinicki, A. J. 1990. The usefulness of the "fail-safe" statistic in meta-analysis. *Educational and Psychological Measurement,* 50(2): 233–243.

Cheung, M. W.-L. 2018. Issues in solving the problem of effect size heterogeneity in meta-analytic structural equation modeling: A commentary and simulation study on Yu, Downes, Carter, and O'Boyle (2016). *Journal of Applied Psychology,* 103(7): 787–803.

Cheung, M. W. L., & Chan, W. 2004. Testing dependent correlation coefficients via structural equation modeling. *Organizational Research Methods,* 7(2): 206–223.

Cohen, J. 1962. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,* 65: 145–153.

Cohen, J. 1983. The cost of dichotomization. *Applied Psychological Measurement,* 7(3): 249–253.

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

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.

Créquit, P., Boutron, I., Meerpohl, J., Williams, H., Craig, J., & Ravaud, P. 2020. Future of evidence ecosystem series: 2. Current opportunities and need for better tools and methods. *Journal of Clinical Epidemiology,* 123: 143–152.

Dahlke, J. A., & Wiernik, B. M. 2019. psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement,* 43(5): 415–416.

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

Davies, H. T. O., Nutley, S. M., & Smith, P. C. 1999. Editorial: What works? The role of evidence in public sector policy and practice. *Public Money and Management,* 19(1): 3–5.

Denyer, D., & Tranfield, D. 2008. Producing a systematic review. In D. Buchanan (Ed.), *The Sage handbook of organizational research methods*: 671–689. London: Sage.

DeSimone, J. A., Köhler, T., & Schoen, J. L. 2019. If it were only that easy: The use of meta-analytic research by organizational scholars. *Organizational Research Methods,* 22(4): 867–891.

Duval, S. J. 2005. The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments*: 127–144. Chichester: Wiley.

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., et al. 2017. Living systematic review: 1. Introduction – The why, what, when, and how. *Journal of Clinical Epidemiology,* 91: 23–30.

Ferguson, C. J., & Brannick, M. T. 2012. Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods,* 17(1): 120–128.

Fine, C., Sojo, V., & Lawford-Smith, H. 2020. Why does workplace gender diversity matter? Justice, organizational benefits, and policy. *Social Issues and Policy Review,* 14(1): 36–72.

Fortune 2017 (August 8). *Google's gender problem is actually a tech problem*. Retrieved from http://fortune.com/2017/08/08/google-gender-struggle-tech/. Retrieved from June 15, 2020.

Friese, M., & Frankenbach, J. 2020. p-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods,* 25(4): 456–471.

Fujimoto, Y., Härtel, C. E., & Azmat, F. 2013. Towards a diversity justice management model: Integrating organizational justice and diversity management. *Social Responsibility Journal,* 9(1): 148–166.

Garner, P., Hopewell, S., Chandler, J., MacLehose, H., Akl, E. A., Beyene, J., et al. 2016. When and how to update systematic reviews: Consensus and checklist. *British Medical Journal,* 354: i3507.

Gibson, C. B., & Gibbs, J. L. 2006. Unpacking the concept of virtuality: The effects of geographic dispersion, electronic dependence, dynamic structure, and national diversity on team innovation. *Administrative Science Quarterly,* 51(3): 451–495.

Gonzalez-Mulé, E., & Aguinis, H. 2018. Advancing theory by assessing boundary conditions with meta-regression: A critical review and best-practice recommendations. *Journal of Management,* 44: 2246–2273.

Gusenbauer, M., & Haddaway, N. R. 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods,* 11(2): 181–217.

Harari, M. B., Parola, H. R., Hartwell, C. J., & Riegelman, A. 2020. Literature searches in systematic reviews and meta-analyses: A review, evaluation, and recommendations. *Journal of Vocational Behavior,* 118: 103377.

Harrison, D., & Klein, K. 2007. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review,* 32(4): 1199–1228.

Harzing, A.-W. 2006. Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management,* 6(2): 243–264.

Havránek, T., Stanley, T. D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., et al. 2020. Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys,* 34: 469–475.

Hedges, L. V. 1982. Estimation of effect sizes from a series of experiments. *Psychological Bulletin,* 92: 490–499.

Hedges, L. V., & Olkin, I. 1985. *Statistical methods for meta-analysis*. Orlando: Academic.

Henmi, M., & Copas, J. B. 2010. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine,* 29(29): 2969–2983.

Hess, C., & Ostrom, E. 2003. Ideas, artifacts, and facilities: Information as a common-pool resource. *Law and Contemporary Problems,* 66(1/2): 111–145.

Hohn, R. E., Slaney, K. L., & Tafreshi, D. 2020. An empirical review of research and reporting practices in psychological meta-analyses. *Review of General Psychology,* 24(3): 195–209.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. 1982. *Meta-analysis: Cumulative research findings across studies*. Beverly Hills: Sage.

Hunter, J. E., Schmidt, F. L., & Le, H. 2006. Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology,* 91(3): 594–612.

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

Ioannidis, J. P., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. 2014. Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences,* 18(5): 235–241.

Jackson, D., Turner, R., Rhodes, K., & Viechtbauer, W. 2014. Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC Medical Research Methodology,* 14(1): 103.

Jak, S., & Cheung, M. W. L. 2020. Meta-analytic structural equation modeling with moderating effects on SEM parameters. *Psychological Methods,* 25(4): 430–455.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. 2011. Again, and again, and again…. *Science,* 334: 1225.

Johnson, B. T., & Hennessy, E. A. 2019. Systematic reviews and meta-analyses in the health sciences: Best practice methods for research syntheses. *Social Science and Medicine,* 233: 237–251.

Johnson, C. D., Bauer, B. C., & Niederman, F. 2017. The automation of management and business science. *Academy of Management Perspectives*. https://doi.org/10.5465/amp.2017.0159.

Judge, T. A., Heller, D., & Mount, M. K. 2002. Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology,* 87(3): 530–541.

Kaufmann, E., Reips, U. D., & Maag Merki, K. 2016. Avoiding methodological biases in meta-analysis: Use of online versus offline individual participant data (IPD) in educational psychology. *Zeitschrift für Psychologie,* 224(3): 157–167.

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. 2012. Publication bias in the organizational sciences. *Organizational Research Methods,* 15(4): 624–662.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. 2013. Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-analytic Reporting Standards). *Journal of Business and Psychology,* 28(2): 123–143.

Kinlock, N. L., Prowant, L., Herstoff, E. M., Foley, C. M., Akin-Fajiye, M., Bender, N., et al. 2019. Open science and meta-analysis allow for rapid advances in ecology: A response to Menegotto et al. (2019). *Global Ecology and Biogeography,* 28(10): 1533–1534.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., et al. 2018. Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science,* 1(4): 443–490.

Kostova, T., Beugelsdijk, S., Scott, W. R., Kunst, V., Chua, C. H., & van Essen, M. 2020. The construct of institutional distance through the lens of different institutional perspectives: Review, analysis and recommendations. *Journal of International Business Studies,* 51(4): 467–497.

Kostova, T., Roth, K., & Dacin, T. 2008. Institutional theory in the study of multinational corporations: A critique and new directions. *Academy of Management Review,* 33(4): 994–1006.

Landis, R. S. 2013. Successfully combining meta-analysis and structural equation modeling: Recommendations and strategies. *Journal of Business and Psychology,* 28(3): 251–261.

Larsen, K. R., & Bong, C. H. 2016. A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Quarterly,* 40(3): 529–551.

Larsen, K. R., Hekler, E. B., Paul, M. J., & Gibson, B. S. 2020. Improving usability of social and behavioral sciences' evidence: A call to action for a national infrastructure project for mining our knowledge. *Communications of the Association for Information Systems,* 46(1): 1.

LeBreton, J. M., & Senter, J. L. 2008. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods,* 11(4): 815–852.

Lee, C. I., Bosco, F. A., Steel, P., & Uggerslev, K. L. 2017. A metaBUS-enabled meta-analysis of career satisfaction. *Career Development International,* 22(5): 565–582.

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. 2018. Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Research Synthesis Methods,* 9(3): 336–351.

Lubinski, D., & Humphreys, L. 1996. Seeing the forest from the trees: When predicting the behavior or status of groups, correlate means. *Psychology, Public Policy, and Law,* 2: 363–376.

Lyness, K. S., & Brumit Kropf, M. 2007. Cultural values and potential nonresponse bias. *Organizational Research Methods,* 10(2): 210–224.

Ma, H. H. 2009. The effect size of variables associated with creativity: A meta-analysis. *Creativity Research Journal,* 21(1): 30–42.

Maassen, E., van Assen, M. A., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. 2020. Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS ONE,* 15(5): e0233107.

Marshall, I. J., & Wallace, B. C. 2019. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews,* 8(1): 163.

Maseland, R., Dow, D., & Steel, P. 2018. The Kogut and Singh national cultural distance index: Time to start using it as a springboard rather than a crutch. *Journal of International Business Studies,* 49(9): 1154–1166.

Maznevski, M. L. 1995. *Process and performance in multicultural teams*, Working Paper, University of Virginia, Charlottesville, VA.

Maznevski, M. L., Davison, S. C., & Jonsen, K. 2006. Global virtual team dynamics and effectiveness. In G. K. Stahl & I. Bjorkman (Eds.), *Handbook of research in international human resource management*: 364–384. Cheltenham: Edward Elgar.

Meehl, P. E. 1990. Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry,* 1(2): 108–141.

Mendes, E., Wohlin, C., Felizardo, K., & Kalinowski, M. 2020. When to update systematic literature reviews in software engineering. *Journal of Systems and Software,* 167: 110607.

Merton, R. K. 1968. The Matthew effect in science: The reward and communication systems of science are considered. *Science,* 159(3810): 56–63.

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

Millard, T., Synnot, A., Elliott, J., Green, S., McDonald, S., & Turner, T. 2019. Feasibility and acceptability of living systematic reviews: Results from a mixed-methods evaluation. *Systematic Reviews,* 8(1): 325.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine,* 151(4): 264–269.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., et al. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews,* 4(1): 1.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., et al. 2017. A manifesto for reproducible science. *Nature Human Behaviour,* 1(1): 1–9.

Newman, M. E. 2009. The first-mover advantage in scientific publication. *Europhysics Letters,* 86(6): 68001.

Oh, I. S. 2020. Beyond meta-analysis: Secondary uses of meta-analytic data. *Annual Review of Organizational Psychology and Organizational Behavior,* 7: 125–153.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. 2017. Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management. *Human Resource Management Review,* 27(1): 201–215.

Page, S. E. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton: Princeton University Press.

Page, M. J., et al. 2020. *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*. https://doi.org/10.31222/osf.io/v7gm2.

Paletz, S. B., Peng, K., Erez, M., & Maslach, C. 2004. Ethnic composition and its differential impact on group processes in diverse teams. *Small Group Research,* 35(2): 128–157.

Park, H. H., Wiernik, B. M., Oh, I.-S., Gonzalez-Mulé, E., Ones, D. S., & Lee, Y. 2020. Meta-analytic five-factor model personality intercorrelations: Eeny, meeny, miney, moe, how, which, why, and where to go. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000476.

Pastor, D. A., & Lazowski, R. A. 2018. On the multilevel nature of meta-analysis: A tutorial, comparison of software programs, and discussion of analytic choices. *Multivariate Behavioral Research,* 53(1): 74–89.

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. 2016. An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. *Journal of Leadership and Organizational Studies,* 23(1): 66–81.

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

Polanin, J. R., Espelage, D. L., Grotpeter, J. K., Valido, A., Ingram, K. M., Torgal, C., et al. 2020a. Locating unregistered and unreported data for use in a social science systematic review and meta-analysis. *Systematic Reviews,* 9: 1–9.

Polanin, J. R., Hennessy, E. A., & Tsuji, S. 2020b. Transparency and reproducibility of meta-analysis in psychology: A meta-review. *Perspectives on Psychological Science,* 15(4): 1026–1041.

Polzer, J. T., Crisp, C. B., Jarvenpaa, S. L., & Kim, J. W. 2006. Extending the faultline model to geographically dispersed teams: How colocated subgroups can impair group functioning. *Academy of Management Journal,* 49(4): 679–692.

Possolo, A., Merkatas, C., & Bodnar, O. 2019. Asymmetrical uncertainties. *Metrologia,* 56(4): 045009.

Revelle, W., & Wilt, J. 2019. Analyzing dynamic data: A tutorial. *Personality and Individual Differences,* 136: 38–51.

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.

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

Rosenthal, R., & Rubin, D. B. 1982. Comparing effect sizes of independent studies. *Psychological Bulletin,* 92: 500–504.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. 2005. *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester: Wiley.

Rousseau, D. 2020. The realist rationality of evidence-based management. *Academy of Management Learning and Education,* 19(3): 415–423.

Scargle, J. D. 2000. Publication bias: The “File Drawer” problem in scientific inference. *Journal of Scientific Exploration,* 14: 91–106.

Schmidt, F. L., & Hunter, J. E. 2015. *Methods of meta-analysis* (3rd ed.). Thousand Oaks: Sage.

Scott, R. W. 2014. *Institutions and organizations* (4th ed.). Thousand Oaks: Sage.

Shemla, M., Meyer, B., Greer, L., & Jehn, K. A. 2016. A review of perceived diversity in teams: Does how members perceive their team’s composition affect team processes and outcomes? *Journal of Organizational Behavior,* 37: 89–106.

Sheng, Z., Kong, W., Cortina, J. M., & Hou, S. 2016. Analyzing matrices of meta-analytic correlations: Current practices and recommendations. *Research Synthesis Methods,* 7(2): 187–208.

Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., & Moher, D. 2007. How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine,* 147: 224–233.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. 2019. How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology,* 70: 747–770.

Smith, P. B., & Fischer, R. 2008. Acquiescence, extreme response bias and culture: A multilevel analysis. In F. J. R. V. de Vijver, D. A. van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures*: 285–314. New York: Taylor & Francis/Lawrence Erlbaum.

Spellman, B. A. 2015. A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science,* 10(6): 886–889.

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.

Stanek, K. C., & Ones, D. S. 2018. Taxonomies and compendia of cognitive ability and personality constructs and measures relevant to industrial, work and organizational psychology. In D. S. Ones, N. Anderson, C. Viswesvaran, & H. K. Sinangil (Eds.), *The SAGE handbook of industrial, work and organizational psychology* (2nd ed., Vol. 1): 366–407. Thousand Oaks: Sage.

Stanley, T. D. 2017. Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science,* 8(5): 581–591.

Stanley, T. D., & Doucouliagos, H. 2014. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods,* 5(1): 60–78.

Steel, P., Johnson, J. W., Jeanneret, P. R., Scherbaum, C. A., Hoffman, C. C., & Foster, J. 2010. At sea with synthetic validity. *Industrial and Organizational Psychology,* 3(3): 371–383.

Steel, P. D., & Kammeyer-Mueller, J. D. 2002. Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology,* 87(1): 96–111.

Steel, P., Kammeyer-Mueller, J., & Paterson, T. A. 2015. Improving the meta-analytic assessment of effect size variance with an informed Bayesian prior. *Journal of Management,* 41(2): 718–743.

Steel, P., Schmidt, J., Bosco, F., & Uggerslev, K. 2019. The effects of personality on job satisfaction and life satisfaction: A meta-analytic investigation accounting for bandwidth–fidelity and commensurability. *Human Relations,* 72(2): 217–247.

Sun, S. 2011. Meta-analysis of Cohen’s kappa. *Health Services and Outcomes Research Methodology,* 11(3–4): 145–163.

Sutton, A. J. 2009. Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*: 435–452. New York: Russell Sage.

Tabachnick, B. G., & Fidell, L. S. 2014. *Using multivariate statistics*. Harlow: Pearson.

Tanner-Smith, E. E., & Tipton, E. 2014. Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods,* 5(1): 13–30.

Taras, V., Kirkman, B. L., & Steel, P. 2010. Examining the impact of culture’s consequences: A three-decade, multilevel, meta-analytic review of Hofstede’s cultural value dimensions. *Journal of Applied Psychology,* 95(3): 405–439.

Taras, V., Rowney, J., & Steel, P. 2009. Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management,* 15(4): 357–373.

Taras, V., & Steel, P. 2009. Beyond Hofstede: Challenging the ten commandments of cross-cultural research. In C. Nakata (Ed.), *Beyond Hofstede: Culture frameworks for global marketing and management*: 40–60. Chicago: Palgrave Macmillan.

Tasheva, S., & Hillman, A. J. 2019. Integrating diversity at different levels: Multilevel human capital, social capital, and demographic diversity and their implications for team effectiveness. *Academy of Management Review,* 44(4): 746–765.

Tranfield, D., Denyer, D., & Smart, P. 2003. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management,* 14(3): 207–222.

Tsuji, S., Bergmann, C., & Cristia, A. 2014. Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science,* 9(6): 661–665.

Versteeg, M., & Ginsburg, T. 2017. Measuring the rule of law: A comparison of indicators. *Law and Social Inquiry,* 42(1): 100–137.

Vicente-Sáez, R., & Martínez-Fuentes, C. 2018. Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research,* 88: 428–436.

Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software,* 36(3): 1–48.

Viechtbauer, W., Lopez-Lopez, J. A., Sanchez-Meca, J., & Marin-Martinez, F. 2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods,* 20: 360–374.

Wasserman, J. D., & Bracken, B. A. 2003. Psychometric characteristics of assessment procedures. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology*: 43–66. New Jersey: Wiley.

Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., et al. 2017. What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. *American Psychologist,* 72(2): 79–117.

Wiernik, B. M., & Dahlke, J. A. 2020. Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science,* 3(1): 94–123.

Wood, B. D., Müller, R., & Brown, A. N. 2018. Push button replication: Is impact evaluation evidence for international development verifiable? *PLoS ONE,* 13(12): e0209416.

Yuan, Z., Morgeson, F. P., & LeBreton, J. M. 2020. Maybe not so independent after all: The possibility, prevalence, and consequences of violating the independence assumptions in psychometric meta-analysis. *Personnel Psychology,* 73(3): 491–516.

ABOUT THE AUTHORS

Piers Steel is a Professor in the Organizational Behaviour and Human Resource department at the Haskayne School of Business, University of Calgary, where he holds the Brookfield Management Research Chair. He received his Ph.D. from the University of Minnesota in I/O Psychology and is a Fellow of SIOP, APA and APS. His research focuses on meta-analysis, integrating theories of motivation, international culture, and determining the national impact of HR practices (e.g., personnel selection).

Sjoerd Beugelsdijk is a Professor of International Business at the University of Groningen, the Netherlands. He earned his Ph.D. at Tilburg University and explores how cultural diversity affects international business. He is a Fellow of the AIB and currently serves as Reviewing Editor for *JIBS* (2016–2022).

Herman Aguinis (Ph.D., 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 research methods. He has published nine books and more than 170 journal articles and is serving as Academy of Management President Elect, President, and Past President during 2020–2023. The 2020, 2019, and 2018 Web of Science Highly Cited Researchers Reports ranked him among the world's 100 most impactful researchers in Economics and Business.

Accepted by Alain Verbeke, Editor-in-Chief, 20 October 2020. This article was single-blind reviewed.