Sartal, Antonio; González-Loureiro, Miguel; Vázquez, Xosé H.

Article

Meta-analyses in management: What can we learn from clinical research?

BRQ Business Research Quarterly

Provided in Cooperation with:
Asociación Científica de Economía y Dirección de Empresas (ACEDE), Madrid

Suggested Citation: Sartal, Antonio; González-Loureiro, Miguel; Vázquez, Xosé H. (2021) : Meta-analyses in management: What can we learn from clinical research?, BRQ Business Research Quarterly, ISSN 2340-9436, Sage Publishing, London, Vol. 24, Iss. 1, pp. 91-111, https://doi.org/10.1177/2340944420916310

Methodological Insight

Business Research Quarterly
2021, Vol. 24(1) 91–111

DOI: 10.1177/2340944420916310
journals.sagepub.com/home/brq
SAGE

Meta-analyses in management: What can we learn from clinical research?

Antonio Sartal[1], Miguel González-Loureiro[1,2] and Xosé H. Vázquez[1]

Abstract

We analyze the weaknesses of meta-analyses (MAs) in management research using as benchmark a scientific field where this technique shows a longer tradition: clinical research. We suggest four areas in which management research MA practices should improve: (1) availability of information and replicability of primary research, (2) correct application of statistical support, (3) execution of heterogeneity analyses, and (4) standardization of result reporting. Using a representative MA on an operations management topic, we identify qualitatively the aspects to be improved at each stage. We show the different results that could have been achieved by following standard procedures in clinical research, incorporating different "good practices" from this research field. Overall, these recommendations aim at improving the transparency and replicability of MAs, which can not only facilitate the accumulation of scientific knowledge but also intensify the dialogue between academia and practitioners.

JEL CLASSIFICATION: C01; M11; M1

Keywords

Meta-analysis, "good practices," transparency, replicability, management, clinical research

Introduction

Meta-analyses (MAs) began to spread in the mid-20th century to integrate and synthesize the results of an increasing number of studies in areas such as psychology and epidemiology (Aguinis et al., 2011c; Geyskens et al., 2009). Essentially, MAs aim at synthesizing the effect of interest by aggregating the estimations of a number of primary studies to estimate a global effect size (Cooper & Hedges, 1994a; Hunter & Schmidt, 1990). This effect size can be correlations (linear associations between pairs of variables), mean differences (between groups), or ratios (hazard ratios, risk ratios, or odds ratios of an event occurrence in a number of groups).

Its subsequent development in other areas, such as management, was the result of the gradual deployment of scientifically mature research practices; however, the very same development of MAs in these fields ran parallel to an increasing number of contradictory results that often hampered scientific progress (Carlson & Ji, 2011). Thus, while the growth of MAs in clinical research has been spectacular, their dissemination in the field of management has not progressed to the same extent (Aguinis et al., 2011a; Grand et al., 2018). In fact, predictions based on a review of the main management journals, which forecasted more than 1,000 MA studies in 2015, have been overly optimistic (Aguinis et al., 2011c): whereas the authors identified 196

[1]University of Vigo, Vigo, Spain
[2]CIICESI, ESTG, Politécnico do Porto, Portugal

Corresponding author:
Antonio Sartal, School of Economics and Business, University of Vigo, Rúa Leonardo Da Vinci, s/n Campus das Lagoas/Marcosende, Vigo, 36310, Spain.
Email: antoniosartal@uvigo.es

studies in their 1982–2009 literature review, only 186 studies were published from 2010 to 2017 (10% of the initial forecast). A number of voices have raised the issue of how management scholars can benefit from the advances in areas such as clinical science, and there are open calls to adapt and adopt those good practices (e.g., Croucher, 2019; Kepes et al., 2012).

The reasons for the underutilization of this methodology in management research have barely been addressed. Indeed, some academics have analyzed the statistical limitations of certain approaches to MAs and have also demonstrated how the results can differ radically depending on the method (Aguinis et al., 2008) or the effect size metrics (e.g., Hedge's g vs. Cohen's q for mean differences). The frequency of use of two of the main statistical approaches—namely, Hunter and Schmidt (H&S; 1990, 2004) and Cooper and Hedges (C&H; 1994a and its subsequent updates, such as Cooper et al., 2009)—is one of the most striking aspects of the comparison between clinical research and management research MAs. While both methods (H&S and C&H) should not be seen as a dichotomous choice, the H&S approach has become the standard in management (Aguinis et al., 2011a), even if it should not always be applied in every MA. Divergences in statistical methods alone cannot explain the different levels of MA development in these two fields, nevertheless. These methods are readily available to both, and, in fact, they inspire various methodological recommendations of the most recent management-related MA updates (Aguinis et al., 2011b; Geyskens et al., 2009). By contrast, compared to the one single roadmap that most MAs in management follow, there are various normative guidelines in clinical research depending on the type of primary studies they synthesize. The reasons for the underutilization of MAs in management could consequently go beyond simple statistical issues.

Our thesis is that the slower adoption of meta-analytic practices in management has to do with the diverse perspectives shown by the analysis and even the way data are presented. We can synthesize the origins of the gap between the two scientific fields into four major areas: (1) limited replicability of MAs in management, mainly due to the difficulty in obtaining information from primary research, (2) the emphasis authors give to the selection and reporting of the statistical method, (3) scarce use of heterogeneity analysis and methods for triangulating results, and (4) a lack of norms in meta-analytical procedures and subsequent reporting in management. All these factors hinder the review processes, slow down the accrual of knowledge, and reduce its usefulness for several groups of business stakeholders (entrepreneurs, managers, labor unions, government, and so on).

Section "Background: meta-analytic procedures in clinical research and management" summarizes the literature to introduce the comparison between meta-analytic procedures in management and those used in clinical research. Once we identified the main differences, we chose what we consider a high-quality and highly cited management research MA (Mackelprang & Nair, 2010), representative of what can be generally observed in the area, easy to understand for a non-specialist and published in a top outlet such as the *Journal of Operations Management*. We use this MA to exemplify our initial diagnosis and delve qualitatively into the different results authors could have obtained if they had followed the standards of clinical research. The study by Mackelprang and Nair (2010) addresses the relation between just-in-time (JIT) and firm performance (Forza & Di Nuzzo, 1998; Thomé et al., 2016). Thus, using the normative guidelines of clinical research as a benchmark, we illustrate in section "A case study" the areas of improvement with a holistic view of the entire meta-analytical process. Then, section "Results and discussion: reviewing and completing the H&S approach" synthesizes the main lessons to offer a proposal for meta-analytic standards in management (MASM). This proposal includes the entire analytical process—beyond the mere statistical analysis—from the systematic literature review and the information gathering to the final reporting of results. Our recommendations target not only MA authors, nevertheless, but also the authors of primary research studies and the editors who publish them.

Background: meta-analytic procedures in clinical research and management

There are several methods to synthesize a body of knowledge on a certain phenomenon related to bibliometric techniques (see Zupic & Čater, 2015 for a list), such as mapping the intellectual structure of research (e.g., Dabic et al., 2014), bibliographic coupling, network analysis of references or co-citations (e.g., Martín-de Castro et al., 2019), or even discursive reviews (e.g., Martínez-Noya and Narula, 2018). Each of them is appropriate to answer different questions, such as delving into the main theoretical approaches, who the experts in a topic are, or what emergent topics are being addressed. Their ultimate goal is to provide the state of the art and, frequently, a discussion on a future research agenda. However, all these bibliometric techniques mainly provide qualitative rather than quantitative responses.

By contrast, MA is a type of quantitative literature review that essentially seeks two goals (Miller & Pollock, 1994): (1) the integration of primary research results by contrasting hypotheses and (2) the presentation of new hypotheses not included in primary research. Thus, in addition to the integration of results or the refutation of established hypotheses, heterogeneity analysis of primary studies opens up new avenues of inquiry (Carlson & Ji, 2011).

Regarding the first goal, meta-analytic integration offers aggregated results about a particular phenomenon—by

Table 1. Breakdown of frequencies of hits citing C&H and H&S, by research domain.

Research domain	# of citing C&H	% Of total hits citing C&H	# of citing H&S	% Of total hits citing H&S	Share of the domain	
					C&H share	H&S share
Clinical: related areas	2,158	33.2	534	10.6	**80.2%**	19.8%
Business Mgmt.: related areas	282	4.3	1,207	23.9	18.9%	**81.1%**

Pearson's chi-square test of independence between rows and columns: chi-square = 1,478.81, 1 d.f. at the alpha level = .05, and p value $< .001$
The table was built with the "search citing works" tool from the Web of Science (WOS) using all citations indexed for C&H, H&S, and their subsequent updates (e.g., Cooper et al., 2009; Hunter & Schmidt, 2004; Schmidt & Hunter, 2015), without restricting the cited work to a certain initial date. We then computed the frequencies within the C&H/H&S matrix and research domains by grouping the research areas related to clinical research and business management research. Significance levels of the test are those of the Pearson's Chi-square test. Bold has been used to highlight the highest proportions in each case.

estimating the target effect size and additional statistics—that are more accurate than those included in each individual study. Very often, management research MAs focus on evaluating the association between some independent and certain dependent variables, so the effect size of interest is a correlation. Clinical research MAs, nevertheless, usually involve the comparison between control and treatment groups, so the effect size of interest is typically the difference of means or a binary variable, and therefore, the effect size is a risk or an odds ratio. In any case, one must keep in mind that to aggregate the findings of multiple, independent studies, researchers need to deal with various measurement methods or sample sizes, among other issues. These "methodological imperfections," which are not attributable to the facts being studied, are called artifacts. Hunter and Schmidt (1990) identify up to 11 artifacts that inflate artificially the estimate of effect size with regard to the real correlation in, primarily, psychometric studies. If the distorting effects of the artifacts are not offset, then the effect size variability between different studies could be mistakenly attributed to moderating variables. In cases where there is high heterogeneity in primary research, MAs help to identify potentially moderating variables that may uncover new hypotheses initially neglected. Hence, the second purpose of MAs stated above.

The process of MAs in the field of clinical research

We posit that the field of clinical research provides an insightful benchmark for good MA practices in management. This statement is based on two major observations: on one hand, we can highlight the maturity of this field in terms of frequency of MA use to produce cumulative knowledge. On the other hand, it is worth noting that MAs in the clinical setting have undergone a significant explicit normative development in recent years (Geyskens et al., 2009). Thus, depending on the type of MA addressed, clinical researchers use different guides such as Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher et al., 2015), Quality of Reporting of Meta-analyses standards–QUORUM and Meta-analyses Of Observational Studies in Epidemiology–MOOSE (Stroup et al., 2000), STrengthening the Reporting of OBservational studies in Epidemiology–STROBE (Vandenbroucke et al., 2007), or the Cochrane Handbook (Higgins & Green, 2011). With regard strictly to the statistics, clinical research also follows the guidelines proposed by Cooper and Hedges (1994a; 80.2% of MAs as shown by Table 1). Overall, the implemented standards reflect an MA process of execution and reporting that follows five stages:

(1) *Problem statement, variables at stake, and formulation of hypotheses*. At this stage, researchers pose the problem statement and identify the studies to be incorporated, as well as the type of information that will be collected. The variables and the research hypotheses must be defined clearly and precisely.
(2) *Literature review and data collection*. This second stage involves a systematic literature review and the collection of data. It is advisable for this stage to be conducted by two or more reviewers. There must also be a clear and explicit definition of the search strategy and the inclusion (and exclusion) criteria. In clinical research, studies not yet published that address the variables of interest are included, along with those published in scientific journals.
(3) *Evaluation of the studies*. This is a key phase because of its influence on the validity of the MA. Beyond clear descriptions of the methodology—in the PRISMA methodology, for example, an informative flowchart is usually included—scholars should codify and include certain parameters from primary studies, such as the measurement of effect, or the information to obtain it; the sample size; and any information about the sample characteristics that may lead to the posterior detection of heterogeneity across studies.
(4) *Analysis of results.* This stage includes the statistical analysis, as well as its interpretation and quality assessment. Once the combined effect measure and its corresponding confidence interval have been obtained, heterogeneity analysis follows. This is crucial for the MA to determine whether the effect

is universal or whether it depends on some contextual variable or classification of the subjects.

(5) *Presentation of results*. In this last stage, researchers must present two main outputs: on one hand, they should report the main results (e.g., combined effect measure, corrections for systematic attenuating artifacts, and correction for interdependent effect sizes). On the other hand, they must also present MA quality assurance measures such as publication bias, outliers, influential studies, and heterogeneity—moderation analyses (Geyskens et al., 2009). Cooper (1990) showed, in fact, that MA reviewers draw heavily on their ability to interpret results, so the organization of content and clarity are key to achieving what Cooper and Hedges (1994b) call *wisdom in integrative reviews*. Although all this information is common in clinical research, MAs in the management field still show a large space for improvement (Schild & Voracek, 2013).

MA in management: main differences with clinical research

The main elements of discussion in management research MAs have converged so far around statistical choices. A non-exhaustive list of papers reflecting these concerns include Cortina (2003), Aguinis et al. (2005), Kisamore and Brannick (2008), Aguinis et al. (2008), Geyskens et al. (2009), Aguinis et al. (2009, 2011), Aytug et al. (2012), and Kepes et al. (2012). Nevertheless, whereas 80.2% of MA studies in the clinical field follow the C&H approach, as recommended by the above-mentioned guides, the share of H&S is 81% in business management MAs (Table 1). This is consistent with the 83% found by Aguinis et al. (2011a) and the 80.8% found by Aytug et al. (2012). Accordingly, given the lack of normative development in management research MAs, this raises the question of whether the implementation of H&S reflects editors' and reviewers' requirements, or whether it is mainly due to a continuous mimicry in the application of MA procedures even when they are not pertinent to the problem and/or data being addressed.

While the emphasis on statistics was probably needed at a time when measurement and analytical diversity generated much confusion, it is necessary now to expand the focus. Following the stages described above, we present a discussion below of the main challenges and how they could be resolved following clinical research.

Hypothesis formulation and data collection: replicability issues. Several authors have identified these first two stages as being the most problematic and associated with the relatively low adoption of MAs in the management field (e.g., Aguinis et al., 2005; Bosco et al., 2017; Geyskens et al., 2009; Kepes et al., 2013). We have decided to merge them because the associated flaws we identify in both of them are inextricable linked.

The main problem is that primary research on management topics does not always report the necessary information, nor does it use transparent practices. The elaboration and contrast of new hypotheses require the availability of sufficient information in primary research to obtain the appropriate measurement of effect size for each subsample. While the primary clinical research studies often include this information due to the required reporting standards, the less frequent use of these normative standards in management results in insufficient relevant information. For instance, Cao and Lumineau's (2015) MAs on the interplay between contractual and relational governance in supply chains had to exclude 402 of the 545 empirical studies because they did not include the required correlations. The exclusion of certain studies from MAs may, therefore, be a cause of publication bias. Another example is the MA on lean production and firm's performance performed by Abreu-Ledón et al. (2018). These authors reported they tried to contact with authors of those primary studies that reported incomplete information, but none of them answered.

Furthermore, there are numerous primary research studies in management that lack sufficient power (sample size) to achieve statistically significant results, which makes management research MAs with adequate sample sizes lack the precision needed to estimate effect size (Dalton & Dalton, 2008). This is the case of the MA performed by Chen et al. (2010) on the antecedents of new product development speed. Their MA presents the adequate sample size, but 11 of the 70 articles were based in sample sizes below 40 cases and some included as low as 24–30 observations in the sample.

Evaluation of studies: conceptualization problem. There is also a problem of semantic confusion in management that prevents MAs from being applied more extensively (Cortina, 2003). While each concept in clinical research is unique and precisely defined, we find conceptual definitions in management that appear to be different, either in the name of the variable or due to definition nuances. In addition, the same concept may be imbued sometimes with different meanings (Mackelprang & Nair, 2010).

This may have been the reason for the enormous propagation of what is currently the most common MA approach in management: Hunter and Schmidt (1990). This type of analysis—called psychometric MAs—comes from the field of operational psychology, and perhaps due to certain similarities with research practices in management (e.g., use of questionnaires, different nomenclatures and metrics for the same concept, and so on), its dissemination has been even identified by some authors as a source of weakness in management research MAs (Aguinis et al., 2011a). Le et al. (2009) ponder the necessary consolidation of the

various conceptual vs. operational definitions and argue that conceptualizing is different from operationalizing its measurement, hence their claim on the need to introduce an attenuation of that artifact. Be as it may, the problem of conceptualization is still pervasive among management research MAs (Cortina, 2003). Rosenzweig and Easton (2010), for instance, present an MA on the tradeoffs that designing a manufacturing strategy entails. They introduce the evolution of the concept *tradeoff in manufacturing* in their background and specifically discuss the problem raised by the different approaches to the conceptualization of these *tradeoffs*. A critical issue follows: the extent to what the papers reviewed may have used the same operationalization while holding on the same conceptualization. In the clinical area, by contrast, there are particular guidelines to transform some measurements into others coherently with the idea expressed by Le et al. (2009).

Analysis of results: the problem of the psychometric approach and evaluation of heterogeneity. Although literature is always evolving in any scientific field, there is a higher consensus in clinical research than in management regarding the meta-analytical procedure or the best statistical technique for a given objective (Kepes et al., 2013). Aguinis et al. (2011a) found 21 different methodological choices that could alter results (e.g., from how to weight the measure of effect size or how to aggregate these measures, to whether fixed or random effects should be used and many others).

A crucial decision that researchers must make is whether to use fixed or random effects to estimate effect size. The fixed-effects model is appropriate when assuming that a large part of the difference in effect size between the primary research studies is due to sample variability (Cooper & Hedges, 1994b; Hunter & Schmidt, 2000). It is, therefore, a *conditional inference* only for the meta-analyzed *k* studies (Hedges & Vevea, 1998). The random-effects model maintains that part of the sample variance is actually due to variations in the effect size, while another part may be due to sampling variance (Overton, 1998). This second option is usually the suggested method for the management field (Kisamore & Brannick, 2008). Nevertheless, authors such as Hedges and Vevea (1998) and Overton (1998) stress that the choice between fixed and random effects should be based on the extent to which the observed variance adjusts to what would be expected if it were distributed *exactly* as predicted by the fixed-effects model (i.e., the same in all the studies) or by the random-effects model (i.e., one part is common and the other is a random variation across the studies). Furthermore, the results of a fixed-effects model can only be valid for the meta-analyzed studies (the sample studies are the population), while random-effects estimations can be extrapolated beyond those studies (the sample studies are only a portion of the total population). In management, this choice is not frequently reported. For instance, the MA by Rosenzweig and Easton (2010) analyzed primary studies that essentially included self-reported measures of tradeoffs in manufacturing, but they did not report whether they used a fixed- or a random-effects model. Both approaches, H&S and C&H, acknowledge this issue.

The other critical decision in this stage is to define the necessary corrections to be made in the effect measurements. The Hunter and Schmidt handbooks provide a detailed technical explanation of the correction of artifacts in the MA of correlations. Hunter and Schmidt (1990) identified up to 11 artifacts in Chapter 2 (e.g., measurement error, dichotomization of variables, or sampling error) that may cause a study's correlation to differ from the actual "true" correlation between independent and dependent variables. Since management research MAs often incorporate psychometric measures, it is advisable to correct for artifacts following the H&S method. The ultimate objective of the study should always be considered, anyway, so researchers should not simply correct the effect measurements systematically neglecting the approach that was used. For example, there is a striking lack of correction for range restriction artifacts in management, for both dependent and independent variables, which Aguinis et al. (2011a) identified in almost 90% of the cases. The C&H approach also advocates for these corrections, although through different computations. Whereas H&S use each study's sample size (ni) and the artifact attenuation factor (A) as an approximation to the optimal weights, clinical research follows essentially the C&H approach, which weights by an estimation of the inverse variance of each effect size.

Once the estimation of effect size is complete, for those cases with high heterogeneity among the primary research studies, the MA helps to identify possible moderating variables and serves as a source of new hypotheses. This is where a very significant problem emerges when scholars follow the H&S approach (Geyskens et al., 2009) because it recommends the use of the well-known Ratio2[1] estimator. This ratio aims at detecting heterogeneity caused by the aggregation of studies that contain different groups based on a certain *moderator* variable. It has been subject to criticisms due to the purely informative, non-probative nature of this test (Borenstein et al., 2009; Schulze, 2004). Accordingly, this literature has proposed additional tests, such as I^2 and H^2—which are widely used in the clinical field—to detect possible moderators. Whereas the I^2 index describes the percentage of variability that is due to heterogeneity rather than to sampling error, the H^2 index is calculated as the ratio of the unaccounted-for variability in effect size to the level of sampling variability (Borenstein et al., 2009). Some authors even recommend using several different methods to triangulate heterogeneity analyses to detect the presence of moderating effects (Kepes et al., 2012, 2013) The difficulties to perform any of these heterogeneity analyses can be found for instance in the MA

Kolev (2016) performed on the relationship between several variables and divestitures. He only tested heterogeneity by means of overlapping in confidence intervals to subgroups and he followed the H&S approach. Another example is the meta-analytic structural equation modeling that Horstmeier et al. (2017) developed in the study of the differential effects of transformational leadership on multiple identifications at work. They only relied on the Hedges and Olkin (1985) *Q* statistic for the heterogeneity analysis and they performed no other triangulation despite the low power of these statistics.

Presentation of results: lack of normative guidance

The lack of normative standards for reporting results is not only a question of aesthetic clarity. It prevents the replicability of MAs, adds complexity for scholars in evaluating their quality, and makes it harder for practitioners to assess their usefulness (Aytug et al., 2012; Grand et al., 2018).

Schild and Voracek (2013) compared the use of graphs across and within the disciplines of medicine, psychology, and business. In the 55 articles analyzed from business, they found that only two had reported a forest plot, a funnel plot, or a flowchart. Meanwhile, among the 523 articles analyzed from medicine, they found 1,037 forest plots (an average of nearly two per paper), 39 funnel plots, and 178 flowcharts. Similarly, Geyskens et al. (2009) found that only 3.5% of the sample analyzed (196 MAs) had incorporated complete information on reliability to ensure the replicability of the studies. The same study by Geyskens et al. (2009) identified other elements that not often appear in the presentation of results, such as analyses of publication bias (only 15% of MAs analyzed), of sensitivity, or of outliers (barely 16%).

The development and dissemination of standardized reporting guidelines are, therefore, essential, not only to provide clarity and transparency to studies but also to allow subsequent updates that support the cumulative build-up of knowledge. In fact, several authors have taken some tentative steps in this direction. Kepes et al. (2013) offer models to summarize the primary research data: the contour-enhanced funnel plot, the funnel plot with trim, or the forest plot of the moderator effects, among others. The recommendations of Kepes et al. (2013), however, have not been implemented in the management research MA literature (this study received a meager average of 13 citations per year in 2016 and 2017). Anzures-Cabrera and Higgins (2010), on the other hand, have offered a guide on how to produce graphic illustrations in MAs with a very similar outcome.

A case study

Selected case and methodology

We use Mackelprang and Nair (2010) as a case study to delve into the strengths and weaknesses of MAs in management research and subsequently develop holistically a proposal for meta-analytic standards beyond statistical considerations. The choice of this article was not random.

To begin with, the article addresses the relationship between JIT manufacturing practices and performance outcomes, which represents an easily comprehensible problem statement even for colleagues working on marketing, finance, international business, strategy, and so on. Furthermore, the article is also a very cited article in its field. A more recent paper would not meet this requirement, but more importantly, it reflects a very relevant issue: we keep basing our analyses on papers that could have altered results if performed differently. Third, the article is of unquestionable quality and has been validated by the reviewers and editors of a prestigious journal, so it is also a good case study to illustrate areas of improvement. To put it bluntly, it would not be very surprising for BRQ readers if we posited areas for improvement using a paper published in a Q4 operations management journal (regardless of its citations), particularly those upon which the replicability of a scientific study depends. Finally, the authors followed Hunter and Schmidt (1990, 2004), the most widespread approach in management as shown in Table 1 and other studies (Aguinis et al., 2011a). This is important because its wide dissemination in management research MA advises to build our proposal upon established strengths, addressing nevertheless its weaknesses by resorting to clinical research common standards.

Together with the insights from section "Background: meta-analytic procedures in clinical research and management," which mainly address areas for improvement in primary studies, replicating Mackelprang and Nair (2010) will allow us to offer a meta-analytical procedure and, specifically, to provide recommendations on how and what information should be included. The result of this effort is an eclectic approach that somehow unifies H&S and C&H as a reflection of current practices in management and potential benefits from imitating clinical MAs.

Data collection

In this stage, we observed two potential areas of improvement that would ensure the replicability of the MA. First, as proposed in clinical research MAs, the first stage should consist of a systematic literature review using different sources to identify all the primary research studies addressing the question at hand. However, similar to many MAs in management, Mackelprang and Nair (2010) limited their search to only certain scientific journals: "the journals in the areas of operations management, management, marketing, and logistics" (Mackelprang and Nair, 2010, p. 286). It is very possible that numerous articles on the subject (published in other journals in the field or related fields) have been overlooked when one considers that criterion. Second, once the global correlation was estimated, it would have been necessary to evaluate the goodness of fit by analyzing

publication bias (which does not appear in the original MA), thus allowing the validity of the MA to be determined.

Furthermore, we could verify the difficulty in replicating management-related MAs because of the limited information contained in the vast majority of the primary studies addressed by Mackelprang and Nair (2010). Following the example of Aguinis et al. (2005), a personal email was sent to all correspondence authors of the 25 primary research studies (listed in Table 2). When emails were returned as undeliverable, they were sent to the other authors. We sent two rounds of emails, but only one author replied. This proves the difficulty in obtaining the necessary information from primary studies, as Mackelprang and Nair (2010) acknowledge in their own paper. They remark not only (p. 296) that seven of the 60 relationships examined did not have sufficient data available but also that when they tried to contact the primary authors, only 13 of the 23 studies that lacked that information replied. We should emphasize the total number of studies they meta-analyzed was 25. This led the authors to substitute reliability measures of JIT practices and performance by the average of those studies reporting them. Coherently, their work stresses the need for journals' editors to demand the complete presentation of results in primary studies (descriptive data, correlations, and confidence measures), as well as the creation of specific forums where authors include supplementary materials or requests for information (e.g., MetaBUS, as reported by Bosco et al., 2017). We should furthermore note that our literature review was not expanded over theirs: we work directly with the set of articles used by Mackelprang and Nair (2010) to avoid possible differences due to a different sample of primary studies.

Regarding statistics, Hunter and Schmidt (2004) was chosen for the first stage. Since management researchers usually incorporate psychometric measures, it is advisable to correct for artifacts (measurement error, dichotomization of variables, or sampling error) that may cause the study of the correlation to differ from the current "true" correlation between independent and dependent variables. In the second stage, we evaluate the combined correlation. Although it is most appropriate from the clinical perspective to employ empirical Bayes or the method of restricted maximum likelihood (REML; Sidik & Jonkman, 2007), we chose the H&S estimator—as Mackelprang and Nair (2010)—to start with the same values. Nevertheless, for comparative purposes, we included the results of both H&S and REML methods. We used the software package R v.3.4.0 (R Core Team, 2017) and the packages metafor 2.0 (Viechtbauer, 2010) and MAc 1.1.0 (Del Re & Hoyt, 2010).

Results and discussion: reviewing and completing the H&S approach

The primary studies to be included in our replication of Mackelprang and Nair (2010) were exactly the same they used. The rationale is comparing the difference in results that merely depends on the approaches themselves. We do include an Appendix 1, however, with the complete search strategy and the criteria for inclusion. The conditions described by Mackelprang and Nair ("journals in operations management, management, marketing, and logistics") could reduce the sample to high-impact journals, which could lead to publication bias.

The overall statistical method can be summarized in four steps: (1) obtaining corrected correlations and variances using both H&S and C&H approaches, (2) aggregating the two families of correlations and variances obtained in the previous step in those cases that reported more than one pair of JIT practices and performance indicators (H&S for correlations, and Borenstein et al. (2009) and Hedges (1989) for variance), (3) estimating tau through REML with variance estimates and compare it to the original H&S approximation with corrections for sampling and measurement errors, including goodness-of-fit indicators (e.g., funnel plot), and (4) evaluating heterogeneity and analyzing moderators following the two methods: H&S and C&H. The main differences lie in how to average independent effect sizes (through sample size for H&S, and from the inverse variance of each effect size for C&H), as well as in the use of more than a single estimator for heterogeneity (Ratio2 for H&S and, in addition to that, I^2 and H^2 for C&H).

Following these steps, we first computed the intra-study correlations—attenuating the necessary artifacts—and then we evaluated the combined correlation between studies. In those cases where the same publication indicates several correlations, we aggregated them into a single measure of effect for each study. It should be noted that we have corrected one mistake (Table 2) in the case of Sim and Curtola (1999). This table also shows the aggregate correlations (*r*) and those corrected for artifacts (*rcorrected*) as computed by Mackelprang and Nair (2010). The main artifacts related to reliability measures of the latent variables were obtained directly from the primary studies (Cronbach's α, composite reliability indicator . . .) and sampling error. With these antecedents, the main differences between the H&S original estimation of Mackelprang and Nair (2010) and our C&H estimation are the following:

1. We did not include any attenuation for variables' reliability. Most of the primary studies did not report them, we could not get them from emails to authors, and their substitution with the average reliability of studies that reported them is not a reliable approximation.
2. We corrected the study correlation to obtain an unbiased estimation of the slight negative correlations, $Yi = ri + (1 - ri^2)/(2*[ni - 3])$ where ri = unattenuated correlation of study *i*. This follows equation 2.7 in Olkin and Pratt (1958), using UCOR in R.
3. We used Hedges (1989) to obtain an approximately unbiased estimate of sampling variance.

Table 2. Information for each study of JIT practices–performance to perform the meta-analysis (H&S and C&H methods).

Study	Raw *r*	*A*	*n*	%Weight H&S	*Rc* H&S	*Vc* H&S	%Weight C&H	*Rc* C&H	*Vc* C&H	Sector	Country	Size
Mehra and Inman (1992)	.193	0.794	114	1.86	0.266	0.005	1.63	0.210	0.005	Other	United States	No info
Sakakibara et al. (1993)	.301	0.727	822	11.25	0.387	0.001	11.74	0.293	0.001	MEAs	United States	No info
Flynn et al. (1995)	.075	0.772	42	0.65	0.101	0.013	0.60	0.076	0.014	MEAs	United States	Other
Lawrence and Hottenstein (1995)	.420	0.745	116	1.66	0.422	0.008	1.66	0.422	0.006	Other	Other	Other
Dean and Snell (1991)	.060	0.760	92	1.38	0.070	0.008	1.31	0.060	0.008	Other	United States	No info
Forza (1996)	.196	0.745	316	3.56	0.333	0.002	4.51	0.263	0.002	MEAs	Other	Other
Jayaram and Vickery (1998)	.132	0.767	53	0.81	0.133	0.018	0.76	0.133	0.020	MEAs	United States	Big
Claycomb et al. (1999)	−.010	0.781	200	3.16	−0.012	0.003	2.86	−0.010	0.004	Other	United States	Big
Sim and Curtola (1999)	.255	0.766	74	0.96	0.188	0.008	1.06	0.149	0.008	MEAs	United States	Other
Callen et al. (2000)	.077	0.767	100	1.52	0.182	0.005	1.43	0.135	0.004	MEAs	Other	Other
Fullerton and McWatters (2001)	.113	0.827	94	1.61	0.215	0.005	1.34	0.157	0.006	Other	United States	Big
McKone et al. (2001)	.304	0.847	117	2.17	0.398	0.005	1.68	0.308	0.004	MEAs	Other	No info
He and Hayya (2002)	.093	0.767	48	0.73	0.184	0.011	0.69	0.137	0.012	Other	United States	Other
Das and Jayaram (2003)	.223	0.804	309	5.17	0.328	0.002	4.41	0.241	0.002	Other	United States	No info
Shah and Ward (2003)	.181	0.748	1508	21.81	0.262	0.000	21.54	0.193	0.000	Other	United States	Other
Ketokivi and Schroeder (2004)	.254	0.730	164	2.26	0.345	0.003	2.34	0.254	0.003	MEAs	Other	Other
Nahm et al. (2004)	.267	0.808	224	3.78	0.268	0.004	3.20	0.268	0.004	MEAs	United States	Big
Challis et al. (2005)	.135	0.616	1024	10.07	0.174	0.001	14.63	0.135	0.001	Other	Other	Other
Li et al. (2005)	.126	0.852	196	3.68	0.154	0.003	2.80	0.126	0.003	Other	United States	Other
Swink et al. (2005)	.279	0.760	57	0.85	0.261	0.010	0.81	0.199	0.010	Other	United States	Big
Narasimhan et al. (2006)	.149	0.776	224	3.49	0.199	0.002	3.20	0.144	0.002	Other	United States	Big
Ward and Zhou (2006)	.169	0.806	769	12.92	0.218	0.001	10.98	0.169	0.001	Other	United States	No info
Avittathur and Swamidass (2007)	−.036	0.764	26	0.39	−0.045	0.025	0.37	−0.037	0.027	Other	United States	Other
Matsui (2007)	.270	0.736	46	0.65	0.491	0.010	0.66	0.351	0.009	MEAs	Other	No info
Dal Pont et al. (2008)	.173	0.724	266	3.61	0.241	0.002	3.80	0.174	0.002	MEAs	Other	Other

Raw r: raw study correlation; *A* attenuation factor. $A=(\alpha x)^{1/2} * (\alpha y)^{1/2}$; α = reliability of JIT practices (x) and performance outcome (y), such as Cronbach's α; *n*: sample size; *Rc* H&S: corrected correlation Hunter and Schmidt method (corrected by *A* and weighted by sample size); *Vc* H&S: estimation of study's sampling variance Hunter and Schmidt method; % weight H&S: percentage of study's weight Hunter and Schmidt method (sample size) $W=n*A^2$; % weight C&H: percentage of study's weight Cooper and Hedges method (inverse variance); *Rc* C&H: corrected correlation Cooper & Hedges method (bias corrected and sample error corrected); *Vc* C&H: estimation of study's sampling variance Cooper & Hedges method; Sector: MEAs: machinery, electronics, and automotive sectors.
We corrected a typographical error in Sim and Curtola (1999): it indicated 0.706 instead of 0.766.

Table 3. Global correlation according to fixed vs. random effects and H&S vs. REML approaches.

Global correlation	Value	*p* value	Credibility interval		Confidence interval
			Lower limit	Upper limit	[Lower limit, Upper limit]
Original M&N (2010) article (fixed effects)	0.25	n.a.	0.12	0.38	[n.a., n.a.]
Our estimate with sampling error correction (random effects, H&S)	0.251	<0.001	0.09	0.41	[0.19, 0.30]
Our estimate with both sampling and error of measurement corrections (random effects, H&S)	0.338	<0.001	0.11	0.56	[0.26, 0.40]
Our estimate with sampling error correction (random effects, REML)	0.196	<0.001	0.07	0.31	[0.15, 0.23]

REML: restricted maximum likelihood; H&S: Hunter and Schmidt method.

These figures allow to assess the global effect measure (global correlation or combined correlation) obtained by the H&S model (0.251, p value $<.001$) under random effects. Our estimate under random effects is the same estimation as that of Mackelprang and Nair (2010) obtained under the assumption of fixed effects and the correction by artifacts (attenuation factor A in Table 2 including the reliability of both the JIT practices and the performance constructs). However, the credibility interval is 1.4 times wider in random than in fixed effects under the same H&S approach (see Table 3). According to Kisamore and Brannick (2008), this wider interval means that uncertainty about the true magnitude of the effect size is higher under random effects than in fixed effects.

A second issue is the approach used to estimate the credibility intervals and, consequently, the estimation of tau, the measurement of between-study variance. Veroniki et al. (2016) provided guidance on which estimation should be used based on the number of studies k, the total number of cases N, how much N varies across studies, and how big tau is. Their overall purpose was to choose the least biased estimator among the usual approaches available (e.g., DerSimonian-Laird, REML, H&S, Hedges, Empirical Bayes, among others). In particular, Veroniki et al. (2016), Novianti et al. (2014), and Viechtbauer (2005) advocate for the use of an REML estimation when the outcomes are continuous, as it is the current case. An additional issue refers to when meta-analysts should use confidence and credibility intervals (Whitener, 1990). Essentially, credibility intervals should be used to address the question of whether moderators are present. A different question is the accuracy of estimations within homogeneous (sub)populations, which should be addressed by confidence intervals. However, confidence intervals should not be used when there are serious doubts on the homogeneity of the population in terms of sampled studies. In heterogeneous samples, a correction for heterogeneity should be introduced (tau2) in the interval computation and this is introduced in the credibility interval.

As Mackelprang and Nair (2010) explained, researchers should use the standard error (SE) in mean correlation for homogeneous studies ($SE=[1-\hat{r}^2]/[N-K]^{1/2}$; $\hat{r}$=corrected correlation; N=total sample size; K=number of studies). If moderators were operating, then the SE formula changes slightly to accommodate the residual variance (SD_{res} in the observed correlations after the variance for sampling error has been removed). SD_{res} should not be corrected, nevertheless, for other artifacts such as measurement error or range restriction because this SE only applies to sample-size-weighted mean correlations, not to the corrected correlations. Whitener (1990) strongly recommends to use credibility intervals to evaluate the potential existence of moderators, while confidence intervals should evaluate the accuracy of estimations within homogeneous (sub)populations.[2]

The estimation under random-effects REML corrected by the same artifacts as the H&S approach is 0.194 (p value $<.001$). This REML approach under a random-effect assumption—inverse variance weighted instead of H&S sample size weighted—is currently preferred in clinical research (Sidik & Jonkman, 2007). In past decades, DerSimonian-Laird was the most common choice just because it was implemented in the main computer programs (Veroniki et al., 2016). Virtually, all the family of moments estimation methods are non-iterative, however, with the exclusion of Paule and Mandel's (1982). By contrast, the family of maximum-likelihood estimators are iterative, which increases the accuracy of the estimation. Consequently, the suitability of using artifact correction and REML to improve estimation accuracy is evident. We should note that the 95% credible interval (0.07, 0.31) is higher than that obtained (0.12, 0.38) by Mackelprang and Nair (2010). Our credibility interval is reached with the *metafor* package of R (Viechtbauer, 2010), whose algorithm is a modification of the method suggested by Higgins et al. (2009) to calculate the prediction interval. The difference observed between both intervals may be due to our choice of random effects instead of fixed effects, and also to the different weighting methods both approaches use. The use of fixed effects assumes that the effect is distributed in the same way in all the studies or, to put it simply, that the population is the same across samples. This represents a controversial assumption

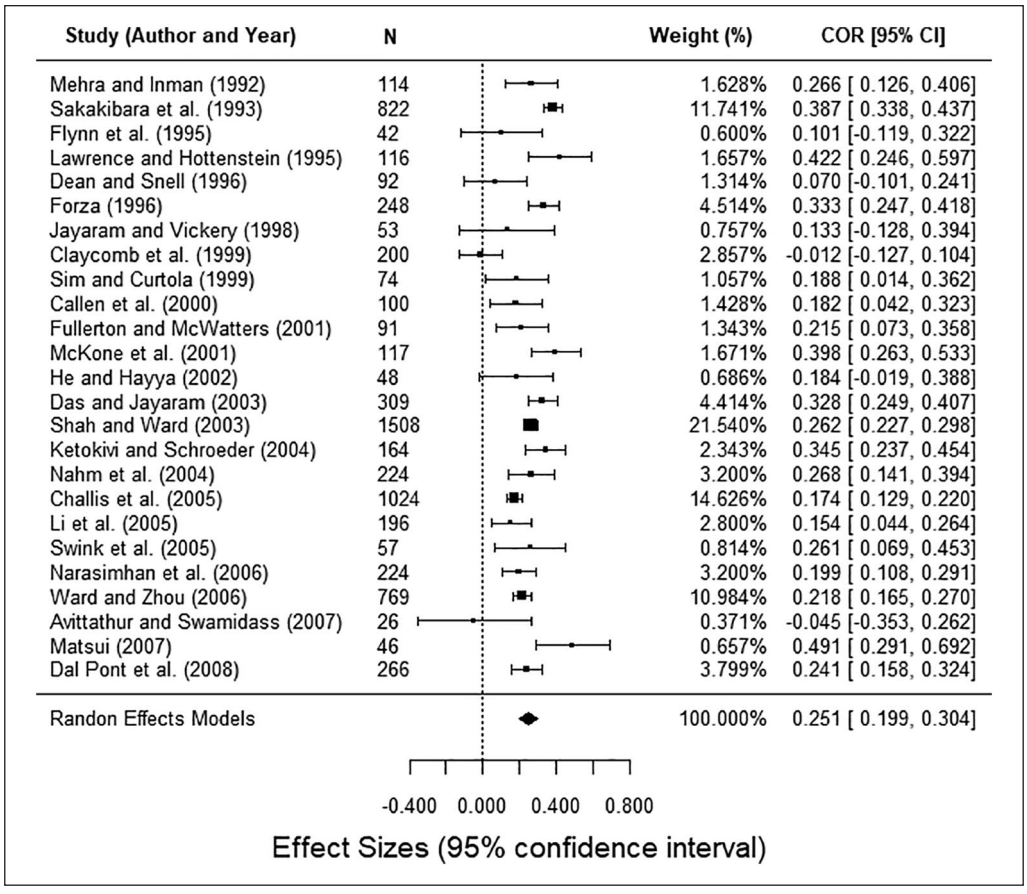

Figure 1. Forest Plot under random-effects model and H&S approach with sampling error correction.

given our later results on moderating variables (Figure 1). The choice of random effects, however, does not presume this assumption, so it probably reflects reality better. Despite this higher interval, the fact that it does not contain the 0 value reinforces the positive relationship between JIT and business performance, although the 0.19 estimate of effect size is lower than the 0.25 original estimate in Mackelprang and Nair (2010). These choices affected how heterogeneity was evaluated, since the estimation of tau differs in both approaches.

Finally, it is worth acknowledging that the variety of statistical software to perform MAs could also become a source of differences. In the H&S approach, the size of each study is the weight to synthesize the results, whereas in the C&H approach, it is the inverse variance. This approach of inverse variance is directly implemented for instance in the meta-essentials package for spreadsheets (Suurmond et al., 2017). It is applied with an additive between-studies variance component based on the DerSimonian and Laird (1986) estimator in the random-effects model. These types of choices may lead to different estimations that, occasionally, may entail different conclusions in terms of heterogeneity and the potential presence of moderators. Anyway, there is a risk in using non-iterative approaches such as the meta-essentials implementation, since spreadsheets are not meant to conduct estimations iteratively. Therefore, researchers cannot choose among the variety of methods to estimate the residual heterogeneity, a feature that is implemented in open-source software R in the *metafor* package. Similarly, the estimation of heterogeneity is another source of differences between C&H and H&S. The lack of normative standards in business management is one of the major differences when compared to clinical research. Scholars and reviewers do not, therefore, know exactly what approach should be chosen for each case, despite the great number of reviews published in methodological journals. We simply replicated the study of Mackelprang and Nair (2010) and chose to evaluate heterogeneity further; our goal was to show the extent to which results could differ, while acknowledging simultaneously that the potential routes a scholar may follow could be very varied.

Reporting of results

In this stage, one of the main contributions of clinical research to the field of management is the use of forest plots (see Figure 1) instead of tables. They allow to see both the amplitude and the variability of the effect measures (of each study and the global effect) in a single diagram, and also spark the possibility of moderating effects

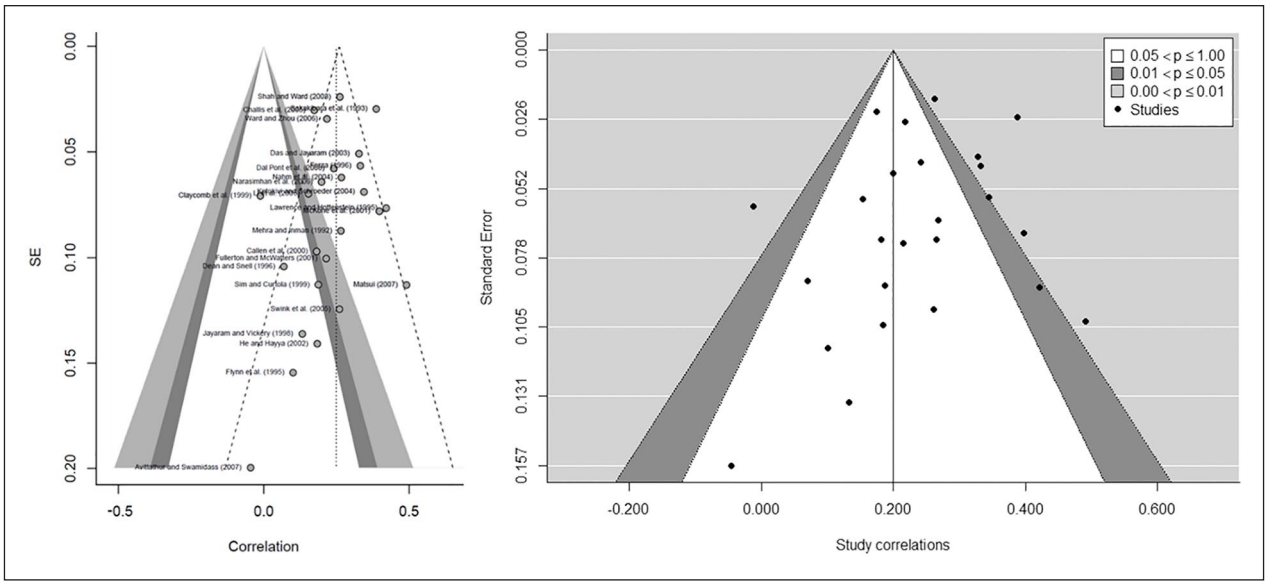

Figure 2. Assessing the existence of publication bias: funnel plot (left) and contour-enhanced funnel plot (right).

when many articles have a 0 in the confidence interval. In addition, organizing the articles ascending by year facilitates the detection of possible temporal patterns. Thus, in Figure 1, all studies are listed by year, *N* indicates the study's sample size, the bar width is the 95% confidence interval of each study's correlation, and the studies' weights follow the H&S approach (in this case, only corrected by sampling error, since the lack of information in primary studies does not allow for any correction for error of measurement).

In the original study, Mackelprang and Nair (2010) include several tables, being very profuse in the detailed MA of the impact of various operative variables on performance (e.g., daily schedule adherence, use of a "pull" system, preventive maintenance, Kanban . . .). Since their goal was an MA of the impact of all those JIT practices on performance, the usual restrictions on paper length probably constrained their capacity to use many graphics. On the other hand, Figure 1 (created with the same data) allows observing the effect that the small sample size has on the accuracy of the estimates, without indicating any particular temporal pattern.

Goodness of fit: analysis of publication bias, heterogeneity, and sensitivity

Once the global correlation is estimated, the next step is to evaluate the goodness of fit to determine the validity of the MA. This evaluation, essential in any quantitative study, is usually ignored in management research MAs. The goodness of fit of the measurements is determined by analyses of publication bias, sensitivity, and heterogeneity.

Publication bias. This analysis determines the validity of the literature review and, consequently, its replicability. We assess the existence of publication bias using a funnel plot (Figure 2). It evaluates whether journals mainly publish papers that contained a significant effect while rejecting to publish non-significant effects. In figure 2, we included the expected triangle under a normal distribution and, with slashed lines, we showed the triangle resulting from the contour-enhanced funnel plot, depicted alone on the right. The right one distinguishes the publication bias from other causes of asymmetry (Peters et al., 2008). The internal white area includes studies with correlation significance above .05 and up to 1.00; it is above .01 and below or equal to .05 in the dark gray area, and equal or below .001 in the clear gray area. Since the graph on the left is not totally symmetrical, it could indicate a certain publication bias; however, the asymmetry regression test does not yield a significant result ($t=-0.915$, $df=23$, $p=.370$). Furthermore, the contour-enhanced funnel plot on the right shows that most of studies are dropped in the white area. Therefore, we cannot confirm that publication bias exists. To put it simply, there is no evidence that the articles published in this topic were only including studies that have detected significant correlations. The result in itself should not be surprising since Mackelprang and Nair (2010) did include studies that had found both significant and insignificant correlations, as recommended by Kepes et al. (2012, 2013).

Sensitivity analysis. Aguinis et al. (2011a) reported that only 16% of management research MAs performed a sensitivity analysis. Mackelprang and Nair (2010) did not perform one. Viechtbauer and Cheung (2010) suggested up to eight plots to assess the potential influence of each study on the overall results. These plots are implemented in R metafor package, which includes externally studentized residuals (rstudent), difference in fits (DFFITS), Cook's distances,

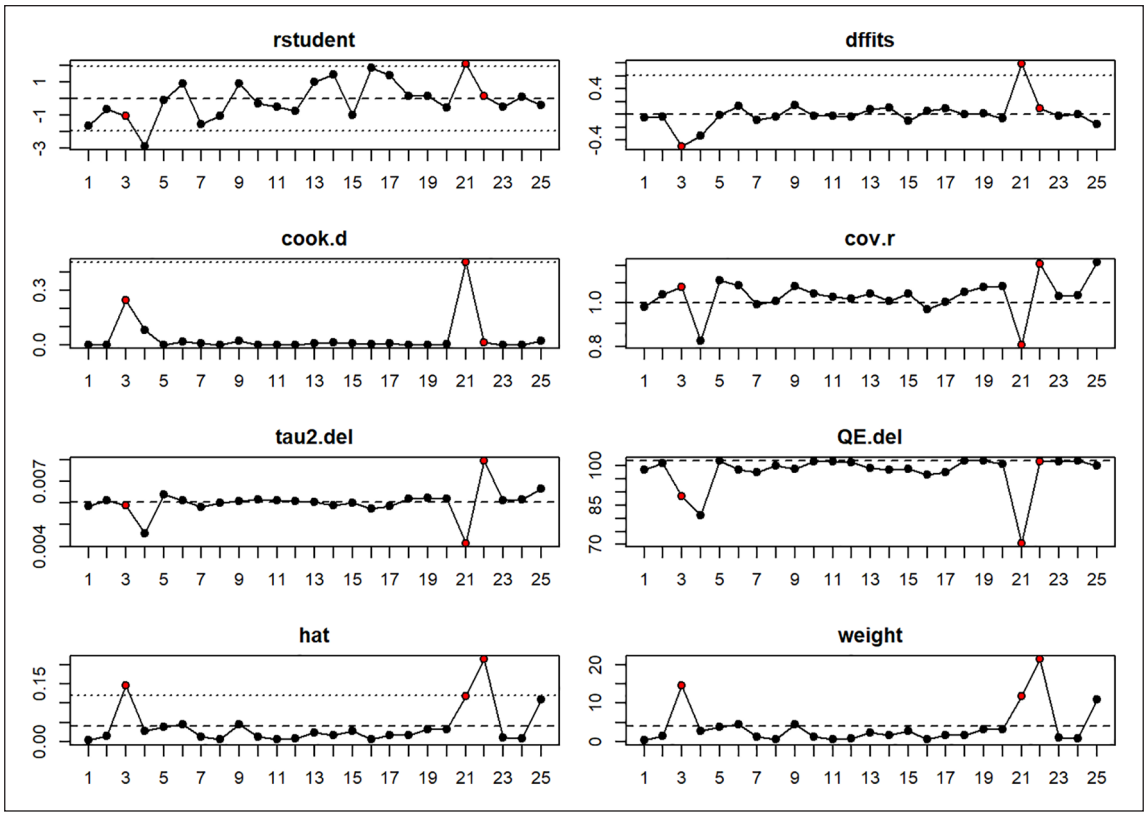

Figure 3. Example of plots for sensitivity analysis.

covariance ratios, leave-one-out estimates of the amount of heterogeneity, leave-one-out values for the test statistics for heterogeneity, hat values, and weights (Figure 3). Thresholds are depicted with dotted lines, whereas slashed ones are used for the average of the metric. The cases identified as potential influencers in any graph are depicted in red in all of them. To evaluate the impact of this influence, we compared the results using leave-one-out analysis. Since we found that none of them influenced the results significantly, they remained in the MA. However, it was possible that the results could have been affected, and their interpretation could have, therefore, changed. It would be useful for the MAs, accordingly, to indicate whether this particular verification has been done.

Heterogeneity analysis. Finally, we performed a heterogeneity analysis. Along with the H&S method's usual Ratio2, we incorporated two additional parameters typical of clinical research: I^2 (total variability due to heterogeneity) and H^2 (total variability/sampling variability; Table 4), as Higgins and Thompson (2002) have defined them. The main weakness of the H&S method is found in the detection of heterogeneity. In fact, even Hunter and Schmidt (1994, p. 335) argue "against excessive confidence in the chi-square significance tests." In that paragraph, they also stated that "The 75 percent rule has been widely misinterpreted. [. . .] If the number of studies in the meta-analysis is small, there is sampling error in the observed variance of sample correlations, that is, second-order sampling error (Hunter & Schmidt, 1990, Chapter 9). Thus, by change, the observed value of Var(r0i)–i.e. the variance of uncorrected correlation–may be larger than that predicted by the sampling error variance formula."

We confirmed this in the following case: if we had only considered the Ratio2 value (1.42), we would conclude, as Mackelprang and Nair (2010) did, that there is no heterogeneity between studies (i.e., all variability is explained by sampling error) and it would not be necessary to assess the influence of possible moderators. In this case, the difficulty faced by the authors was that they followed the widespread practice in the management field of using chi-square-based tests (Geyskens et al., 2009), the result of which indicates the absence of heterogeneity. Abreu-Ledón et al. (2018) triangulated this possibility in their MAs on the impact of lean production on firm's performance, and, in fact, after using the Ratio2, they performed partial meta-regressions by using sector, time, and the country's level of economic development as potential moderators. Thus, triangulation will yield a more holistic perspective of potential heterogeneity.

By contrast, the I^2 index shows that the percentage of unexplained variance in the global effect measure is 62.97% and the H^2 is greater than 1.5 (2.7). Heterogeneity is consequently high (Higgins & Thompson, 2002). In addition, the upper limit of the confidence interval is very

Table 4. Heterogeneity test.

Test	Value	95% CI
Ratio2 (H&S method)	1.42[a]	
Tau^2 (estimated amount of total heterogeneity)	0.0057 ($SE = 0.0029$)	[0.023, 0.0203]
Tau (square root of estimated tau^2 value)	0.0752	[0.0474, 0.1425]
I^2 (total heterogeneity/total variability)	62.97%	[40.3638, 85.9269]
H^2 (total variability/sampling variability)	2.70	[1.6768, 7.1058]
Heterogeneity test Q ($df = 24$)	70.67 ($p < .001$)	

H&S: Hunter and Schmidt method; CI: confidence interval; SE: standard error.
[a]A Ratio2 greater than or equal to 0.75 implies that there is only one population correlation and that the relationship is not subject to moderating factors (Mackelprang & Nair, 2010).

high, indicating that a large part of the heterogeneity is due to actual differences in the studies, rather than to sampling error. The possible moderators are tested next.

Post hoc analysis: identification of potential moderators

Geyskens et al. (2009) recommend the evaluation of potential moderators even if some of the heterogeneity ratios seem to preclude their existence. Because Mackelprang and Nair (2010) followed the H&S method, they could not have done anything further, and consistent with the Ratio2 result, they did not evaluate possible moderators. This highlights again the problem arising from the fact that H&S is the most widespread method in management. In this regard, the authors were consistent with their initial objective of evaluating the impact of JIT management variables on performance. In addition, their approach was also consistent with the results of the heterogeneity test. Therefore, their results guided their decisions and there was no reason to choose another path, despite the low power of this test and the fact that they used a single test (Chowdhry et al., 2016; Sangnawakij et al., 2017). It is worth noting that there are additional tests of moderating effects and there is a general recommendation to evaluate potential moderators even when an initial test such as Ratio2 may point to homogeneity across studies (Aguinis et al., 2011a, 2011b; Geyskens et al., 2009).

To illustrate some potential moderators, we decided to evaluate some of the most common control variables in the lean manufacturing literature: firm size (large vs. others), sector (machinery, electronics, and automotive [MEA] vs. others), and country (United States vs. others), which is consistent with the variables tested by Abreu-Ledón et al. (2018). The results allow us to determine whether any of these factors influence the relationships (Table 5). To conduct this analysis, Mackelprang and Nair (2010) would have needed to codify additional characteristics of the primary study samples, which are aspects that do not always appear as control variables in the statistical analyses of these studies. As explained above, seven studies had no information about the firms' size, and we were unsuccessful in obtaining the missing information from our effort to contact authors.

We verified that the sector is the only moderator with an effect on the relationship between JIT and performance, QM ($df = 1$) $= 6.09$, $p < .05$, according to the Raudenbush's (1994) method. The QM stands for the Q statistic of model sum of squares (Lipsey & Wilson, 2001), which follows a chi-square distribution with $p - 1$ degrees of freedom, p being the number of predictors in the model. It tests whether at least one of the regression coefficients (not including the intercept) is different from 0. In fact, heterogeneity (I^2) decreases from the global 62.9% (see Table 4) to 56.0% (Table 5) in the MEA sector, when sector is introduced as a potential moderator. However, the "*others*" sector continues to maintain a relatively high degree of heterogeneity ($I^2 = 65.7\%$). Thus, the effect is higher in MEA than in other sectors. Neither size nor country was detected as moderators, however. The correlation between JIT and performance in the MEA sector (0.32) is 28% higher than the global average previously obtained (0.25), while that correlation value falls to .21 in the "others" sector. The impact of JIT in the MEA sector is, therefore, 46% higher than that in other sectors, which corroborates the previous results. This could be due to a wider use and tradition of *lean* tools in these sectors. Regarding firm size and country, they do not have a significant moderating effect on the relationship between JIT and performance ($pval_{size} = 0.056$ and $pval_{Country} = 0.057$). This could mean that the *relative* improvement on performance that can be attributed to JIT practices does not depend on the firm or where (country) its facilities are located. Figure 4 presents further detail, as it displays a forest plot by sector (MEA vs. other sectors).

It is worth citing the research of Abreu-Ledón et al. (2018) to illustrate, and to warn, about how the stratification of the data can affect the findings. Contrary to our results, they found a significant moderating effect with country but not with the sector. This could happen because they compared manufacturing vs. services sectors, causing mixed effects in the case of the aggregation of manufacturing in a single sector. Similarly, these authors used the country's economic development as a moderator, while we decided to classify countries into two groups (United States and others). Be as it may, when there is evidence in other studies, this reflection recommends to perform a

Table 5. Meta-regression analysis with potential moderators.

Potential moderator		*K*	Cor.	95% CI	*Q*	Tau^2	I^2 (%)
Size (*)	Large	6	.174	[0.087, 0.261]	10.02	0.0044	50.1
	Others	12	.243	[0.197, 0.290]	21.21	0.0021	48.1
Country	United States	17	.217	[0.162, 0.272]	48.11	0.0066	66.7
	Others	8	.305	[0.233, 0.378]	23.64	0.0061	70.4
Sector	MEA	11	.322	[0.258, 0.385]	27.07	0.0041	56.0
	Others	14	.218	[0.164, 0.272]	44.98	0.0035	65.7

MEA: machinery, electronics, and automotive sectors; *K*: number of observations; Cor.: correlation; *Q*: Cochran heterogeneity statistic; CI: confidence interval.
(*) In seven studies, we could not find the information on size.

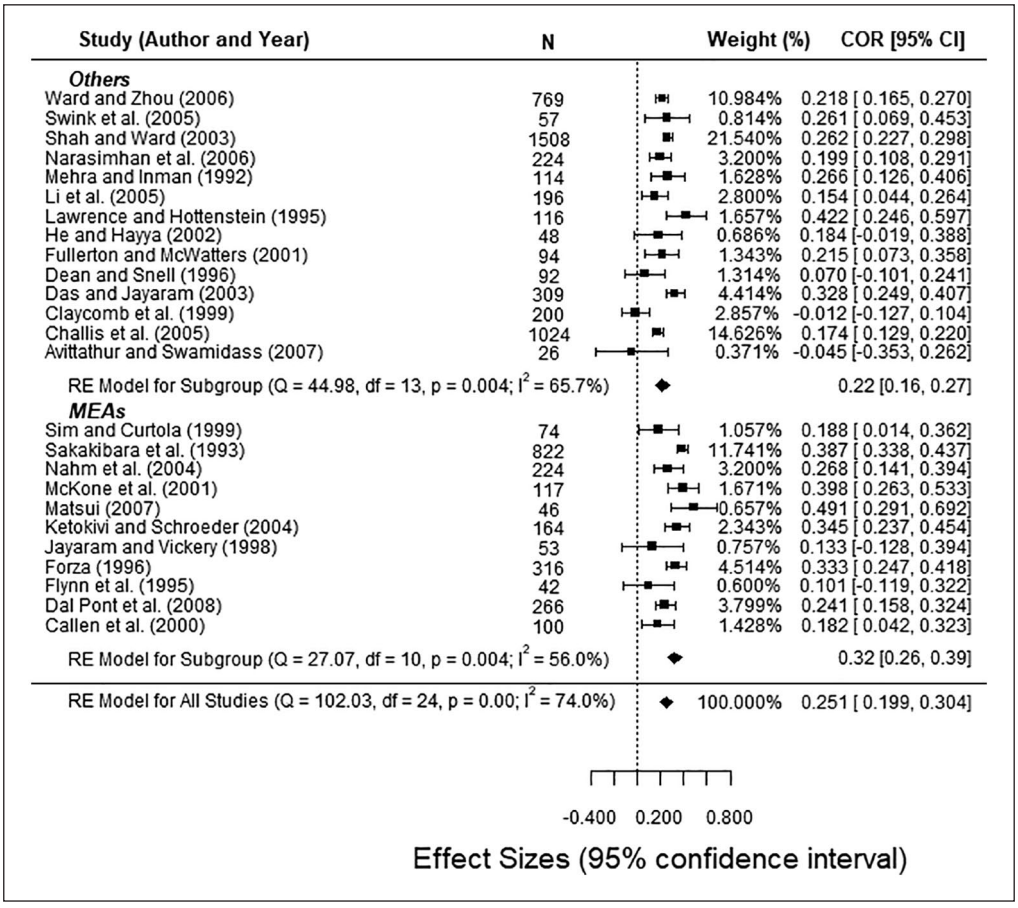

Figure 4. Forest plot with combined correlations for MEAs vs. other sectors.

moderator analysis even in those cases where a statistic is signaling in the opposite direction.

Conclusion: recommendations and "good practices" for MAs in management

Our "background" section has identified different weaknesses in management MAs compared to MAs in clinical research (or H&S vs. C&H) in two ways: some are related to primary studies, due mainly to a lack of information needed, and others are due to non-correct MA execution. The case study, in turn, shows how the analysis of a relevant management MA could change if clinical research procedures were followed. There is, therefore, a wide space for several improvements in management research MA that goes beyond statistical considerations: from data collection to information reporting, including the use of MAs as a means of generating new hypotheses. Particularly, we verified the lack of data in primary studies, as well as the fact

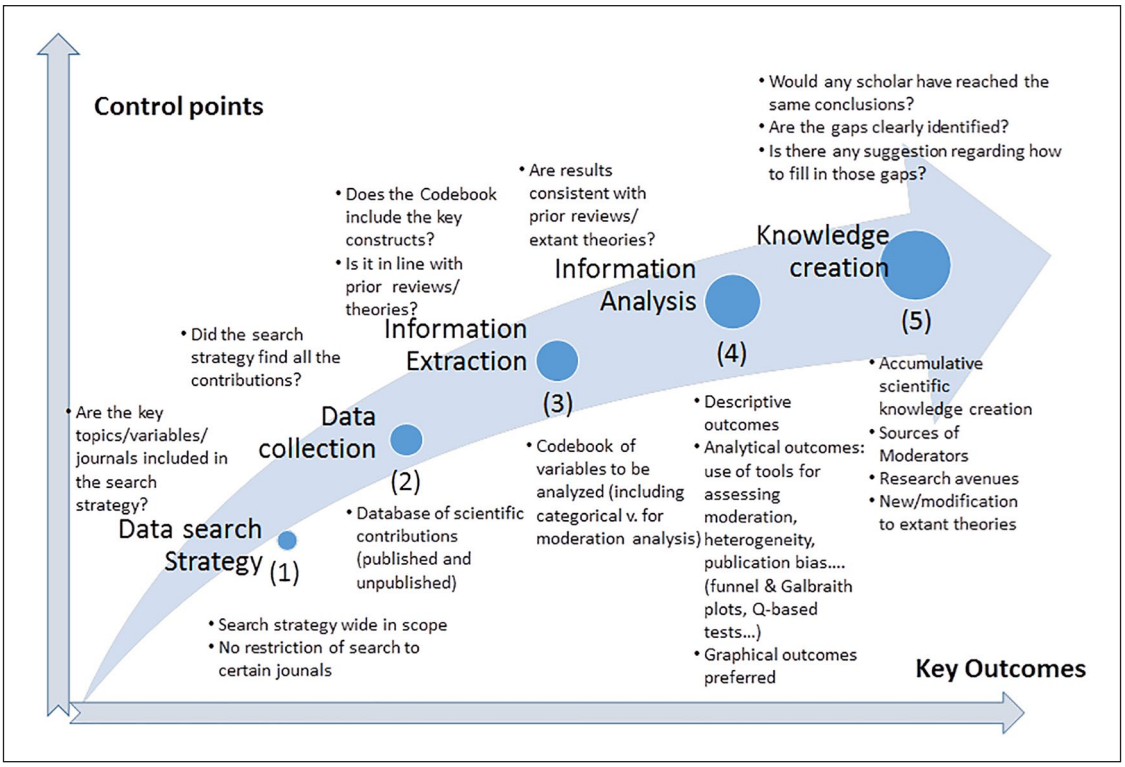

Figure 5. Roadmap for conducting MAs in management based on lessons learned from the clinical research field.

that there are no normative standards to collect information or for the subsequent assessment of MA quality. Furthermore, the lack of reporting standards diminishes the credibility of these studies among scholars and their usefulness for managers, workers, or policy makers. In fact, our analysis has showed the simplicity of detecting moderating effects that can lead to new working hypotheses.

These arguments call for the need to avoid ambiguity in management research MAs. Authors, reviewers, and specially editors need to develop a set of good practices for the correct execution and reporting. Figure 5 illustrates a summary of the five-step process that should inspire this path. The first stages are fundamental for any study with a systematic literature review, as any aggregating study will only be as good as the primary studies on which it is based. Similarly, an MA tends to be more useful the better authors can communicate results. From this point of view, scholars should avoid an excessive use of tables and rely more on illustrative graphics, as in clinical research (e.g., a forest plot organized by the publication year or a funnel plot to evaluate publication bias).

Table 6 summarizes the main challenges (areas of improvement) in conducting a management research MA for each step in the process illustrated in Figure 5. It provides a comparison with the current status in clinical research analysis and ends with a series of recommendations aligned with previous studies (e.g., Aguinis et al., 2011a; Marin-Garcia, 2015).

The adoption of these practices can help other agents out of academia to use the results of management research MAs more frequently. The estimation under REML rather than H&S will yield more accurate estimations of effect size, while the reporting of figures and the triangulation of heterogeneity with additional estimators will offer additional insights on potential moderators. This will enable firms and public policy makers to apply more fine-grained strategies and policies. On one hand, these good practices allow a more efficient analysis since they synthesize (and weigh) all the previous analyses on a certain topic. On the other hand, they allow considering certain key moderating factors, with a large number of studies, which perhaps the previous studies did not address.

With this background, implicitly, our article not only offers several arguments to explain why the use of MAs is not as widespread in management as in other scientific areas but also provides direct evidence of the consequences that current flaws in their execution and reporting cause. All agents involved in the publication process should, therefore, put more effort into adhering to certain work standards and bring transparency and credibility to the studies we carry out and publish.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Table 6. Areas of improvement for MAs in management and recommendations from clinical research MAs.

Areas of improvement	Status of clinical research MAs	Recommendations (to authors and reviewers)
(1) Information availability in primary research: enhancing replicability		
Limited information included in the published studies, both regarding data (correlation matrix, descriptive . . .) and reliability measures. Lack of forums on this topic. Difficulties in contacting authors of primary studies (see Appendix 1).	The necessary measures (odds ratio, relative risk, and hazard ratio) are usually included in primary studies. Requests for information, corrections, or inquiries via forums (e.g. PubPeer) are common.	Journals should require a complete presentation of results (e.g. descriptive, meaningful correlations, and confidence measures). Creation of specific forums where authors and journal editors can link to raw data to provide supplementary material (for example, MetaBUS, as reported by Bosco et al. (2017)).
Semantic confusion in the use of terminology and jumps between the theoretical definition vs. operational definition. Inconsistency in the aggregation to constructs and use of "custom" psychometric scales.	Use of precise and accurate terminology without ambiguity for interpretation: clinical terminology.	Use only scientific measures (valid, reliable, shared, and psychometric scales that have already been tested). When there are different names for similar variables or constructs, acknowledge and report this clearly in the rules for inclusion part of the methods section. Be more explicit regarding the link between theoretical and operational definitions.
Lack of systematic literature reviews. Tendency to publish in only top or high-impact journals (which could lead to publication bias).	Standards exist for literature reviews and data extraction forms (e.g. PRISMA, CONSORT, MOOSE). Criteria for inclusion/exclusion and coding established prior to data collection. Two reviewers for applying exclusion rules and coding.	Scientific associations should promote standards for literature reviews and data extraction (e.g. the MAER-Net workgroup on economics or CAMPBELL for the social sciences). Inclusive perspective: consider also including "gray literature" (doctoral dissertations, works in progress, and so on). Include in Appendix 1 (or website) the complete search strategy, with criteria for inclusion and coding.
(2) Analytical methods used: improving statistical application and heterogeneity analyses		
Meta-analysts do not often justify their choice of a specific meta-analytic approach. Mimicry in the application of statistical support.	A specific meta-analytic approach is not always justified in the study, but the authors always follow the established methods for the field (trials, experiments . . .).	Journals should require justifications for the methods used in accordance with the study objectives. Describe the effect(s) that will be meta-analyzed and the rationale for the study.
Massive use of the H&S method motivated by the use of psychometric scales and the need to correct artifacts that eliminate an important portion of variability. The H&S Ratio2 to evaluate heterogeneity is not very powerful. Analyses of moderators are infrequent.	Low use of artifacts (the use of psychometric scales is not common). Heterogeneity estimators (I^2 and H^2) are very powerful. Analyses of moderators are very common.	Heterogeneity analysis using more powerful tests (I^2, H^2). Encode all the variables available in primary studies. Request separate descriptions for each group of units analyzed in the primary studies. If there are theoretical indications of moderation, continue with the study even if initial tests indicate an absence of heterogeneity.
(3) Standardization of results reporting: increasing the use and dissemination of MAs in management		
Deficiencies in the assessment of MA quality: no guidelines exist.	Guidelines exist for reporting results according to the type of research (e.g. QUORUM, MOOSE, PRISMA).	The scientific associations in the management field should promote the development of standards for reporting results.
MAs are commonly used as compilations of studies.	In addition to compiling information, MAs are also a source of new hypotheses.	Advocate for post hoc studies that generate new hypotheses (e.g. meta-regression), even when only low heterogeneity has been detected.
Low level of usage by businesses and public entities.	Highly useful tool to keep the body of knowledge updated for clinical practitioners or pharmaceutical studies. Extensive use of visual images (e.g. forest plots and funnel plots).	Greater dissemination of MA studies to practitioners, highlighting their usefulness as a synthesis of results and as a source of new areas of research. Journals should put a high priority on the use of graphics and diagrams instead of tables.

MAs: meta-analyses; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Funding

Our work has received financial support from the Spanish and the Galician Governments through grants ECO2016-76625-R and ED431C 2018/46, respectively. Miguel González-Loureiro acknowledges support of Portuguese national funds from FCT – Fundação para a Ciência e Tecnologia through project UIDB/04728/2020.

Notes

1. Ratio2 = $\bar{e}/\sigma^2_{r'}$ where $\bar{e}$ is the weighted mean sampling error variances, and $\sigma^2_{r'}$ is the variance of the corrected correlations (Hunter & Schmidt, 2004).
2. The formula for confidence interval of a global correlation β is $95\%\text{CI} = \beta \pm 1.96\sqrt{SE^2}$, with SE^2 the standard error. The formula for credibility interval entails the estimation of tau (the estimator of heterogeneity): 95% $Cred.Int. = \beta \pm 1.96\sqrt{SE^2 + tau^2}$.

References

References marked with an asterisk indicate studies included in the meta-analysis

Abreu-Ledón, R., Luján-García, D. E., Garrido-Vega, P., Escobar-Pérez, B. (2018). A meta-analytic study of the impact of Lean Production on business performance. International Journal of Production Economics, 200, 83–102.

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, 94–107.

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, 9–34.

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

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, 5–38.

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

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, 306–331.

Anzures-Cabrera, J., Higgins, J. P. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. Research Synthesis Methods, 1, 66–80.

*Avittathur, B., Swamidass, P. (2007). Matching plant flexibility and supplier flexibility: Lessons from small suppliers of U.S. manufacturing plants in India. Journal of Operations Management, 25, 717–735.

Aytug, Z. G., Rothstein, H. R., Zhou, W., Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. Organizational Research Methods, 15, 103–133.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., Rothstein, H. R. (2009). Introduction to meta-analysis. John Wiley.

Bosco, F. A., Uggerslev, K. L., Steel, P. (2017). metaBUS as a vehicle for facilitating meta-analysis. Human Resource Management Review, 27, 237–254.

*Callen, J. L., Fader, C., Krinsky, I. (2000). Just–in–time: A cross–sectional plant analysis. International Journal of Production Economics, 6, 277–301.

Cao, Z., Lumineau, F. (2015). Revisiting the interplay between contractual and relational governance: A qualitative and meta-analytic investigation. Journal of Operations Management, 33, 15–42.

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

*Challis, D., Samson, D., Lawson, B. (2005). Impact of technological, organizational and human resource investments on employee and manufacturing performance: Australian and New Zealand evidence. International Journal of Production Research, 43, 81–107.

Chen, J., Damanpour, F., Reilly, R. R. (2010). Understanding antecedents of new product development speed: A meta-analysis. Journal of Operations Management, 28, 17–33.

Chowdhry, A. K., Dworkin, R. H., McDermott, M. P. (2016). MA with missing study-level sample variance data. Statistics in Medicine, 35, 3021–3032.

*Claycomb, C., Droge, C., Germain, R. (1999). The effect of just-in-time with customs on organizational design and performance. International Journal of Logistics Management, 10, 37–58.

Cooper, H., Hedges, L. V. (1994a). The handbook of research synthesis and meta-analysis. Russell Sage Foundation.

Cooper, H., Hedges, L. V. (1994b). Potentials and limitations of research synthesis. In H. Cooper, L. V. Hedges (Eds.), The handbook of research synthesis and meta-analysis (pp. 521–530). Russell Sage Foundation.

Cooper, H., Hedges, L. V., Valentine, J. C. (2009). The handbook of research synthesis and meta-analysis (2nd ed.). Russell Sage Foundation.

Cooper, H. M. (1990). On the social psychology of using research reviews. In Wachter K. W., Straf M. L. (Eds.), The future of meta-analysis (pp. 75–88). Russell Sage Foundation.

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. Organizational Research Methods, 6, 415–439.

Croucher, R. (2019). Research methods and management. British Journal of Management. https://doi.org/10.1111/1467-8551.12347

Dabic, M., González-Loureiro, M., Furrer, O. (2014). Research on the strategy of multinational enterprises: Key approaches and new avenues. BRQ Business Research Quarterly, 17, 129–148.

*Dal Pont, G., Furlan, A., Vinelli, A. (2008). Interrelationships among lean bundles and their effects on operational performance. Operations Management Research, 1, 150–158.

Dalton, D. R., Dalton, C. M. (2008). Meta-analyses—Some very good steps toward a bit longer journey. Organizational Research Methods, 11, 127–147.

*Das, A., Jayaram, J. (2003). Relative importance of contingency variables for advanced manufacturing technology. International Journal of Production Research, 41, 4429–4452.

*Dean, S., Jr., Snell, A. (1991). Integrated manufacturing and job design: Moderating effects of organizational inertia. Academy of Management Journal, 34, 776–804.

Del Re, A. C., Hoyt, W. T. (2010). MAc: Meta-analysis with correlations. R Package Version 1.0.5 [Computer software]. http://CRAN.R–project.org/package_MAc

DerSimonian, R., Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7, 177–188.

*Flynn, B. B., Sakakibara, S., Schroeder, R. G. (1995). Relationship between JIT and TQM: Practices and performance. Academy of Management Journal, 38, 1325–1360.

*Forza, C. (1996). Achieving superior operating performance from integrated pipeline management: An empirical study. International Journal of Physical Distribution and Logistics Management, 26, 36–63.

Forza, C., Di Nuzzo, F. (1998). Meta-analysis applied to operations management: Summarizing the results of empirical research. International Journal of Production Research, 36, 837–861.

*Fullerton, C., McWatters, S. (2001). The production performance benefits from JIT implementation. Journal of Operations Management, 19, 81–96.

Geyskens, I., Krishnan, R., Steenkamp, J. B. E., Cunha, P. V. (2009). A review and evaluation of meta-analysis practices in management research. Journal of Management, 35, 393–419.

Grand, J. A., Rogelberg, S. G., Allen, T. D., Landis, R. S., Reynolds, D. H., Scott, J. C., Tonidandel, S., Truxillo, D. M. (2018). A systems–based approach to fostering robust science in industrial–organizational psychology.' Industrial and Organizational Psychology, 11, 4–42.

*He, J., Hayya, C. (2002). The impact of just–in–time production on food quality. Total Quality Management, 13, 651–670.

Hedges, L. V. (1989). An unbiased correction for sampling error in validity generalization studies. Journal of Applied Psychology, 74, 469–477.

Hedges, L. V., Olkin, I. (1985). Statistical methods for meta-analysis. Academic Press.

Hedges, L. V., Vevea, J. L. (1998). Fixed–and random effects models in meta-analysis. Psychological Methods, 3, 486–504.

Higgins, J. P., Green, S. (Eds.) (2011). Cochrane handbook for systematic reviews of interventions (Vol. 4). John Wiley & Sons.

Higgins, J. P., Thompson, T. S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21, 1539–1558.

Higgins, J. P., Thompson, S. G., Spiegelhalter, D. J. (2009). A reevaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172, 137–159.

Higgins, J. P., Whitehead, A., Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. Statistics in Medicine, 30, 903–921.

Horstmeier, C. A., Boer, D., Homan, A. C., Voelpel, S. C. (2017). The differential effects of transformational leadership on multiple identifications at work: A meta-analytic model. British Journal of Management, 28, 280–298.

Hunter, J. E., Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. SAGE.

Hunter, J. E., Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper, L. V. Hedges (Eds.), The handbook of research synthesis and meta-analysis (pp. 323–336). Russell Sage foundation.

Hunter, J. E., Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. International Journal of Selection and Assessment, 8, 275–292.

Hunter, J. E., Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research findings. SAGE.

*Jayaram, J., Vickery, S. K. (1998). Supply-based strategies, human resource initiatives, procurement leadtime, and firm performance. Journal of Supply Chain Management, 34, 12–24.

Kepes, S., Banks, G. C., McDaniel, M., Whetzel, D. L. (2012). Publication bias in the organizational sciences. Organizational Research Methods, 15, 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, 123–143.

*Ketokivi, R., Schroeder, G. (2004). Manufacturing practices, strategic fit and performance: A routine–based view. International Journal of Operations and Production Management, 24, 171–191.

Kisamore, J. L., Brannick, M. T. (2008). An illustration of the consequences of meta-analysis model choice. Organizational Research Methods, 11, 35–53.

Kolev, K. D. (2016). To divest or not to divest: A meta-analysis of the antecedents of corporate divestitures. British Journal of Management, 27, 179–196.

*Lawrence, M., Hottenstein, P. (1995). The relationship between JIT manufacturing and performance in Mexican plants affiliated with US companies. Journal of Operations Management, 13, 3–18.

Le, H., Schmidt, F. L., Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct–level relationships. Organizational Research Methods, 12, 165–200.

*Li, S., Rao, S., Ragu-Nathan, T. S., Ragu–Nathan, B. (2005). Development and validation of a measurement instrument for studying supply chain management practices. Journal of Operations Management, 23, 618–641.

Lipsey, M. W., Wilson, D. B. (2001). Practical meta-analysis. SAGE.

Mackelprang, A. W., Nair, A. (2010). Relationship between just–in–time manufacturing practices and performance: A meta-analytic investigation. Journal of Operations Management, 28, 283–302.

Marin-Garcia, J. A. (2015). Publishing in two phases for focused research by means of "research collaborations." WPOM-Working Papers on Operations Management, 6, 76–80.

Martín-de Castro, G., Díez-Vial, I., Delgado-Verde, M. (2019). Intellectual capital and the firm: Evolution and research trends. Journal of Intellectual Capital, 20, 555–580.

Martínez-Noya, A., Narula, R. (2018). What more can we learn from R&D alliances? A review and research agenda. BRQ Business Research Quarterly, 21, 195–212.

*Matsui, Y. (2007). An empirical analysis of just–in–time production in Japanese manufacturing companies. International Journal of Production Economics, 108, 153–164.

*McKone, R., Schroeder, G., Cua, K. O. (2001). The impact of total productive maintenance practices on manufacturing performance. Journal of Operations Management, 19, 29–58.

*Mehra, R., Inman, A. (1992). Determining the critical elements of just–in–time implementation. Decision Sciences, 23, 160–174.

Miller, N., Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. Cooper, L. V. Hedges (Eds.), The handbook of research synthesis and meta-analysis (pp. 457–486). Russell Sage Foundation.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A. (2015). Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement. Systematic Reviews, 4, 1–9.

*Nahm, M., Vonderembse, A., Koufteros, X. A. (2004). The impact of organizational culture on time–based manufacturing and performance. Decision Sciences, 35, 579–607.

*Narasimhan, R., Swink, M., Kim, S. W. (2006). Disentangling leanness and agility: An empirical investigation. Journal of Operations Management, 24, 440–457.

Novianti, P. W., Roes, K. C., van der Tweel, I. (2014). Estimation of between-trial variance in sequential meta-analyses: A simulation study. Contemporary Clinical Trials, 37, 129–138.

Olkin, I., Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics, 29, 201–211.

Overton, R. C. (1998). A comparison of fixed–effects and mixed (random–effects) models for meta-analysis tests of moderator variable effects. Psychological Methods, 3, 354–379.

Paule, R. C., Mandel, J. (1982). Consensus values and weighting factors. Journal of Research of the National Bureau of Standards, 87, 377–385.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. Journal of Clinical Epidemiology, 61, 991–996.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper, L. V. Hedges (Eds.), The handbook of research synthesis (pp. 301–321). Russell Sage Foundation.

R Core Team. (2017). R: A language and environment for statistical computing. R foundation for statistical computing [Computer software].

Rosenzweig, E. D., Easton, G. S. (2010). Tradeoffs in manufacturing? A meta-analysis and critique of the literature. Production and Operations Management, 19, 127–141.

*Sakakibara, B., Flynn, B., Schroeder, R. G. (1993). A framework and measurement instrument for just–in–time manufacturing. Production and Operations Management, 2, 77–194.

Sangnawakij, P., Böhning, D., Adams, S., Stanton, M., Holling, H. (2017). Statistical methodology for estimating the mean difference in a MA without study-specific variance information. Statistics in Medicine, 36, 1395–1413.

Schild, A. H., Voracek, M. (2013). Less is less: A systematic review of graph use in meta-analyses. Research Synthesis Methods, 4, 209–219.

Schmidt, J., Hunter, E. (2015). Methods of meta-analysis: Correcting error and bias in research findings (3rd ed.). SAGE.

Schulze, R. (2004). Meta-analysis: A comparison of approaches. Hogrefe & Huber.

*Shah, P., Ward, T. (2003). Lean manufacturing: Context, practice bundles, and performance. Journal of Operations Management, 21, 129–149.

Sidik, K., Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. Statistics in Medicine, 26, 1964–1981.

*Sim, A., Curtola, P. (1999). Time–based competition. International Journal of Quality and Reliability Management, 16, 659–674.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A., Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. Journal of American Medical Association, 283(15), 2008–2012.

Suurmond, R., van Rhee, H., Hak, T. (2017). Introduction, comparison, and validation of meta-essentials: A free and simple tool for meta-analysis. Research Synthesis Methods, 8, 537–553.

*Swink, M., Narasimhan, R., Kim, S. W. (2005). Manufacturing practices and strategy integration: Effects on cost efficiency, flexibility, and market–based performance. Decision Sciences, 36, 427–475.

Thomé, A. M. T., Scavarda, L. F., Scavarda, A. J. (2016). Conducting systematic literature review in operations management. Production Planning & Control, 27, 408–420.

Vandenbroucke, J. P., Von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. PLOS Medicine, 4, Article 297.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D., Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods, 7, 55–79.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. Journal of Educational and Behavioral Statistics, 30, 261–293.

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

Viechtbauer, W., Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. Research Synthesis Methods, 1, 112–125.

*Ward, P., Zhou, H. (2006). Impact of information technology integration and lean/ just–in–time practices on lead–time performance. Decision Sciences, 37, 177–203.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. Journal of Applied Psychology, 75, 315–321.

Zupic, I., Čater, T. (2015). Bibliometric methods in management and organization. Organizational Research Methods, 18, 429–472.

Appendix 1

Full meta for syntax

```
####==================Dataframe creation
library(dplyr)
library(gsl)
library(metafor)
library(compute.es)
library(psych)
library(MAd)
library(MAc)

ProjectLocation <- getwd()
DataLocation <- paste0(ProjectLocation, "/Data")
load (paste0(DataLocation, "/META_DF.Rdata"))

####=====================================
##===========Functions for analysis

#====Transformation for approximation population correlation coefficient:
# Function of r (raw correlation coeffcient) and n (sample size used in publication)
Approx_pop_r <- function(r,n){
 G_r <- r+ ((r*(1-r^2))/(2*(n-3)))
 G_r
}
#====Variance estimation for r (Hunter and Schmidt approximation)
hunter_var <- function(r, n){
((1 - r^2)^2)/(n - 1)
}
#====Credibility interval for random effect estimates (Higgins et al., 2011; Viechtbauer, 2010)
Cred_interval <- function(estimate, se, tau2, alpha=0.05)
{
int <- qnorm(1-(alpha/2))*(sqrt(se^2+tau2))
cred_int <- c(estimate-int, estimate+int)
return(cred_int)
}

####=====================================
##===========Creating dataframe for analysis

#===Creating dataframe with effect sizes and sampling variances

# META_DF[[7]]: raw dataframe
# r: raw correlation coefficient, n: sample size
# vtype="UB": approximately unbiased estimates of the sampling variances (Hedges, 1989)

transf_data <- META_DF[[7]]%>%
 dplyr::mutate(G=Approx_pop_r(r=META_DF[[7]][,6], n=META_DF[[7]][,8]))%>%
 dplyr::mutate(SVar_Hedges=escalc(measure="UCOR", ri=G, ni=N, data=META_DF[[7]], vtype="UB")$vi)%>%
 dplyr::mutate(ES_r=escalc(measure="UCOR", ri=G, ni=N, data=META_DF[[7]], vtype="UB")$yi)

####=====================================
##===========Aggregating effect sizes

#====Agreggating within-study effect sizes with Borenstein et al. (2009) procedure
BHHR_agg <- MAd::agg(id=ID, es=ES_r, var=SVar_Hedges, n.1=N, method="BHHR", cor=.50, mod=NULL, data=transf_datos)
# Adding random-effects weights and confidence intervals to the dataframe
BHHR_agg <- data.frame(BHHR_agg, MAc::wgts(es=es, var=var, data=BHHR_agg))

#====Agreggating within-study effect sizes with Hunter and Schmidt (2004) procedure
HS_r <- MAc::agg(id=ID, r=G, n=N,
        cor=.50, mod=NULL, data=transf_datos)$r
# var=hunter_var(transf_datos$ES_r, transf_datos$N)

####=====================================
##===========Random effects models

#===Adjusting models
# META_DF[[6]]: simplified dataframe –> one row per publication. All estimated parameters estimated
# in the previous step have been added
# Random effects model with artefacts (REML estimation)
mod1 <- mareg(es ~ 1, var=var, method="REML",
         data=META_DF[[8]])
# Random effects model with artefacts attenuation (Hunter & Schmidt method)
mod2 <- rma(ri=HS_r, ni=N, measure="COR", var=var_hunter, method="HS", weights=N-1, data=META_DF[[8]])

#===Reporting results: random forest plot
forest_plot <- forest(mod2, test.overall.fixed=FALSE, test.overall.random=TRUE,addcred=F,
      slab=paste(META_DF[[8]]$Study), showweights=TRUE, digits=3,
      mlab="Randon Effects Models", xlab="Effect Sizes",
      cex=0.8, cex.lab=1.2, ilab.xpos=c(-0.65),
      ilab=paste(META_DF[[8]]$N), ilab.pos=4,
      order=order(META_DF[[8]]$Year))

#===Credibility intervals for estimates
CredInt <- Cred_interval(estimate=mod1$beta, se=mod1$se, tau2=mod1$tau2, alpha=0.05)
```

```
#===Heterogeneity estimates and confidence intervals
H_est <- confint(mod2, digits=4)
#===Test fot heterogeneity
summary(mod2)

####=======================================
##==========Goodness of fit analysis

#===Publication bias: funnel plot
library(meta)
HS_meta <- metacor(cor=HS_r, n=N, studlab=Study,
sm="COR", comb.fixed=FALSE, comb.random=TRUE,
            hakn=F, method.tau="HS,," data=META_
DF[[8]])
funnel_plot <- funnel(HS_meta, comb.fixed=FALSE,
comb.random=T, studlab=TRUE,
      cex.studlab=0.5,contour.levels=c(0.9, 0.95, 0.99),
      bg="darkgray",xlim=c(-0.8,0.8),
      xlab="Correlation", ylab="SE")

#===Regression test for funnel plot assimetry
# Egger test:
regtest(HS_metafor, model="lm")
# Random effects version of the Egger test
regtest(HS_metafor, model="rma")

#===Sensitivity analysis
# Leave-one-out analysis (pooled effect recalculated, with
one study omitted each time)
leave1out(mod2)

# Outlier and influential case diagnostics plots (Viechtbauer
& Cheung, 2010): six plots
# (1) Standardized residuals
# (2) Difference in fits (DFFITS)
# (3) Cook's distances
# (4) Covariance ratios
# (5) Leave-one-out estimates of the amount of
heterogeneity
# (6) Leave-one-out values of the test statistics for
heterogeneity
# (7) Hat values
# (8) Weights


influence.plot <- influence(mod2)
plot(influence.plot, plotdfb=TRUE)


####========Post-Hoc analysis: potential moderators
Size_HS <- update(HS_meta, byvar=Size, print.byvar
=TRUE,
            method.tau="HS", level.comb=0.95)
Country_HS <- update(HS_meta, byvar=Country, print.
byvar=TRUE,
            method.tau="HS", level.comb=0.95)
Sector_HS <- update(HS_meta, byvar=Sector, print.
byvar=TRUE,
            method.tau="HS", level.comb=0.95)


#==Forest plot for sector
windows(width=20, height=20)
par(mar=c(1,2,0,2))
forest(Sector_HS, comb.random=TRUE, test.subgroup.
random=TRUE,
   bylab="Sector", comb.fixed=F, test.subgroup.fixed=F,
   fontsize=8, squaresize=0.5, addspace=FALSE,
   fs.heading=7, fs.study=7, fs.random=7,
   fs.predict=7, fs.predict.labels=6, fs.axis=5.5,
   overall.hetstat=FALSE
```