Meta-Psychology, 2019, vol 3, MP.2018.935
https://doi.org/10.15626/MP.2018.935
Article type: Commentary
Published under the CC-BY4.0 license

Open data: Not relevant
Open materials: Not Relevant
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: N/A

Edited by: Rickard Carlsson
Reviewed by: Gustav Nilsonne, Guillaume Rousselet
Analysis reproduced by: Tobias Mühlmeister
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/AYUXZ

An Extended Commentary on Post-Publication Peer Review in Organizational Neuroscience

Guy A. Prochilo*
University of Melbourne

Winnifred R. Louis
University of Queensland

Stefan Bode
University of Melbourne

Hannes Zacher
Leipzig University

Pascal Molenberghs
Institute for Social Neuroscience

Abstract

While considerable progress has been made in organizational neuroscience over the past decade, we argue that critical evaluations of published empirical works are not being conducted carefully and consistently. In this extended commentary we take as an example Waldman and colleagues (2017): a major review work that evaluates the state-of-the-art of organizational neuroscience. In what should be an evaluation of the field's empirical work, the authors uncritically summarize a series of studies that: (1) provide insufficient transparency to be clearly understood, evaluated, or replicated, and/or (2) which misuse inferential tests that lead to misleading conclusions, among other concerns. These concerns have been ignored across multiple major reviews and citing articles. We therefore provide a post-publication review (in two parts) of one-third of all studies evaluated in Waldman and colleague's major review work. In Part I, we systematically evaluate the field's two seminal works with respect to their methods, analytic strategy, results, and interpretation of findings. And in Part II, we provide focused reviews of secondary works that each center on a specific concern we suggest should be a point of discussion as the field moves forward. In doing so, we identify a series of practices we recommend will improve the state of the literature. This includes: (1) evaluating the transparency and completeness of an empirical article before accepting its claims, (2) becoming familiar with common misuses or misconceptions of statistical testing, and (3) interpreting results with an explicit reference to effect size magnitude, precision, and accuracy, among other recommendations. We suggest that adopting these practices will motivate the development of a more replicable, reliable, and trustworthy field of organizational neuroscience moving forward.

Keywords: organizational neuroscience, confidence intervals, self-corrective science, effect sizes, null hypothesis significance testing (NHST), parameter estimation, Pearson correlation, post-publication peer-review, reporting standards.

Note: Correspondence concerning this article should be addressed to Guy A. Prochilo, Melbourne School of Psychological Sciences, University of Melbourne. E-mail: guy.prochilo@gmail.com

Introduction

Organizational neuroscience is a domain of research that draws heavily on social and cognitive neuroscience traditions, but which examines specifically how neuroscience can inform our understanding of people and organizing processes in the context of work (Waldman, Ward, & Becker, 2017). Marked progress has been made at the theoretical level in the decade since its inception. For example, this has included a maturing discussion on the ethics, reliability, and interpretation of neuroscience data and how this applies to organizing behavior and the workplace (Healey & Hodgkinson, 2014; Lindebaum, 2013, 2016; Niven & Boorman, 2016). However, the same level of progress has not been made with respect to careful and consistent critical evaluation of empirical works beyond the point of initial publication. The standards within psychological science (including organizational behavior research) are changing to reflect concerns over the transparency of reporting practices, appropriate use of inferential statistics, and the replicability of published findings (Cumming, 2008, 2014; Cumming & Maillardet, 2006; Nichols et al., 2016; Nichols et al., 2017; Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016). In this extended commentary, we argue that scholars of organizational neuroscience are not considering these implications often enough, especially in major reviews of the literature.

This commentary takes as an example the major review piece by Waldman and colleagues (2017) published in *Annual Review of Organizational Psychology and Organizational Behavior*. In this article, the authors critically evaluate the state-of-the-art of organizational neuroscience, including its methods and findings, and provide recommendations for investing in neuroscience-informed practices in the workplace. However, in what should be an evaluation of the field's empirical work, the authors uncritically summarize a series of studies that: (1) provide insufficient transparency to be clearly understood, evaluated, or replicated, and/or (2) which misuse inferential tests that lead to misleading conclusions, among other concerns. It is customary for scientists and practitioners to cite information from the most recent review pieces, meaning that such reviews (especially *Annual Reviews*) and the references cited therein can wield a disproportionate impact on the future of a field of study. Omission of satisfactory post-publication review in the above work is therefore unfortunate, and may motivate poor decisions that waste scarce time, effort, and financial resources for both researchers and organizational practitioners alike.

This commentary will not be a systematic review of all studies conducted in organizational neuroscience. Instead, to bring explicit attention to the concerns we raise above, we provide a focused post-publication review of five of the 15 empirical studies critically evaluated in Waldman and colleagues (2017). Our commentary therefore dissects a full one-third of studies that were deemed methodologically and statistically sound as part of an evidence base for guiding organizational research and practice (see Table 1 for a list of these studies and justification for their selection). Our motivation for this format, in contrast to a general pooling of findings via systematic review, is threefold. First, at least two of these studies represent seminal works that are among the most influential and highly cited in the field (see Figure 1 for a citation distribution). Second, these studies present with critical methodological or interpretational concerns that have been overlooked in multiple major reviews of the literature. And third, on the basis of these concerns, it is not entirely clear that these studies are being evaluated beyond what is reported in their abstracts by those who cite them. These studies deserve a close and detailed scrutiny and we provide this here.

The primary aim of our commentary is to push the field in a positive direction by encouraging a more critical review of research findings in organizational neuroscience. In doing so, we seek to promote the development of a more replicable, reliable, and trustworthy literature moving forward[1]. First, we contextualize our publication evaluation criteria by discussing what has come to be known as the *replication crisis* in psychological science. While many solutions to this crisis have been offered, we focus on two easily implementable criteria that are likely to have a broad impact: (1) complete and transparent reporting of empirical findings, and (2) statistical inference that considers the magnitude and precision of research findings beyond mere statistical significance. Second, we conduct a post-publication review of selected empirical works in two parts. In Part I, we comprehensively and systematically evaluate the fields' two seminal works with respect to our evaluation criteria. And in Part II, we provide focused reviews of secondary works that each center on a single specific methodological concern that we feel must be a point of discussion as the field moves forward. These concerns are: (1) fMRI statistical analyses that preclude inferences from sample to population, (2) unsubstantiated claims of convergent validity between neuroscience and psychometric measures, and (3) the impact of researcher degrees of freedom on the inevitability of reporting statistically significant results.

[1] *Note*: The concerns we discuss in this commentary are in no way unique to organizational neuroscience. We single out this field, not because it represents a special case, but because we have contributed work to this field (Molenberghs, Prochilo, Steffens, Zacher, & Haslam, 2017).

Table 1

Publications selected for post-publication peer review

	Publication	Justification for selection
	Seminal Works	
1	Peterson et al. (2008)	• This study represents one of the earliest works to apply neuroscience methods to organizing phenomena. It has also been described as the first study do to so following the first theoretical writings in organizational neuroscience (see Ward, Volk, & Becker, 2015). It is one of the most highly cited publications of all those evaluated in Waldman and colleagues' (2017) review ($N = 98$) and is discussed in most reviews of the literature since its publication (e.g., Butler, O'Broin, Lee, & Senior, 2015; Waldman & Balthazard, 2015; Waldman, Balthazard, & Peterson, 2011b; Ward, Volk, & Becker, 2015). On the basis of precedence, citations, and inclusion in multiple reviews, this study would be considered seminal.
2	Waldman et al. (2011a)	• This study is the most highly cited publication of all those evaluated in Waldman and colleagues' (2017) review ($N = 177$) and is included in most reviews of the literature (e.g., Ashkanasy, Becker, & Waldman, 2014; Becker & Menges, 2013; Becker, Volk, & Ward, 2015; Waldman & Balthazard, 2015; Waldman, Balthazard, & Peterson, 2011b; Waldman, Wang, & Fenters, 2016; Ward, Volk, & Becker, 2015). It also cited in several systematic reviews (e.g., Butler, O'Broin, Lee, & Senior, 2015; Nofal, Nicolaou, Symeonidou, & Shane, 2017). It is *the* seminal work of the field.
	Secondary Works	
1	Boyatzis et al. (2012)	• This study represents one of the earliest fMRI studies of the field, and is a highly cited work ($N = 99$). It has also been used as part of the evidence base for guiding research and organizational practice decisions in extended theory pieces (e.g., coaching; Boyatzis & Jack, 2018). This study raises an important methodological concern that we suggest must be a point of discussion among scholars: fMRI analyses that preclude inferences from sample to population.
2	Waldman et al. (2013a, 2013b, 2015)	• This publication represents a single empirical study that has been reported through conference proceedings (Waldman et al., 2013a), as an unpublished pre-print (Waldman et al., 2013b), and within a textbook chapter that discusses it at length (Waldman, Stikic, Wang, Korszen, & Berka, 2015). Cumulatively, these publications have received 32 citations. We include this study for evaluation for several reasons. First, the reporting of this study across multiple venues makes it challenging for scholars to clearly understand and critically evaluate the work. Second, this work is discussed in several major reviews of the literature, and within the textbook that adopts the field's name: *Organizational Neuroscience* (Waldman & Balthazard, 2015). This gives the impression that the work is of high quality. And finally, this study raises an important methodological concern that we suggest must be a point of discussion among scholars: unsubstantiated claims of convergent validity between neuroscience and psychometric measures.
3	Kim and James (2015)	• This study represents one of the most recent fMRI studies conducted in the field and has received 6 citations. While this is low with respect to other works, it is discussed at length in Waldman and colleagues' (2017) major review, and is represented as high quality work. It also raises a specific methodological concern that we suggest must be a point of discussion among scholars: the impact of researcher degrees of freedom on the inevitability of reporting statistically significant results.

Note: Citations were obtained from Google Scholar on Jan 23, 2019. The above studies represent one-third of all studies critically evaluated in Waldman and colleagues (2017) review of the state-of-the-art of organizational neuroscience.

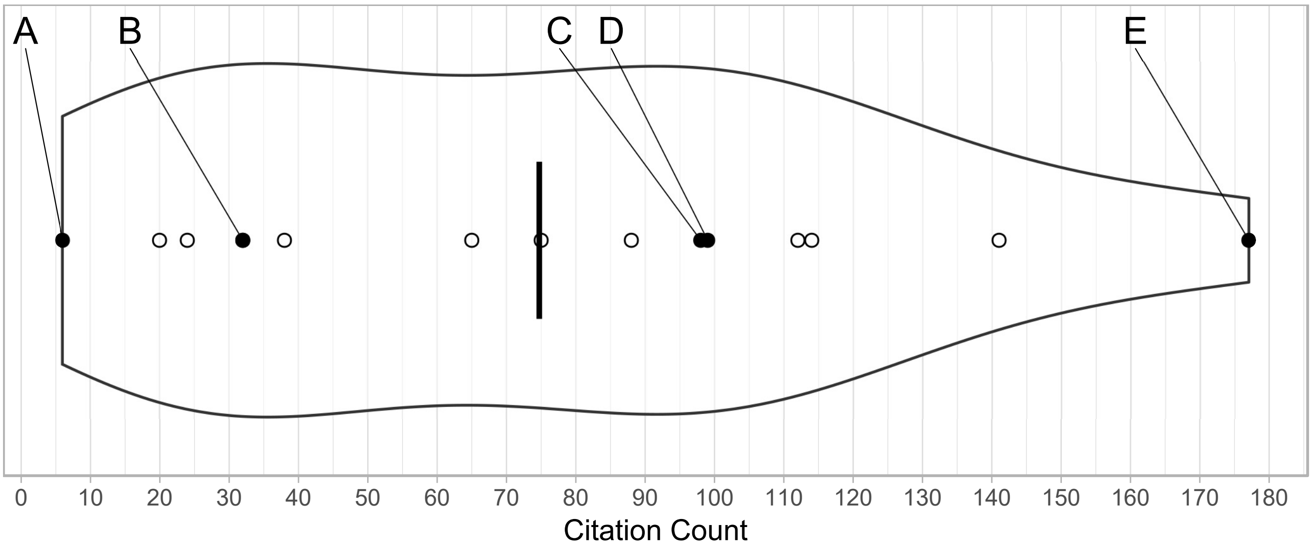

Figure 1. Dot plot with rotated probability distribution on each side of the data showing the number of Google Scholar citations for all 15 empirical studies evaluated in Waldman and colleagues (2017). Citations range from 6 to 177 (M = 74.7, SD = 49.62). The publications under review in this commentary are represented by filled dots and an associated label: (A) = Kim and James (2015); citations = 6; (B) = Waldman et al. (2013a, 2015); cumulative citations = 32; (C) = Peterson et al. (2008); citations = 98; (D) = Boyatzis et al. (2012); citations = 99; (E) = Waldman et al. (2011a); citations = 177. The vertical line marks the mean. With respect to citation impact, these studies capture a cross-section of all studies that were subject to review by Waldman and colleagues (2017). Data were acquired from Google Scholar on Jan 23, 2019.

Publication evaluation criteria

Criteria I: Completeness and transparency of reporting practices

Science has been described as a cumulative and self-corrective process (Merton, 1973; Popper, 1962). Published empirical findings are not taken as unquestionable fact, but rather, all findings are subject to verification through systematic critical evaluation and replication. In doing so, these efforts may provide support that a finding is credible, or that it is wrong, and that the scientific record should be corrected. However, there are growing concerns that a large number of published empirical findings in psychological science are false or at least misleading (Benjamin et al., 2018; Button et al., 2013; Cumming, 2014; Ioannidis, 2005, 2012; Leslie, George, & Drazen, 2012; Munafò et al., 2017; Nelson, Simmons, & Simonsohn, 2018; Simmons et al., 2011; Wicherts et al., 2016). Empirical findings are failing to replicate (Camerer et al., 2018; Klein et al., 2018; Open Science Collaboration, 2015), and these failed replications appear to be invariant to the context and culture in which the replication is attempted (e.g., Klein et al., 2018). This alarming problem has become known as the *replication crisis*, and there is now open discussion that the self-correcting ideal is not performing as well as it should across different areas of science (Ioannidis, 2012).

There are many impediments to self-correction in psychological science, including publication bias, selective reporting of results, and fabrication of data, among others. However, one of the most basic impediments for evaluating published findings as part of a self-corrective science is that researchers do not consistently provide a complete and transparent report of how exactly their research has been conducted and analyzed (Appelbaum et al., 2018). In this commentary we will argue that this has been true of at least some organizational neuroscience work, and that it is particularly prevalent in the seminal works that appear in multiple reviews of the literature.

Complete and transparent reporting is key to systematically communicating what was done in any empirical study. There are multiple systematized reporting standards in psychological science that target various subdisciplines and types of experimental design. For example, the Consolidated Standards of Reporting Trials (CONSORT) is a 25-item checklist that standardizes the way in which authors prepare reports of randomized controlled trial findings (www.consort-statement.org). The OHBM Committee on Best Practice in Data Analysis and Sharing (COBIDAS) guidelines describes how to

Table 2
Outline of publication evaluation criteria

Evaluation Criteria	Description
Method	
Data collection	• Describe methods used to collect data.
Quality of measurements	• Describe the quality of the measurements (e.g., the training of data collectors).
Psychometric properties of instruments	• Describe the reliability and validity of the measurement instruments (e.g., internal consistency for composite scales; inter-rater reliability for subjectively scored ratings; etc.).
Data diagnostics	• Describe the methods used for processing the data (e.g., defining and dealing with outliers; determining if test assumptions are satisfied; and the use of data transformations, if required).
Analytic strategy	
Inferential statistics	• Describe the analytic strategy for how inferential statistics were used to test each hypothesis.
Results	
Statistics and data analysis	• Report descriptive statistics that are relevant to interpreting data (e.g., measures of central tendency and dispersion). • Report appropriate inferential statistics obtained from statistical tests of each hypothesis, including exact *p* values if null hypothesis significance testing (NHST) has been used. • Report effect size estimates and confidence intervals on estimates, where possible. • Report whether statistical assumptions for each test were satisfied.
Interpretation of findings	
Discussion	• Provide an interpretation of results that is substantiated by the data analysis strategy and other aspects of the study (e.g., adequacy of sample size; sampling variability; generalizability of results beyond the sample; etc.). • Consider specifically effect magnitude, accuracy, and precision when interpreting results.

Note: These criteria were adapted from the JARS-Quant guidelines with a specific focus on methods, results, and interpretation of findings. The JARS-Quant guidelines can be found in full in Appelbaum et al. (2018).

plan, execute, report, and share neuroimaging research in a transparent fashion (Nichols et al., 2017; Nichols et al., 2016). And the APA Working Group on Journal Article Reporting Standards (JARS-Quant) covers reporting of all forms of quantitative empirical work, regardless of subdiscipline (Appelbaum et al., 2018).

The JARS-Quant have been designed with the intent of being a gold standard for reporting quantitative research across all of the psychological sciences. This includes research incorporating neuroscience methods such as organizational neuroscience. Therefore, we have adapted a subset of these guidelines to systematically guide our post-publication review of seminal works (see Table 2). These guidelines pertain to criteria that guide the reporting of methods, results, and interpretation of findings. These are elements of research work that are essential for enabling empirical claims to be clearly understood and evaluated by readers, and to allow findings to be replicated with reasonable accuracy.

Criteria II: Appropriateness of statistical inferences

The cause of the replication crisis is multifaceted, and inadequate reporting practices are just a single factor among many contributing to the failure of self-correction in psychological science. A growing number of scholars are also raising concerns that a key theme in this crisis is an overreliance on the null hypothesis significance testing (NHST) approach when conducting research and interpreting results (e.g, Calin-Jageman & Cumming, 2019; Cumming, 2014; Peters & Crutzen, 2017). That is, researchers have traditionally prioritized all-or-none decisions (i.e., a finding is either statistically significant or non-significant) to the exclusion of information that describes the magnitude and precision of a finding, or whether that finding is likely to replicate. For these reasons, findings that are highly variable, imprecise, or which have been selectively reported (or manipulated) based on all-or-none decision criteria have flourished. And these findings are not replicating. We believe similar concerns regarding NHST are influencing the quality of organizational neuroscience.

The NHST approach has been well described elsewhere (see Frick, 1996; Nickerson, 2000). Briefly, an effect (or effect size) describes a quantification of some difference or a relationship that is computed on sample data. As one example, this may include the magnitude and direction of a Pearson correlation coefficient (*r*). In the NHST tradition, a researcher begins by stating a prediction regarding the direction of an effect that they believe to be true of the population from which they are sampling. This is then tested against a null hypothesis which specifies that the true population effect may actually be zero. This test yields a *p* value that quantifies the probability of obtaining a test statistic (e.g., *t*) of a given magnitude or greater when sampling from a population where the null hypothesis is true. In statistical terminology, this is the probability of making a Type I error. In order to minimize such errors, a significance level called alpha (α) is used as the threshold for an all-or-none decision. If the obtained *p* value is less than a prespecified α level, we consider ourselves sufficiently confident to assert that an effect is statistically significant and different from zero. In psychological science this threshold by convention is .05, which entails that in the long-run (i.e., after many replications of a study) we are only willing to accept Type I errors at most 5% of the time.

One of the major criticisms of this approach is that it simply does not provide researchers with the full information they need to describe the relationship between an independent and dependent variable (Calin-Jageman & Cumming, 2019; Cumming, 2014; Cohen, 1990). NHST and *p* values only provide evidence of whether an effect is statistically significant, and of the direction of an effect. Scholars also cite concerns that NHST and its associated *p* values are too often misconstrued or misused by its practitioners, thereby leading to claims that are not substantiated by the data (e.g., Gelman & Stern, 2006; Nickerson, 2000; Nieuwenhuis, Forstmann, & Wagenmakers, 2011; McShane, Gal, Gelman, Robert, & Tackett, 2019). As an alternative (or adjunct) to NHST, proponents of what has been called *parameter estimation* (Kelley & Rausch, 2006; Maxwell, Kelley, & Rausch, 2008; Woodson, 1969) or the *New Statistics* (Calin-Jageman & Cumming, 2019; Cumming, 2014) have argued that inference should focus on: (1) the magnitude of a finding through reporting of effect size, (2) the accuracy and precision of a finding through reporting of confidence intervals on an effect size, and (3) an explicit focus on aggregate evidence through meta-analysis of multiple studies.

On an individual study basis, the parameter estimation approach yields an identical all-or-none decision to that provided by *p* values. However, the focus shifts from a dichotomous all-or-none decision to information regarding the magnitude of an effect, and its accuracy and precision as quantified by confidence intervals.

Accuracy refers to the long run probability that a confidence interval of a given length will contain the true population value. For example, a 95% confidence interval is an interval of values that, if a study were to be repeated many times with different samples from the same population and under the same conditions, the true population value would be included in this interval 95% of the time. It is therefore plausible (although, not certain) that any particular 95% confidence interval will contain the true population parameter. Precision refers to a measure of the statistical variability of a parameter, and is quantified by the width of a confidence interval (or, alternatively, the half width of the confidence interval: the margin of error). For example, a narrow 95% confidence interval is said to have high precision in that there are a limited range of plausible values which the population parameter could take. Conversely, a wide 95% confidence interval is not very precise because the population parameter can take on a very wide range of plausible values.

Some scholars have advocated completely abandoning NHST and *p* values in favor of a parameter estimation approach to statistical inference (e.g., Calin-Jageman & Cumming, 2019; Cumming, 2014). We don't go so far. Instead, in the style of Abelson (1995), we believe that all statistics (including *p* values, confidence intervals, and Bayesian statistics, among others) should be treated as aids to principled argument. However, to limit the scope of our commentary, our evaluations will have an explicit focus on effect size magnitude and, as an indication of accuracy and precision, the confidence intervals on these effects. In doing so, we will argue that NHST and *p* values have been misused across many organizational neuroscience works, and that reviewers of this literature too often accept statistical analyses and interpretations of data uncritically (see Table 2 for our full evaluation criteria).

Post-publication peer review

In the following sections we provide a concise overview of each study listed in Table 1. In Part I, we follow this by a systematic post-publication review of the methods, analytic strategy, results, and interpretation of findings of the fields' two seminal works. In Part II, our post-publication review is focused (and restricted) to specific concerns in secondary works that we suggest must be a point of discussion as the field moves forward. A summary of recommendations for improving post-publication review based on this commentary is given in Table 3.

Part I

Systematic post-publication review of seminal works

Peterson et al. (2008). Neuroscientific Implications of Psychological Capital: Are the Brains of Optimistic, Hopeful, Confident, and Resilient Leaders Different?

The purpose of Peterson et al. (2008) was to examine the neural basis of psychological capital: a composite trait comprised of hope, resilience, self-esteem, and optimism, and which has been linked to effective leadership. Using a sample of 55 business and community leaders, participants were asked to engage in a 'visioning task', in which they were required to create a spoken vision for the future of their business or organization while EEG measures were recorded. As the authors describe, this visioning task was theorized to evoke an emotional response that is aligned with psychological capital. Expert opinions on the affective behavior witnessed during the EEG task were combined with psychometric measures of psychological capital and leadership, and these measures were used to dichotomize participants as high or low on this trait. Following this, differences in EEG activity were assessed between each group.

The authors reported that analysis of their EEG data revealed that high psychological capital was correlated with greater activity in the left prefrontal cortex. This was interpreted as activity associated with greater happiness, as well as having successful interpretation, meaning, construction, and sense-making skills. The authors further reported that low psychological capital was correlated with greater activity in the right frontal cortex and right amygdala. This was interpreted as activity associated with difficulty in displaying and interpreting emotions, as well as a greater likelihood to display negative affectivity or avoidance behaviors in social situations. A primary conclusion provided by the authors (which has been repeated in subsequent reviews) is that these findings are a demonstration of the importance of emotions in the study of psychological capital. The authors further suggest that future research should look more closely at the role of negative affect (e.g., fear) as a mechanism underlying low psychological capital.

Critical review. The critical evaluation of Peterson et al. (2008) first requires qualification based on the venue in which it has been published. *Organizational Dynamics* is a journal that publishes content primarily aimed at organizational practitioners (e.g., professional managers), and therefore restricts full and transparent reporting of methods, results, and analyses in favor of narrative readability for practitioner audiences (Elsevier, 2018). For this reason, the journal encourages publication of supplementary material (which may include detailed methods and results), as well as sharing of data in data repositories that can be directly linked to the article itself. These latter standards may not have been in effect at the time of publication of this early work. In any case, Peterson et al. (2008) does not report any data, or link to any external dataset or supplementary information that can be used to evaluate the content of what is reported. This is problematic because this study has been repeatedly and explicitly cited as an example of high-quality empirical work in almost every review of the literature since its publication (e.g., Butler, O'Broin, Lee, & Senior, 2015; Waldman et al., 2017; Ward, Volk, & Becker, 2015). Because this study is so consistently raised to the status of a high-quality empirical study, Peterson et al. (2008) must be evaluated according to the same standards as any other empirical publication. That is, with adequate post-publication review.

Methods. We first consider the psychometric measures used in this study. The authors claim to have assessed psychological capital using a self-report questionnaire, yet, no information is given regarding what psychological instrument was employed. Furthermore, no information is given regarding the conditions under which this instrument was administered, the psychometric properties of the instrument, or how data acquired from this instrument were processed with respect to outliers or other test assumptions. These same concerns relate to the instrument which was used to assess appraisals of participant leadership characteristics. It is also unclear by what process scores on these instruments were combined to dichotomize participants into groups that were considered representative of high and low psychological capital. And following this process, it is also unclear by what method the dichotomization was performed. Several possibilities include the mean, median, cut-points based on previous literature, or even selective testing of all quantiles and choosing those that yield the smallest p value in a subsequent inferential test, among others. A further complication is that, in addition to each psychometric measure, the dichotomization was also based on affective behavior demonstrated during a visioning task. No information is provided regarding how these ratings were determined, or whether this was implemented correctly. This includes no information on whether coders had the requisite expertise to perform this task, or to what extent dichotomization decisions were consistent across coders. And no information is provided on how this information was weighted alongside psychometric measures to perform the group dichotomization.

We now turn our attention to the EEG measures. The authors provide no information regarding: (1) how EEG data were recorded (e.g., number of channels, electrode configuration, reference electrodes, and sampling rate), (2) how the data were pre-processed (i.e., how artefacts from eye movements, blinks, muscle artefacts, and sweating were identified and removed if necessary, or what filters were applied to remove frequencies of no interest), and (3) whether and how the experimenters controlled for typical artefacts resulting from bodily movements during the experiment. The latter is particularly important given that participants were instructed to talk while EEG recordings were obtained. Movement during EEG recordings can create substantial artefacts (Urigüen & Garcia-Zapirain, 2015).

Altogether, it is extremely difficult for readers to evaluate whether any of the reported measures or methods of data processing were valid, reliable, or implemented correctly. The authors do not provide sufficient methodological detail to the standard that is required of scientific reporting. Because of this, it is unlikely that Peterson et al. (2008) could be replicated with any reasonable level of accuracy.

Analytic strategy. The authors describe that they compared the brain maps of participants who were categorized as high versus low psychological capital. However, they do not specify what analytic strategy was used to perform this test. It is therefore not possible for readers to evaluate whether this analytic strategy was appropriate or implemented correctly. A further concern relates to use of dichotomization itself. Dichotomization of continuous data reduces the efficiency of experimental design, and can lead to biased conclusions that do not replicate across different samples (Altman & Royston, 2006; MacCallum, Zhang, Preacher, & Rucker, 2002; Royston, Altman, & Sauerbrei, 2006; Senn, 2005).

Results. The authors report no statistics. That is, the authors report no measures of central tendency, no measures of dispersion, no inferential statistics, no measures of effect size, no measures of accuracy or precision, and do not report on whether statistical assumptions were satisfied. It is therefore not possible for readers to evaluate any empirical claims on the basis of test statistics.

Interpretation of findings. Peterson et al. (2008) report two main findings: (1) greater activity in the left prefrontal cortex of participants with high psychological capital was indicative of happiness, and (2) greater activity in the right prefrontal cortex and amygdala of participants with low psychological capital was indicative of negative affectivity. This interpretation, however, relies heavily on reverse inference and a highly modular interpretation of regional brain function (for discussion, see Poldrack, 2006). The prefrontal cortex is an incredibly large and diverse region, and is involved in a variety of executive functions, including, but not limited to: top-down regulation of behavior, generating mental representations, goal-directed behavior, directing attention, reflecting on one's intentions and the intentions of others, and regulation of the stress response (Arnsten, Raskind, Taylor, & Connor, 2015; Blakemore & Robbins, 2012; Goldman-Rakic, 1996; Robbins, 1996). Similarly, the amygdala is presently considered a complex and diverse structure that is involved in emotion regulation, motivation, and rapidly processing sensory information of both positive and negative valence (for review, see Janak & Tye, 2015). These regions support functions that lack the specificity to be decomposed into the interpretations provided by the authors, particularly with respect to the methods and analytic strategy that was employed. Additionally, because EEG mainly detects signals originating from sources close to the scalp, activity of deep brain structures such as the amygdala cannot be detected without sophisticated source localization analysis (Grech et al., 2008). It is not clear that this localization analysis has been conducted, and had it been conducted, whether it would be possible to localize the signal to the amygdala specifically. And finally, effect size magnitude, accuracy, and precision are not given any consideration. Altogether, the interpretation provided by Peterson et al. (2008) is not substantiated by their methods and analytic strategy. Subsequent claims regarding the importance of emotion and negative affect in psychological capital may therefore be misleading or entirely false.

Summary. Peterson et al. (2008) lacks an adequately transparent account of what was conducted in their empirical study for it to be clearly understood, evaluated, or replicated with reasonable accuracy. It is a sobering reflection on the field that this work has been cited 98 times in the Google Scholar database with little discussion of what are severe and extreme limitations. It is even more sobering that it has been referenced in almost every major review of the literature since its publication without considering these limitations. Indeed, these limitations are so severe and manifest that it is incomprehensible any reasonable scholar is reading this work before citing it.

In the interest of a self-corrective and cumulative science, we recommend that the findings and conclusions by Peterson et al. (2008) should not be repeated as part of the evidence base for organizational neuroscience in any future literature reviews. Furthermore, given that this is a highly cited and discussed work in the current literature, we also call on the authors to amend their reports following the full JARS-Quant guidelines, and to publish their data and methods openly to allow for re-analysis.

Waldman et al. (2011a). Leadership and Neuroscience: Can We Revolutionize the Way That Inspirational Leaders Are Identified and Developed?

Waldman et al. (2011a) is an EEG study that investigated the neural basis of inspirational leadership, which is a form of leadership that is implicated in desirable organizational outcomes. In a sample of 50 business leaders, participants were asked to engage in a 'visioning task' while undergoing EEG assessment. Vision statements articulated by each leader were coded on a continuum from non-socialized/personalized (rating of '1': self-focused and self-aggrandizing) to socialized (rating of '3': collective-oriented with a positive focus). Visions higher in socialized content were considered to be demonstrative of inspirational leadership. Additionally, three to five followers of each leader (e.g., colleagues or employees), respectively, were asked to rate how inspirational their leader was based on two subscales of the Multifactor Leadership Questionnaire (Bass & Avolio, 1990). The subsequent analysis was restricted to a measure of coherence in the high-frequency beta rhythm above the right frontal cortex. As the authors describe, this measure may have theoretical implications for emotion regulation, interpersonal communication, and social relationships.

The obtained EEG and behavioral data were analyzed through a correlation analysis. Right frontal coherence was positively correlated with socialized vision content coding ($r = .36$, $p < .05$), and follower perceptions of inspirational leadership were positively correlated with the socialized vision content coding ($r = .39$, $p < .01$). However, coherence was unrelated to follower perceptions of inspirational leadership ($r = .26$, $p < .10$). Based on these data the authors draw two main claims. First, they assert that these data indicate their neurophysiological measure of inspirational leadership was more strongly related to an explicit inspirational leadership behavior (i.e., socialized content in vision creation) than to an indirect measure made through follower perceptions of inspirational leadership. This difference in magnitude of correlations was considered indicative of a causal mechanistic chain: right frontal coherence forms the basis of socialized visionary communication, which in turn, builds follower perceptions of inspirational leadership. And second, the authors claim that the correlation between coherence and socialized vision ratings represent a meaningful, neural distinction between leaders who espouse high versus low visionary content. Specifically, they argue that this has implications for leadership development. The particular example discussed by the authors relates to targeted training through EEG-based neurofeedback (i.e., use of an operant conditioning paradigm with real-time EEG feedback). Here, they contend that neurofeedback may be used to enhance ideal brain states associated with effective leadership, such as right frontal coherence.

Critical review. As in Peterson et al. (2008), the critical evaluation of Waldman et al. (2011a) requires qualification based on the venue through which it has been published. The *Academy of Management Perspectives* publishes empirical articles that are aimed at the non-specialist academic reader (Academy of Management, 2018). For this reason, full and transparent reporting of key aspects of empirical work are sometimes eschewed in favor of readability to the non-specialist audience. However, Waldman et al. (2011a) is potentially the most influential work in all organizational neuroscience (see Table 1; Figure 1). Therefore, this publication deserves a comprehensive evaluation of its methods, results, analytic strategy, and claims.

Methods. We first consider the psychometric measures. The authors report that perceptions of inspirational leadership were obtained from three to six followers of each participant using the Multifactor Leadership Questionnaire. The authors also describe that an overall measure of inspirational leadership was computed by summing these responses, which is a practice they describe is consistent with prior research. A measure of internal consistency is also provided, which demonstrated high scale reliability (i.e., $\alpha = .91$). However, the authors do not provide a description of data diagnostics. For example, it is unknown how outliers in the data were identified, whether or not they were removed (and by what method they were removed), or how data were to be treated if it did not meet statistical test assumptions. The method by which participants were coded on the socialized vision rating scale is also unclear. While the authors describe the criteria by which two expert coders categorized participants, no information is provided on how the coders were trained, or the extent to which there was inter-rater agreement between the coders.

Turning our attention to the EEG measures, the authors report the that the 10/20 system has been used, the number of electrodes, and the three electrode locations specific to their analysis. However, there is no information regarding the sampling rate, reference electrodes, or of the general setup. This includes a lack of information on fixation and movement control, and if none were used, how the impact of potential artefacts on the EEG signals have been accounted for. This issue is particularly important given that EEG was recorded while participants were engaged in an active task. As described previously, movement during EEG can cause substantial artefacts. There is also no information provided relating to pre-processing and the use of filters.

Altogether, Waldman et al. (2011a) report a greater depth of information than Peterson et al. (2008). However, Peterson et al. (2008) sets a low standard. Waldman et al. (2011a) requires further detail for an adequate evaluation of the validity and reliability of its reported methods. It may be possible to replicate this method, but the accuracy with which the replication would be conducted may be inadequate.

Analytic strategy. The authors describe that they focused on coherence between three electrodes in the right frontal region of the brain. However, the analytic strategy is not explicitly described and must be inferred from a summary of their findings. Here, coherence data were extracted and subjected to a correlation analysis with ratings of inspirational leadership and socialized vision content coding. The authors do not report the specific correlation analysis that was conducted (i.e., Pearson correlation, Kendall rank correlation, or Spearman correlation). However, the authors use the notation for Pearson correlation (r), and computation of exact p values using Fisher's method (see *Interpretation*) are consistent with those reported in their publication. It can therefore be inferred that the authors have subjected their data to Pearson correlation under the assumption of bivariate normality and no bivariate outliers. Although, the authors do not report whether these latter statistical assumptions were satisfied. Altogether, the lack of transparency of the analytic strategy makes it difficult to evaluate whether it was appropriate or implemented correctly.

Results. The authors do not report descriptive statistics (i.e., central tendency or dispersion) required for interpretation of the psychometric data and socialized vision ratings. This makes it difficult to assess whether the distribution of these data were appropriate for the statistical tests that were performed, and their subsequent interpretation. For example, a restriction of range on either of these measures may influence the subsequent inferential test, the representativeness of the sample, or the generality of conclusions. The authors do report the mean coherence and its range. However, the range is a measure of dispersion that may not be typical of the dataset as a whole, and other measures would be more informative (e.g., standard deviation). Measures of effect size magnitude are reported as Pearson's correlation coefficient, and are accompanied by inexact p values. Reporting inexact p values make it difficult to assess Type I error probability (although, these may be computed from the summary data; see *Interpretation*). Confidence intervals are also absent. Finally, the authors do not report scatterplots of their data. This is problematic because summary correlation coefficients could have been generated by a variety of distributions of data, some of which may render the statistical test inappropriate. For example, Pearson's correlation is not robust, meaning that a single extreme value can have a strong influence on the coefficient (Pernet, Wilcox, & Rousselet, 2013). As a graphical demonstration of this, in Figure 2 we provide examples of 8 distributions consistent with the correlation between frontal coherence and socialized vision content coding reported by the authors (i.e., r = .36, N = 50; plot_r function: cannonball package [v 0.0.0.9] in R Vanhove, 2018). Altogether, the authors do not report their results in adequate detail to fully describe the data.

Interpretation. We consider separately the authors two main claims below.

Claim 1: A difference in correlation magnitudes. A critical conclusion in this publication relates to a comparison of effect size estimates, which in this case involves the Pearson correlation coefficient (r). The authors suggest that the correlation between right frontal coherence and socialized vision content (r = .36, p <.05) is greater than the correlation between right frontal coherence and perceptions of inspirational leadership (r = .26, p < .10). Based on this observation, the authors draw a theoretical conclusion relating to the mechanistic basis of inspirational leadership. This claim appears to be motivated on the basis of eyeballing a difference in the absolute magnitude between these correlations, as well as a difference in the all-or-none decision criteria based on the p values. However, the claim that an r of .36 is greater than .26 assumes that each r is equal to the correlations we would obtain if we were to sample the entire population of relevant business leaders and their followers. That is, not just this sample.

In statistical terminology, this is the assumption that each r is equal to the respective population effect size, rho (ρ). However, r represents the best estimate of ρ in a probability distribution of rs that lie below and above each r estimate (Zou, 2007). Therefore, to determine if one r is greater than another, we must examine the distribution of probable scores within which each ρ may plausibly fall. This can be assessed using a parameter estimation approach by computing a 100(1 - α)% confidence limit on each r. In psychological science, α by convention is .05 which necessitates a 95% confidence interval (95% CI). This interval can be obtained using Fisher's r to z transformation by first calculating the con-

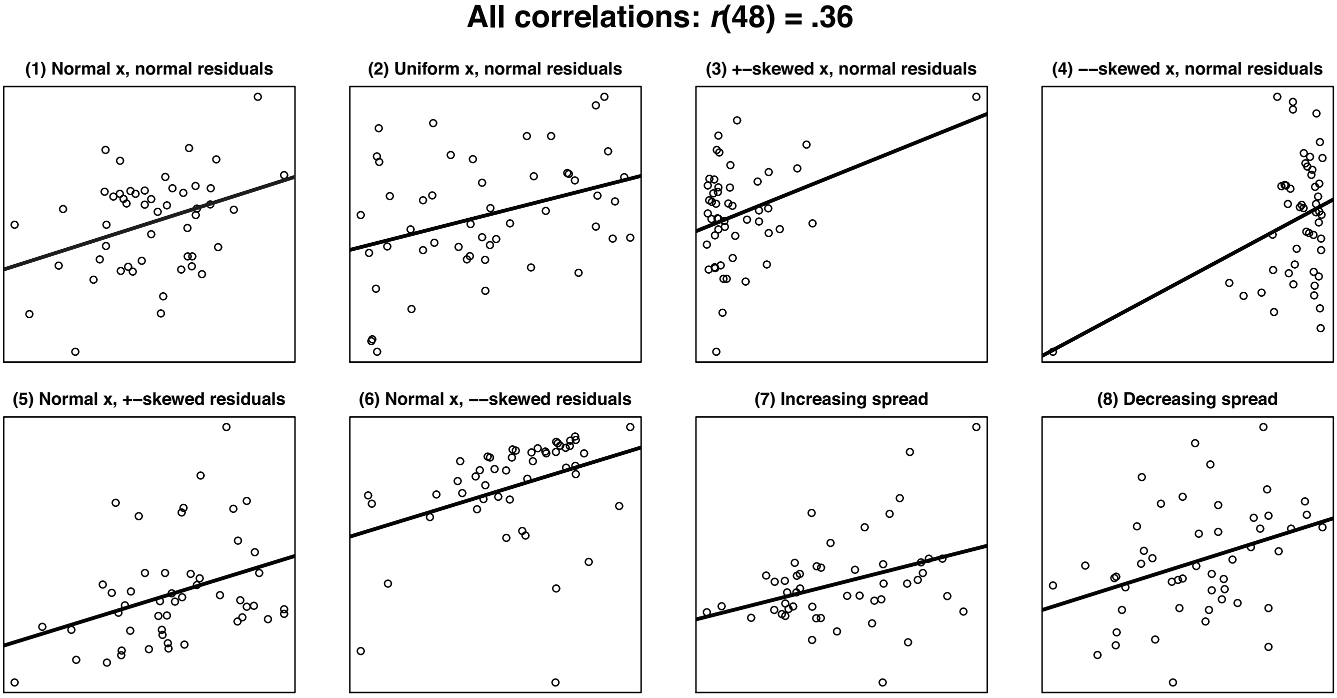

Figure 2. Examples of 8 scatterplots consistent with $r = .36$ in a sample of $N = 50$. The correlation between right frontal coherence and the socialized vision scale could have been plausibly generated by any of these data distributions. This demonstrates the importance of reporting scatterplots in order to verify whether Pearson's correlation analysis was justified, and whether the correlation coefficient is representative of the data that has generated it.

fidence limits for $z(\rho)$ and then back-transforming the limits to obtain a confidence interval for ρ (Zou, 2007). We conduct these analyses below[2]. For completeness, we also report the exact *p* values given by the *t* statistic from each combination of *r* and *N*.

For the correlation between right frontal coherence and socialized vision content we obtain: $r(48) = .36$, 95% CI [.09, .58], $p = .010$ (Figure 3A; lower). For the correlation between right frontal coherence and follower perceptions of inspirational leadership we obtain: $r(48) = .26$, 95% CI [-.02, .50], $p = .068$ (Figure 3A; upper). Focusing on the 95% CI (and ignoring the α criterion for evaluating *p* values), it can be seen that the distribution of possible values of ρ overlap. To test this statistically, however, we must conduct a statistical test of the difference between these correlations using the null hypothesis of a zero difference (i.e., the 95% CI on the difference contains zero). Zou's (2007) method has been recommended for testing the statistical difference between correlations[2] (Cumming & Calin-Jageman, 2016, p. 320). In this case, Zou's method takes into consideration the overlapping dependent correlation between socialized vision content and ratings of inspirational leadership ($r = .39$). Using this method we obtain: $r_1 - r_2 = .10$, 95% CI [-.19, .39] (cocor package for R [v1.1-3]; Diedenhofen & Musch, 2015). The best estimate of $r_1 - r_2$ is .10, however, the 95% CI on this effect is consistent with an interval of values ranging from -.19 to .39 (Figure 3B).

These analyses indicate that we are insufficiently confident to conclude that $r = .36$ is greater than $r = .26$ in these data, and that the difference between these correlations may be zero. Any theoretical claim that relies on a difference between these correlations is therefore not an accurate reflection of the data. Claiming that a difference between a statistically significant correlation and a non-significant correlation is itself statistically significant, is a common misinterpretation of NHST. This is referred to as the *interaction fallacy*, and has been discussed comprehensively elsewhere (Gelman & Stern, 2006; Nickerson, 2000; Nieuwenhuis et al., 2011).

[2] *Note*: These analyses assume bivariate normality, meaning robust alternatives may yield more accurate intervals (Pernet et al., 2013). However, as the author's analyses and claims are performed under this assumption, we also proceed with an assessment of claims assuming this is satisfied.

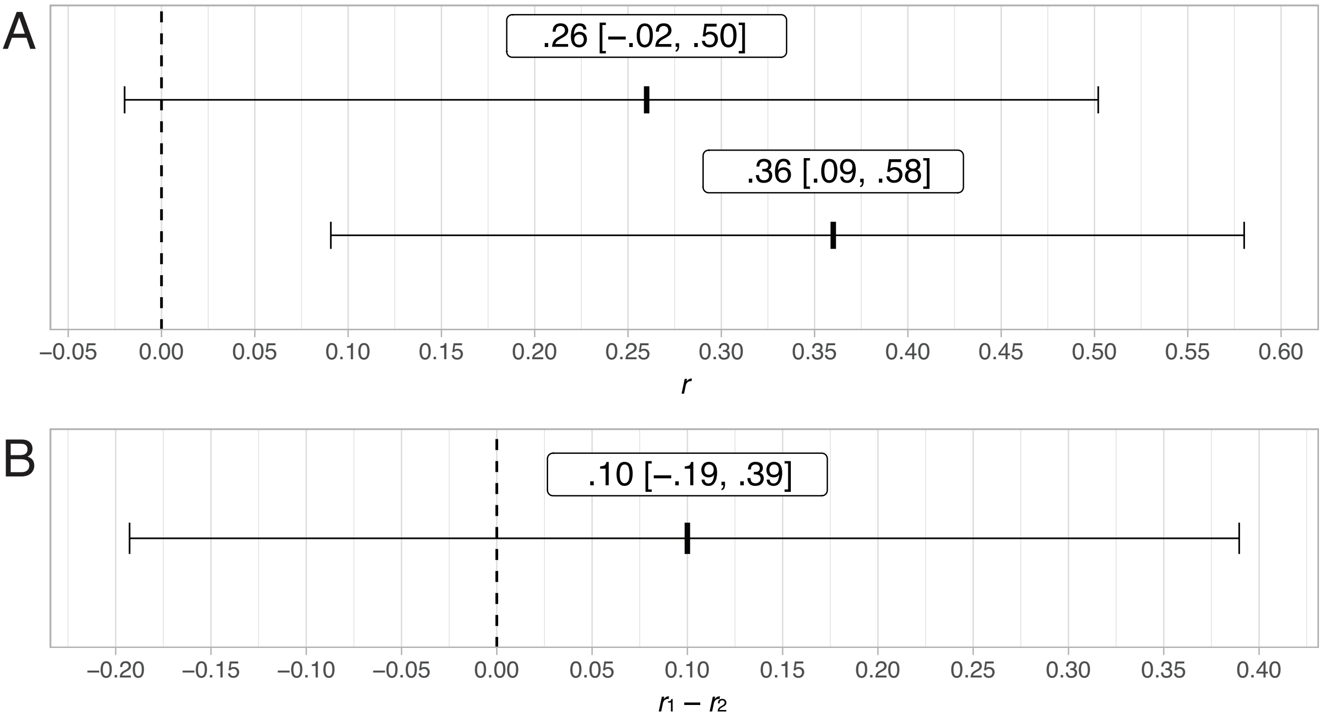

Figure 3. Plot (A): graphical representation of Pearson's r = .26 and r = .36 and their 95% confidence intervals (CI) with a sample of N = 50 (values are given in each label). The 95% CI has been computed using Fisher's r to z transformation under the assumption of bivariate normality. While r = .26 is not statistically different from zero, its 95% CI has considerable overlap with r = .36. Plot (B): a statistical test of the difference between Pearson's r = .26 and r = .36 and its 95% CI, accounting for the overlapping dependent correlation of r = .39 (tails = two-sided, null hypothesis = zero) using Zou's (2007) method. The difference between each correlation is compatible with zero difference.

Claim 2: The meaningfulness of a correlation magnitude. The second critical conclusion in this work is that there is a meaningful, neural distinction between leaders who were considered high versus low in espousing socialized visions. Statistically speaking, this claim rests on the strong assumption that the correlation between right frontal coherence and the socialized vision scale, r = .36, is equal to the population effect size, ρ. However, as we have shown in our calculations above, r = .36 is consistent with an interval of values specified by its confidence interval. The 95% CI on a correlation coefficient tells us that in 95% of random samples from the same population, the 95% CI will contain the population parameter, ρ. It follows logically that in 5% of cases the 95% CI will miss this value. Therefore, we can deduce that it is plausible (although, not certain) that a 95% CI will contain the true value of ρ. In this case, ρ could plausibly range from .09 to .58 (see Figure 3A; lower).

There is indeed a statistically significant neural distinction between leaders who espouse different quantities of socialized content in their visions. However, the precision of the estimate of this association is extremely low. That is, the 95% CI is so wide that the margin of error (quantified as the half-width of a confidence interval) is approaching the magnitude of the estimate itself (i.e., margin of error = .24). This means that there are a great many plausible values for which ρ could take. These values may potentially be negligible for everyday purposes (.09) or even very large (.58). The claim that this correlation is sufficient evidence for use of neurofeedback to enhance coherence is therefore staggeringly disproportionate to the precision with which this effect has been measured.

Drawing substantive conclusions based on a dichotomous all-or-none decision in the absence of effect size magnitude, accuracy, and precision is one of the most widespread misuses of NHST (Cumming, 2014). One of the main reasons that scholars conduct empirical studies is to learn about an effect of interest. When a p value describes an effect as statistically different from zero, yet the confidence interval is very wide, we understand very little about the effect beyond its direction (i.e., positive or negative). For this reason, there has been increasing interest in recent years for planning studies to estimate the magnitude of an effect within a confidence interval that is adequately narrow in width. This has been referred to as accuracy in parameter estimation (AIPE; Kelley & Rausch, 2006; Maxwell et al., 2008; Peters & Crutzen, 2017) or precision for planning (Cumming, 2014). In this paradigm, a key question prior to conducting a study is: what sample size is required to provide a sufficiently precise estimation of an expected effect size of interest? What is *sufficiently precise* depends on one's research objectives. However, one suggestion is to target a margin of error of at least half of the expected effect size (although this is not always a practical solution; Cumming & Calin-Jageman, 2016, p. 266). If Waldman et al. (2011a) considered $r = .36$ to be their best estimate of the expected population effect, ρ, they may consider designing a study that yields a margin of error of no more than .18. This would require a minimum of 92 participants to attain this level of precision with 95% confidence[3] (confIntR: userfriendlyscience package [v 0.7.2] in R; Peters, Verboon, & Green, 2018). It has been suggested that taking a parameter estimation approach to research planning may assist in the production of empirical work that is accurate, precise, and more likely to replicate (Peters & Crutzen, 2017).

Summary. Like Peterson et al. (2008), Waldman et al. (2011a) lacks an adequately transparent description of their study for it to be clearly understood, evaluated, or replicated with reasonable accuracy. Furthermore, the level of description that can be extracted from this report reveals that NHST has been misused or misinterpreted, and has led to interpretations of findings that are not substantiated by the data. This is *the* seminal work in organizational neuroscience: it has been cited 177 times in the Google Scholar database and is discussed at length in most reviews of the literature. Yet, little attention has been given to what are important limitations in methods, analytic strategy, and interpretation of results.

On the basis of this review, we recommend that scholars familiarize themselves with the interaction fallacy and other misuses of NHST (see Gelman & Stern, 2006; Nickerson, 2000; Nieuwenhuis et al., 2011). We also recommend that the results of NHST be considered with explicit reference to effect size and precision to allow for a more informative judgment of research findings. And we further recommend that researchers consider precision for planning in order to attain sufficiently narrow confidence intervals that allow for meaningful conclusions to be drawn from findings. Finally, in the interest of a self-corrective and cumulative science, we suggest that scholars do not carelessly recite the contents of Waldman et al. (2011a) in future reviews of the literature without sufficient critical evaluation. Given that Waldman et al. (2011a) is the seminal work of the field, we also call on the authors to amend their reports following the JARS-Quant guidelines and to publish their data openly for re-analysis.

Part II

Focused post-publication review of secondary works

Boyatzis et al. (2012). Examination of the neural substrates activated in memories of experiences with resonant and dissonant leaders.

Boyatzis and colleagues (2012) is one of the earliest fMRI studies conducted in organizational neuroscience. This study was an exploratory investigation into the neural basis of the personal and interpersonal consequences of interacting with resonant and dissonant leaders, with the implication that such knowledge may inform leadership training and practice. As the authors describe, resonant leaders are considered those whose relationships are characterized by mutual positive emotions, while dissonant leaders are those who invoke negative emotions.

Using a sample of eight individuals with extensive employment experience, participants were interviewed to describe two distinct interactions with two leaders they considered resonant or dissonant, respectively (i.e., four leaders, describing eight interactions total). Audio statements based on each of these eight interactions were created for each participant (8 – 10 s) to be used as cues to recreate an emotional memory of the interaction while undergoing fMRI (5 s). As a manipulation check, participants were also presented with a 4-item question that gauged the valence of their emotional response from strongly positive to strongly negative (2 – 3 s), where recall of resonant and dissonant leaders were expected to yield positive and negative affective responses, respectively. Using an event-related design, each of the eight different cues were randomly

[3] It is important to note from our discussion that $r = .36$ may not be the best estimate of ρ. The authors may therefore take a conservative approach and choose to plan for precision based on the lower limit of the plausible range of values for which ρ could take (i.e., .09). To estimate $\rho = .09$ with a margin of error of no more than half of this expected effect size (i.e., .045) and with 95% confidence, a study would require 1867 participants (confIntR: userfriendlyscience package in R). Conducting studies with precision will require more resources than researchers are accustomed to, particularly when an expected effect size is very small.

presented six times across three runs which resulted in 48 trials in total.

Results of the manipulation check confirmed that emotional responses were all in the predicted direction. Following preprocessing, fMRI data were then analyzed using a fixed-effects analysis. For the contrast between the resonant and dissonant conditions (i.e., resonant > dissonant), the authors reported greater activation in seven regions of interest (ROIs): the left posterior cingulate, bilateral anterior cingulate, right hippocampus, left superior temporal gyrus, right medial frontal gyrus, left temporal gyrus, and left insula. Because the authors tested no hypotheses in this exploratory study, results were interpreted through reverse inference based on existing social, cognitive, and affective neuroscience research. For example, some of these regions have been implicated in the putative mirror neuron system. This system comprises a class of neurons that modulate their activity when an individual executes motor and emotional acts, and when these acts are observed in other individuals (Molenberghs, Cunnington, & Mattingley, 2012). As the authors describe, several regions implicated in this network were activated in response to memories of resonant and dissonant leaders. However, some of these regions were less active during the dissonant memory task. The authors interpreted this as a pattern of avoidance of negative affect and discomfort that was experienced during moments with dissonant leaders, and which may indicate a desire to avoid these memories.

Critical review. Boyatzis and colleagues (2012) describe this study as an exploratory study. Therefore, we critically evaluate this publication as an example of a pilot research and overlook limitations that characterize such works. Such limitations may include (although, not necessarily) a lack of directional a priori hypotheses and a strong reliance on reverse inference. Here we focus specifically on the type of fMRI statistical analysis that has been performed, and the implications this has for drawing inferences from a sample to the whole population.

Drawing inferences about the population from an fMRI analysis. Organizational behavior researchers are typically interested in what is common among a sample of participants in order to permit generalizability of an effect to the full population from which they are sampled. That is, scholars wish to predict and explain organizing behavior beyond the random sample that is included in their study in order to inform organizational theory and practice decisions. The same principle applies to fMRI data analysis. In any fMRI study, the blood oxygen-level (BOLD) response to a task will vary within the same participant from trial-to-trial (within-participant variability) and from participant-to-participant (between-participant variability). Therefore, in order to draw inference from a sample group to the full population of interest, a mass univariate fMRI analysis must account for both within- and between-participant variability (Penny & Holmes, 2007). This is what is referred to as a random-effects (or mixed-effects) analysis, which allows for formal inference about the population from which participants have been drawn.

Boyatzis et al. (2012) report that only a fixed-effects analysis has been conducted on their fMRI data. Fixed-effects analyses account only for within-subject variability, and for this reason, inferences from such analyses are only relevant to the participants included in that specific fMRI study. In this case, inferences therefore only describe the eight participants recruited in Boyatzis et al. (2012). Because between-participant variance is much larger than within-participant variance, fixed-effects analyses will typically yield smaller *p* values that overestimate the significance of effects. For this reason, fixed-effects analyses are not typically reported in the absence of a corresponding random-effects analysis, particularly since the very early days of neuroimaging research (Penny & Holmes, 2007).

The results of fixed-effects analyses are useful if a researcher is interested in the specific participants included in a sample (e.g., a case study), or if it can be justified that the sample represents the entire population of interest. However, because Boyatzis et al. (2012) conducted only a fixed-effects analysis, this means it would be uncertain if the same pattern of activations would be observed if an additional participant were to be included in the study, or if a replication were to be performed. Indeed, the authors report that the exclusion of the single female participant rendered eight regions of interest non-significant, demonstrating the instability of their reported effects and the strong influence of outliers when using fixed-effects analyses. A random-effects analysis is the appropriate analysis to perform if researchers seek to generalize their findings to the population at large.

Summary. Boyatzis et al. (2012) aimed to explain the neuronal basis of interactions with dissonant and resonant leaders, with the implication that such knowledge could improve leadership training and practices. The authors take this step in an extensive review piece describing the neural basis of leadership (Boyatzis, Rochford, & Jack, 2014) and provide explicit recommendations on leadership practice on the basis of these exploratory findings (Boyatzis & Jack, 2018).

This study has been cited 99 times in the Google Scholar database. However, almost no attention has been directed to the inadequacy of the analytic strategy, and the implications this has for generalizing findings from sample to population. While random-effects analyses are consistently reported in fMRI work in the broader social and cognitive neurosciences, we recommend that scholars remain vigilant of this practice in organizational neuroscience. We also recommend that scholars should be aware of this concern when discussing Boyatzis et al. (2012) in future reviews of the literature. Finally, in the interest of a self-corrective and cumulative science, we also call on the authors to repeat their analyses using a random-effect analysis. If these results do not replicate, we call on the authors to correct potentially misleading claims based on these data, and, if necessary, amend their recommendations for leadership practices accordingly. It is also noteworthy that in any case, an adequately powered replication study is required to confirm these findings given that the sample size was very small.

Waldman et al. (2013a). Emergent leadership and team engagement: An application of neuroscience technology and methods.

Waldman et al. (2013a) is an empirical study that used real-time EEG recordings to examine emergent leadership and team engagement. The aim of this study was to investigate whether individual self-reports of engagement could predict whether that individual is likely to be appraised as an emergent leader in a team context. A second aim was to examine whether fellow team members were likely to be more engaged when an emergent leader (compared to a non-leader) used verbal communication during a group-based problem-solving task. To assess these research questions, the authors used psychometric measures of engagement and leadership, and an EEG-based measure of engagement that could be determined in real-time and on a second-by-second basis.

To this end, 146 business administration students were allocated to 31 teams of 4-5 individuals and given 45 minutes to solve a corporate social responsibility case problem in a team setting. During this task, EEG was measured continuously from each participant and time-matched to individual speaking times using video recordings. As the authors describe, the EEG measure was based on a discriminant function that has been used to classify an individual's cognitive state into different levels of engagement (Berka et al., 2004; Berka et al., 2005). At the conclusion of the task, participants were asked to assess their level of engagement retrospectively using the Rich et al. (2010) job engagement measure, and to assess fellow team member levels of emergent leaders behaviors using items from the Multifactor Leadership questionnaire (Bass & Avolio, 1990).

In the analysis that followed, emergent and non-emergent leaders were identified in each of the 31 groups based on the extreme (i.e., highest and lowest) follower ratings of emergent leadership. Using logistic regression (and controlling for gender, age, and the number of friends in each team) self-reports of individual engagement were found to be a significant predictor of categorization as an emergent leader on behalf of other team members ($b = 0.97, p < .05$).

Having demonstrated that self-reports of engagement predicted emergent leader status, the authors conducted a test to determine if other team members were more engaged during periods of emergent leader verbal communication (compared to individuals who scored lowest on follower ratings of emergent leadership). In their methods section, the authors indicated that the team sample size was reduced from $N = 31$ to 26 in the following analyses, due to technical problems with EEG recordings. Pearson correlation analysis was performed between aggregate measures of team-level engagement using the self-report and the EEG measure, which revealed a positive relationship ($r = .32, p < .05$) that the authors interpreted as evidence of moderate convergent validity for their EEG measure. A one-tailed dependent-samples *t*-test revealed no difference in real-time (EEG-based) team engagement for the total time an emergent leader vs. non-leader was communicating during the task ($t = 1.33$, $p > .05$). However, when restricting the analysis to solely the final instance of emergent leader and non-leader communication, real-time team (EEG-based) engagement was found to be greater during emergent leader communication compared to non-leader communication ($t = 2.24, p < .05$).

The authors concluded that individuals who are highly engaged (as measured by self-report) are likely to be appraised by fellow group members as an emergent leader. In turn, the claim is made that emergent leaders may be responsible, in part, for team engagement. This is because the EEG-measure of engagement was greater during the last period of emergent leader versus non-leader communication. From these results, the authors also claim that real-time EEG recordings using their discriminant function represents a valid measure of engagement. It is asserted that such measures may be particularly useful to organizational behavior research investigating ongoing team processes.

Critical review. Waldman et al. (2013a) represents a single dataset that has been reported through multiple venues, where each venue provides a different level of detail regarding methods, analytic strategy, and results. Our critical review is therefore guided by the initial work that was published through conference proceedings (Waldman et al., 2013a), the unpublished preprint available on the ResearchGate repository (Waldman et al., 2013b), and the published textbook chapter within which it is discussed at length (Waldman, Stikic, Wang, Korszen, & Berka, 2015). In this commentary we focus on the neuroscience component of the study and examine the claim that the EEG measure represented a valid index of organizational engagement.

A parameter estimation approach to assessing convergent validity. The claim that emergent leaders generate the greatest level of engagement during verbal communication among fellow team members requires an assessment of the validity of the EEG measure. The authors report that an r of .32 represents moderate convergent validity between the aggregate team-level EEG measure and self-report measures of engagement. However, this interpretation relies on the assumption that the effect size estimate, r, is equal to the population parameter, ρ. As demonstrated earlier in our review, it is plausible (although, not certain) that a 95% CI will contain the true value of ρ. To estimate ρ in this study we therefore apply the Fisher r to z transformation using $r = .32$ and $N = 26$ (noting that the sample size was reduced to 26 because of problems with the EEG recording). We report the exact p values given by the t statistics in these calculations, and for completeness, conduct both a one-sided and two-sided test because it is unclear what test has been performed.

For this assessment of convergent validity, we obtain: $r(24) = .32$, 95% CI [-.01, 1], $p = .056$ (one-sided test), and $r(24) = .32$, 95% CI [-.08, .63], $p = .111$ (two-sided test). Assuming that a one-sided test was reported, a NHST approach to inference tells us that a correlation of the magnitude of $r = .32$ is expected to occur 5.6% of the time when the population parameter, ρ, is actually zero. That is, we are insufficiently confident to comment on the direction and magnitude of this correlation coefficient, and cannot rule out that the true convergent validity is zero. Reporting $p = .056$ as $p < .05$ appears to be a generous rounding of a p value to satisfy an all-or-none decision criterion for publication. Beyond this observation, however, the 95% CI tells us there are a great many plausible values for which ρ could take, including a value of zero.

For validity claims, one suggestion has been that we should consider r magnitudes of .80 and .90 as good (rather than by Cohen's classic definition as very large), and r magnitudes of .60 or .70 as small and inadequate (rather than large; Cumming & Calin-Jageman, 2016, p. 319). On this basis, $r = .32$ may be quite problematic for claiming convergent validity. When we consider the precision with which this correlation has been measured, the plausible value of $r = 0$ (or even close to zero) makes any claim of convergent validity very unpersuasive.

The calculations above also draw into question the claim that the real-time EEG measure used here represents a valid measure of organizationally-relevant engagement, as it has been described in multiple reviews of the literature (e.g., Waldman, Wang, & Fenters, 2016; Waldman et al., 2017). If we suspect that a measure does not have sufficiently high validity, we must be wary of any subsequent conclusions that have been made on the basis of this measure. In this case, this includes the claim that emergent leaders are responsible for generating team engagement.

A possible explanation of the lack of convergent validity observed in this study may be that the EEG measure is simply tracking alertness. In their original validation study, Berka et al. (2004) describe that this EEG measure was developed for the purpose of monitoring mental workload during cognitive tasks. Specifically, they describe that this measure can classify real-time EEG epochs into one of four states of alertness: 'sleepy', 'relaxed wakefulness', 'low vigilance', and 'high vigilance'. In contrast to this, Rich et al.'s (2010) self-report measure defines engagement as an individual's complete physical, cognitive, and emotional investment into a work role or job. This includes items such as "I feel proud of my job" (emotional engagement), and "I devote a lot of attention to my job" (cognitive engagement).

High self-report ratings on measures of pride and attention with respect to one's job or task may covary with moment-to-moment alertness or vigilance, but they do not necessarily need to do so. Ongoing assessments of alertness may be of interest to organizational scholars. However, the relationship between this EEG measure of alertness and organizationally-relevant engagement (and consequently, emergent leadership) is not entirely clear in this study.

Summary. Waldman et al. (2013a) and the book chapter in which it is discussed extensively (Waldman et al., 2015) have been collectively cited a total of 32 times in the Google Scholar database. This work has also been given ample discussion space in several major reviews of the literature (e.g., Ashkanasy, Becker, & Waldman, 2014; Waldman & Balthazard, 2015; Waldman et al., 2017). However, these reviews provide little discussion of whether the authors' claims are a correct reflection of their analyses and results. We recommend that scholars carefully and consistently evaluate whether measures used in organizational neuroscience are methodologically and statistically valid. To do so, we recommend taking a parameter estimation approach to evaluating assessments of convergent validity by considering effect size magnitude and quantifying precision via confidence intervals.

Kim and James (2015). Neurological Evidence for the Relationship between Suppression and Aggressive Behavior: Implications for Workplace Aggression.

The final organizational neuroscience study we critically evaluate is an fMRI experiment conducted by Kim and James (2015). In this study, the authors set out to examine whether brain regions differed in activity during suppression (i.e., a maladaptive emotion regulation strategy) and passive viewing of negatively-laden affective images. Having established this, the primary aim of this study was to determine whether activity in these regions were associated with aggressive behaviors. As the authors discuss, such research may provide insight into factors leading to a reduction of workplace aggression.

Prior to scanning, ratings of aggressive behavior were obtained from two significant others of each participant and averaged, respectively, to give ratings of five different forms of aggression (i.e., physical, property, verbal, relational, and passive aggression). Seventeen participants were then subjected to an fMRI task in which they were exposed to negatively valenced or neutral images from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 2008). Negatively valenced images were preceded by instructions (4 s) to either suppress negative emotional reactions or passively watch each image, while neutral images were always preceded by instructions to passively watch. The latter condition served as a baseline. Following this, four images were presented, and participants employed the instructed emotion regulation strategy (20 s). A manipulation check question was then asked to examine the intensity of negative emotions experienced on a 4-point scale (i.e., neutral to strong; 4 s). To recover from the potential effects of experiencing negative affect, a set of four grey-tone pattern images were presented (20 s), serving as a rest period. This was followed by another manipulation check question, which again measured the intensity of negative emotions experienced (4 s). The task was repeated over 48 trials.

Results of the manipulation check confirmed that participants experienced greater negative emotions following negatively valanced images compared to the rest period. Participants also reported greater negative emotions when using suppression compared to passive observation. According to the authors' discussion (which implicates the use of suppression in the intensification of negative affective experiences) this indicated that the manipulation was successful. FMRI data was subsequently analyzed using a group-level analysis at the whole-brain level. Compared to baseline, the authors reported broadly overlapping areas of activation for suppression and passive watching (e.g., the bilateral visual cortex and insula). For their primary contrasts of interest (i.e., the difference between experimental conditions) the authors reported greater activity in the cingulum and both the left and right insula when engaged in suppression (versus passive watching), and in the calcarine sulcus when engaged in passive watching (versus suppression). Average *t* scores were extracted from each of these four regions for each participant, respectively. These values were then used in a Pearson's correlation analysis against each of the five (psychometrically assessed) aggression ratings. The authors reported a significant negative correlation between average *t* scores in the calcarine sulcus for the passive watching > suppression contrast and property aggression (r = -.49, $p < .05$). However, the authors report there was no significant relationship with any other type of aggression.

Based on these findings, the authors conclude there was a significant association between suppression (i.e., the neural substrate) and aggression (i.e., the psychometric measure), and that this association has implications for organizational practice and research. Specifically, they suggest that use of suppression as an emotion regulation strategy in the workplace will be related to counterproductive behavior in the form of aggression, and that managers should focus on building an organizational climate or set of norms that preclude use of suppression. In their limitations section, the authors acknowledge that their small sample size provides only preliminary evidence for the relationship between suppression and aggression. However, they also claim that the correlation magnitude reported in this study indicates that an equivalent or even larger effect should be observed in studies with larger samples.

Critical review. Kim and James (2015) presents with several important limitations. This includes no provision of reliability statistics for their aggression measures, and no description of how the fMRI data were modeled (i.e., a fixed-effects versus a random-effects analysis), among other concerns. However, our focus here will be on a concern that has not yet received attention in our commentary, and which we feel must be a focal point of discussion in the field moving forward: *researcher degrees of freedom*.

Research designs that will inevitably yield statistical significance. Researcher degrees of freedom refers to the number of arbitrary decisions available to a researcher when formulating hypotheses, designing experiments, analyzing data, and reporting results (de

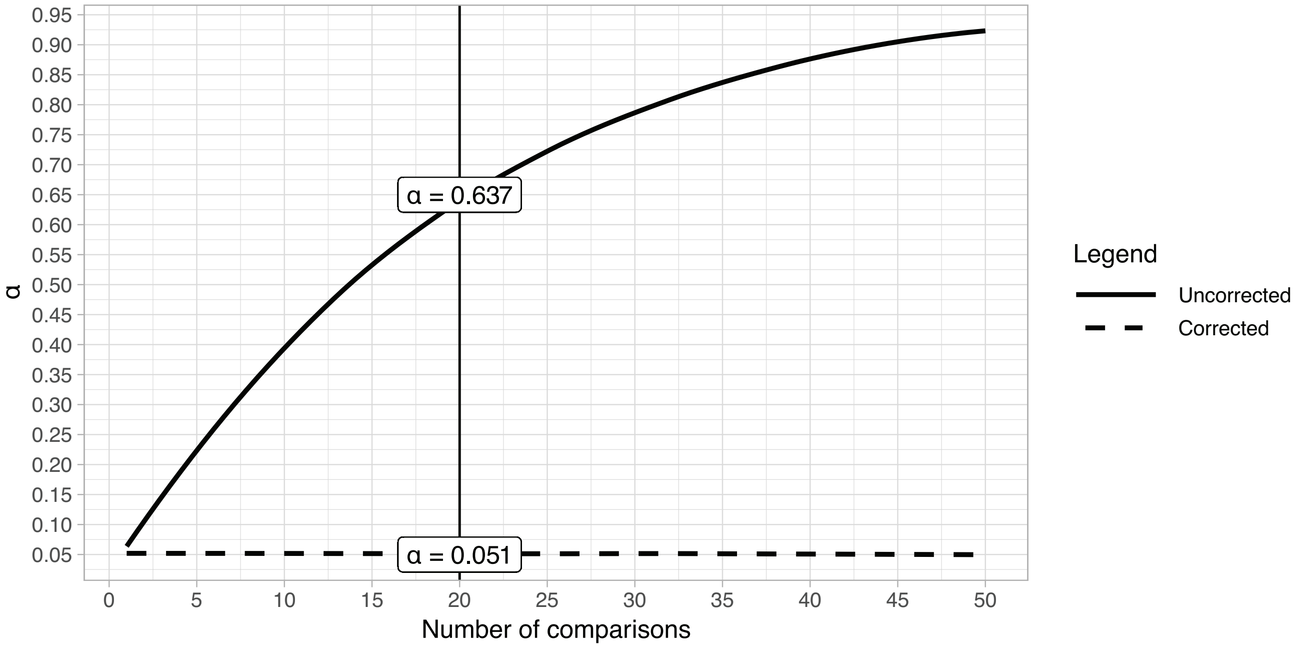

Figure 4. Plot demonstrating the multiple comparisons problem in Kim and James (2015) by way of simulation. These data were generated from 10,000 replications of 50 studies that each computed between 1 and 50 correlation coefficients (r) in the same study, respectively. These correlations were computed from a sample of $N = 17$ and drawn from a bivariate normal distribution where the population parameter (ρ) is zero. Because $\rho = 0$, any statistically significant correlation is a false positive (i.e., a Type I error). The y axis quantifies the false positive rate (α), and the x axis quantifies the number of correlations performed in any individual study. The solid curved line represents the long-run false positive rate (uncorrected) in each of the 50 studies. The false positive rate increases with increasing numbers of comparisons. For example, when comparisons (hereby: i) = 1, $\alpha = .05$; $i = 2$, $\alpha = .10$; $i = 3$, $\alpha = .14$; $i = 4$, $\alpha = .19$, and so on. In Kim and James (2015) a minimum of 20 comparisons must be performed to test a single hypothesis, which increases α from .05 to .64 (see vertical line and upper α label). When the same simulation is performed with Bonferroni corrections, the long-run α is restricted to approximately .05 regardless of how many comparisons are made (see dotted horizontal line).

Winter & Dodou, 2015; Ioannidis, 2008; Nelson et al., 2018; Simmons et al., 2011; Simonsohn, Nelson, & Simmons, 2014; Wicherts et al., 2016). Opportunistic discretion in the decisions that occur at each step of the research process can increase the probability of attaining sufficiently small p values in favor of the existence of an effect, even when none exists. That is, researcher degrees of freedom can substantially increase the probability of reporting false positives (i.e., Type I errors). This phenomenon is therefore an important contributor to the production of research findings that do not replicate. In the present study, Kim and James (2015) adopt a research design that has been described as highly prone to bias due to researcher degrees of freedom: collection of multiple dependent measures of the same construct. The decision to measure five alternate forms of aggression creates multiple opportunities for observing a statistical relationship between these data and the fMRI data. In combination with two fMRI contrasts yielding average t scores in four significantly active regions of interest, this study demands a minimum of 20 Pearson correlation analyses to assess a single primary research question. On the basis of this chosen research design, scholars of organizational neuroscience should recognize that it is unsurprising that at least one of the 20 correlation tests would be significantly different from zero.

When multiple opportunities exist to reject a single hypothesis of no relationship, the Type I error under the null hypothesis is inflated from $\alpha = .05$ to $\alpha = 1-(1-.05)^i$, where i refers to the number of comparisons in a single hypothesis family (Veazie, 2006). Because 20 tests have been performed in Kim and James (2015), the probability of incorrectly rejecting the null hypothesis was therefore approximately 64% instead of 5% (see Figure 4 for a confirmation of this computation via a simulation approach). In this case, because a single positive finding was sufficient to reject the null, one option for returning Type I error to an acceptable 5% threshold would be to perform a Bonferroni adjustment by dividing the α criterion by the number of tests performed (see Figure 4). Alternatively, if this study was construed as an exploratory investigation, and missing an effect was of prime concern (i.e., making a Type II error), no adjustment to α may be necessary. Instead, reporting of all tests alongside their 95% confidence intervals would be informative to provide a descriptive estimation of the range of possible effects (Lee, Whitehead, Jacques, & Julious, 2014, 41). This would then require assessment in a future confirmatory study of sufficient power with multiple comparison adjustment. Although, the informativeness of an exploratory correlational study with such a small sample size may still be questionable. For example, if we construct a 95% confidence interval on the correlation between average t scores extracted from the calcarine sulcus and property aggression, we obtain: $r(15) = -.49$, 95% CI [-0.79, -0.01], $p = .046$. The precision of this estimate is so poor that the plausible range for the population parameter covers almost all negative correlation values, and approaches zero.

Correlational analyses with such small samples (even in exploratory studies) are also rarely desirable for another reason: there is a high probability that correlation coefficient will be inflated (Button et al., 2013; Ioannidis, 2008; Yarkoni, 2009). In a correlation study with a sample of 17 participants and an α threshold of .05, the critical r value is ± .48. That is, any given correlation will only be statistically significant if the sample correlation is greater than $r = .48$ or less than $r = -.48$. If the population correlation between any two measures is $\rho = .30$, for example, a study of this sample size will systematically inflate any significant r estimate to a minimum of .48. The reason for this inflation may be due to one of a number of researcher degrees of freedom, but is also possible simply as a result of random sampling error. Correlational studies in small samples can therefore expect massive inflations of statistically significant r values for all but very large population effects. This is sometimes referred to as the “winner’s curse”: scientists lucky enough to discover a statistically significant finding in a small sample study are likely to overestimate its magnitude by chance (Button et al., 2013; Ioannidis, 2008). Indeed, artificial inflation of the correlation coefficient is a highly plausible explanation for Kim and James’ (2015) single significant correlation. The critical r for statistical significance in this study was ± .48 ($p = .050$) and the single reported significant r was -.49 ($p = .046$). In contrast to the author’s claim that an equivalent or even larger effect should be observed with a larger sample, their correlation may be substantially inflated or simply an artefact of sampling error.

Summary. Researchers can (and often) make what appear to be reasonable design and analysis decisions that can increase researcher degrees of freedom. We therefore have no reason to believe that Kim and James (2015) was purposefully designed to draw out significant findings. However, it is unfortunate that this work is discussed at length in Waldman and colleagues’ (2017) major review piece with no consideration of the inevitability of reporting at least one statistically significant result. The decision to guide organizational research and practice decisions on the basis of a study that had a 64% probability of falsely rejecting one or more tests is a subjective judgement. While this was deemed reasonable by Waldman and colleagues (2017), it may be considered inappropriate by researchers and organizational practitioners who are risking scarce time, effort, and financial resources with respect to investment in organizational neuroscience practices.

Wicherts et al. (2016) provide an extensive 34-item checklist of different degrees of freedom that researchers have available in formulating hypotheses, and in the design, analysis, and reporting of research results. We recommend that scholars of organizational neuroscience familiarize themselves with these items and be vigilant of researcher degrees of freedom in their evaluations of the literature. We also recommend that scholars consider preregistering their own empirical work before conducting a study in order to avoid these problems themselves. Preregistration is the specification of a research design, hypotheses, and analytic strategy ahead of conducting a study (Munafò et al., 2017; Nosek, Ebersole, DeHaven, & Mellor, 2018), and is usually accomplished by posting these specifications to an independent registry that makes it discoverable (e.g., the Open Science Framework: https://osf.io/). Registered reports are a specific type of preregistration that are receiving increasing interest, where peer review occurs prior to data collection (Chambers, Dienes, McIntosh, Rotshtein, & Willmes, 2015; Nosek & Lakens, 2014). In registered reports, high quality protocols are assessed on their methods and analytic strategy, and are provisionally accepted for publication regardless of the magnitude, direction, and statistical significance of an experimental result (so long as the authors follow through with the registered methodology). The use of preregistration and registered reports is likely to substantially reduce questionable research practices in organizational neuroscience, and address many of the concerns we have raised throughout this review.

Conclusion

In 2013, Ashkanasy and colleagues appealed to scholars to not ‘throw the baby out with the bathwater’ in

Table 3
Summary of recommendations for improving post-publication review in organizational neuroscience

	Recommendation	Summary
1	Evaluate the transparency and completeness of reporting in an empirical work before accepting its claims.	• A lack of transparency and completeness in reporting makes it difficult to evaluate whether aspects of the study were valid, reliable, or implemented correctly, and whether interpretations are substantiated by the data. Consider the APA Working Group on Journal Article Reporting Standards (JARS-Quant; Appelbaum et al., 2018) as a guide for best practices in reporting, among other existing systematized guidelines.
2	Become familiar with common misuses and misinterpretations of null hypothesis significance testing (NHST).	• Improper use of NHST can lead to misleading or erroneous conclusions that are not substantiated by the data. Fallacies of NHST have been described in detail elsewhere (e.g., Nickerson, 2000).
3	Interpret results with an explicit reference to effect size magnitude, accuracy, and precision.	• NHST only provides information about whether an effect is statistically significant and its direction. A more informative interpretation of findings can be had through computation of effect sizes, and construction of confidence intervals to specify accuracy and precision.
4	Consider planning for precision.	• When a *p* value is small but the confidence interval on an effect is wide, we know very little about the effect beyond its direction. Consider planning studies to attain a pre-specified margin of error or confidence interval width and evaluate whether existing studies have been designed in a way that yield sufficiently precise detail about an effect of interest.
5	Ensure appropriate statistical tests have been employed if generalizations have been made from sample to population.	• If seeking to generalize from sample to population, ensure that an appropriate analysis has been conducted. For example, inference beyond the sample in mass univariate fMRI analyses requires a random-effects analysis.
6	Evaluate claims of convergent validity between neuroscience measures and psychometric constructs using a parameter estimation approach.	• The point estimate alone is not enough to specify our confidence in the convergence between a neuroscience and psychometric measure. By computing a confidence interval on measures of convergent validity (e.g., Pearson's correlation) we can specify its accuracy and precision to make a more informed judgement.
7	Consider researcher degrees of freedom in the evaluation of published findings and when designing empirical studies.	• Arbitrary decisions when designing experiments, analysing data, and reporting results can increase the likelihood of false positives. Checklists of researcher degrees of freedom are available for consultation when evaluating and designing studies (e.g., Wicherts et al., 2016).
8	Consider preregistration and registered reports in order to build a replicable and trustworthy organizational neuroscience.	• Preregistration involves specification of a research design, hypotheses, and analytic strategy ahead of conducting a study, and posting these specifications to an independent registry. Registered reports involve peer review of protocols ahead of data collection. These methods may substantially improve the replicability and trustworthiness of the literature.

response to unfavorable critiques of how organizational neuroscience may impact organizational research and practice. Six years on, there is now substantial concern that a number of studies with methodological and/or interpretational problems are being uncritically and habitually recited in multiple reviews of the literature. Whereas Ashkanasy expressed worry that we may lose something valuable by dismissing organizational neuroscience altogether, the research climate may have reversed to an extent that all organizational neuroscience work is now considered valuable, indiscriminately.

Waldman and colleagues (2017) is a recent *Annual Review* that has been presented as a critical evaluation of the state-of-the-art in organizational neuroscience. For this reason, the references cited therein may therefore wield a disproportionate impact on the future of the field and influence organizational practice and investment decisions. Because we seek the development of a replicable, reliable, and trustworthy organizational neuroscience moving forward, in this commentary we therefore have provided a comprehensive post-publication review of one-third of all works evaluated in Waldman and colleagues (2017). In doing so, we have identified several research themes that we propose scholars must engage with in future evaluations of the literature. These include: (1) evaluation of the transparency and completeness of an empirical work before accepting its claims, (2) familiarization with misuses or misconceptions surrounding NHST, including the interaction fallacy, (3) interpreting results with explicit reference to effect size magnitude, accuracy, and precision, (4) considering planning analyses for precision so that we can make an informed judgement about an empirical finding, (5) using appropriate statistical tests that allow for generalizability from sample to population, (6) using parameter estimation to evaluate claims of convergent validity between neuroscience and psychometric measures, (7) considering researcher degrees of freedom when evaluating published findings and designing empirical studies, and (8) considering preregistration of studies and registered reports in the interest of a replicable and trustworthy organizational neuroscience. We summarize these recommendations in Table 3.

Organizational neuroscience has emerged as a field because it has been theorized that assessing organizing behavior at multiple levels of analysis, including the neural level, will be a valuable endeavor for organizational theory and practice (e.g., see Healey & Hodgkinson, 2014). We endorse this conclusion. However, a long-acknowledged concern in organizational behavior research is that theories based on studies with fundamental limitations can sometimes persist, propagate, and motivate organizational practice and the behavior of individuals for decades (Ghoshal, 2005; Lindebaum & Zundel, 2013). It is beyond the scope of this commentary to detail examples of studies that would exemplify best practices, but we provide some recommendations and refer multiple times to comprehensive guidelines that describe what such studies would look like (e.g., Cumming, 2008, 2014; Cumming & Maillardet, 2006; Nichols et al., 2017; Nichols et al., 2016; Simmons et al., 2011; Wicherts et al., 2016).

Science is a cumulative and self-corrective endeavor. As organizational neuroscience matures, the trustworthiness of its literature should gradually improve as findings are found to be credible or are refuted and the scientific record is corrected. This process is most efficient when empirical works are reported with sufficient transparency and completeness to allow for critical evaluation, and when scholars are consistently applying a critical eye to the existing literature. Post-publication review, which in some cases challenges the conclusions of published work, will play an important part in the development of sound theory and organizational practice decisions that may emerge as a result of organizational neuroscience.

Conflict of interest

All authors declare they have no conflict of interest.

Funding

This work was partly supported by a Melbourne Research Scholarship awarded to G.A.P. by the University of Melbourne, and a Heart Foundation Future Leader Fellowship (1000458) awarded to P.M.

Author contributions

G.A.P. conceived the idea for this commentary, critically evaluated each publication, performed the analyses, generated the figures, scripted all analyses in R, wrote the paper, and compiled the paper in LaTeX. W.R.L provided guidance on framing the paper and cross-checking the analyses. W.R.L., S.B., H.Z., and P.M. contributed to the paper development, editing, and provided feedback.

Open Science Practices

This type of article does not have any associated data or materials to be shared. The R statistical software analysis scripts and instructions for reproducing all analyses, simulations, and figures are available on GitHub at: https://github.com/gprochilo/org_neuro_com. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

References

Abelson, R. (1995). *Statistics as principled argument*. Taylor & Francis Inc.

Academy of Management. (2018). Academy of Management Perspectives. Accessed on May 9, 2018. Retrieved from https://journals.aom.org/journal/amp

Altman, D. G. & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ (Clinical research ed.) 332*(7549), 1080–1080. doi:10.1136/bmj.332.7549.1080

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. doi:10.1037/amp0000191

Arnsten, A. F. T., Raskind, M. A., Taylor, F. B., & Connor, D. F. (2015). The effects of stress exposure on prefrontal cortex: Translating basic research into successful treatments for post-traumatic stress disorder. *Neurobiology of Stress*, *1*, 89–99. doi:10.1016/j.ynstr.2014.10.002

Ashkanasy, N. M. (2013). Neuroscience and leadership: Take care not to throw the baby out with the bathwater. *Journal of Management Inquiry*, *22*(3), 311–313. doi:10.1177/1056492613478519

Ashkanasy, N. M., Becker, W. J., & Waldman, D. A. (2014). Neuroscience and organizational behavior: Avoiding both neuro-euphoria and neurophobia. *Journal of Organizational Behavior*, *35*(7), 909–919. doi:10.1002/job.1952

Bass, B. M. & Avolio, B. J. (1990). *Transformational leadership development: Manual for the Multifactor Leadership Questionnaire.* Menlo Park, CA: Mind Garden.

Becker, W. J. & Menges, J. I. (2013). Biological implicit measures in HRM and OB: A question of how not if. *Human Resource Management Review*, *23*(3), 219–228. doi:10.1016/j.hrmr.2012.12.003

Becker, W. J., Volk, S., & Ward, M. K. (2015). Leveraging neuroscience for smarter approaches to workplace intelligence. *Human Resource Management Review*, *25*(1), 56–67. doi:10.1016/j.hrmr.2014.09.008

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. doi:10.1038/s41562-017-0189-z

Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., … Olmstead, R. (2004). Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human–Computer Interaction*, *17*(2), 151–170. doi:10.1207/s15327590ijhc1702_3

Berka, C., Levendowski, D., Westbrook, P., Davis, G., N Lumicao, M., Olmstead, R., … K Ramsey, C. (2005). EEG quantification of alertness: Methods for early identification of individuals most susceptible to sleep deprivation. In *Proceedings of the SPIE Defense and Security Symposium, Biomonitoring for Physiological and Cognitive Performance during Military Operations.* (Vol. 5797, pp. 78–89). doi:10.1117/12.597503

Blakemore, S.-J. & Robbins, T. (2012). Decision-making in the adolescent brain. *Nature Neuroscience*, *15*, 1184–91. doi:10.1038/nn.3177

Boyatzis, R. E. & Jack, A. I. (2018). The neuroscience of coaching. *Consulting Psychology Journal: Practice and Research*, *70*(1), 11–27. doi:10.1037/cpb0000095

Boyatzis, R. E., Passarelli, A. M., Koenig, K., Lowe, M., Mathew, B., Stoller, J. K., & Phillips, M. (2012). Examination of the neural substrates activated in memories of experiences with resonant and dissonant leaders. *The Leadership Quarterly*, *23*(2), 259–272. doi:10.1016/j.leaqua.2011.08.003

Boyatzis, R. E., Rochford, K., & Jack, A. I. (2014). Antagonistic neural networks underlying differentiated leadership roles. *Frontiers in Human Neuroscience*, *8*(114), 1–15. doi:10.3389/fnhum.2014.00114

Butler, M. J. R., O'Broin, H. L. R., Lee, N., & Senior, C. (2015). How organizational cognitive neuroscience can deepen understanding of managerial decision-making: A review of the recent literature and future directions. *International Journal of Management Reviews*, 1–18. doi:10.1111/ijmr.12071

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi:10.1038/nrn3475

Calin-Jageman, R. J. & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, *73*(sup1), 271–280. doi:10.1080/00031305.2018.1518266

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*. doi:10.1038/s41562-018-0399-z

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2. doi:https://doi.org/10.1016/j.cortex.2015.03.022

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psycholog-*

ical Science, *3*(4), 286–300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The New Statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi:10.1177/0956797613504966

Cumming, G. & Calin-Jageman, R. (2016). *Introduction to the New Statistics: Estimation, open science, and beyond*. London: Taylor and Francis.

Cumming, G. & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, *11*(3), 217–27. doi:10.1037/1082-989x.11.3.217

de Winter, J. C. & Dodou, D. (2015). A surge of *p*-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, *3*, 1–44. doi:10.7717/peerj.733

Diedenhofen, B. & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, *10*(4), 1–12. doi:10.1371/journal.pone.0121945

Elsevier. (2018). Organizational dynamics. Accessed May 9, 2018. Retrieved from https://www.journals.elsevier.com/organizational-dynamics

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*(4), 379–390. doi:10.1037/1082-989X.1.4.379

Gelman, A. & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. doi:10.1198/000313006X152649

Ghoshal, S. (2005). Bad management theories are destroying good management practices. *Academy of Management Learning & Education*, *4*(1), 75–91. Retrieved from http://ezproxy.lib.monash.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=16132558&site=ehost-live&scope=site

Goldman-Rakic, P. S. (1996). The prefrontal landscape: Implications of functional architecture for understanding human mentation and the central executive. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1346), 1445–53. doi:10.1098/rstb.1996.0129

Grech, R., Cassar, T., Muscat, J., Camilleri, K. P., Fabri, S. G., Zervakis, M., . . . Vanrumste, B. (2008). Review on solving the inverse problem in EEG source analysis. *Journal of NeuroEngineering and Rehabilitation*, *5*(25), 1–33. doi:10.1186/1743-0003-5-25

Healey, M. P. & Hodgkinson, G. P. (2014). Rethinking the philosophical and theoretical foundations of organizational neuroscience: A critical realist alternative. *Human Relations*, *67*(7), 765–792. doi:10.1177/0018726714530014

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), 0696–0701. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–8. doi:10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–54. doi:10.1177/1745691612464056

Janak, P. H. & Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature*, *517*(7534), 284–292. doi:10.1038/nature14188

Kelley, K. & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–85. doi:10.1037/1082-989x.11.4.363

Kim, M. Y. & James, L. R. (2015). Neurological evidence for the relationship between suppression and aggressive behavior: Implications for workplace aggression. *Applied Psychology*, *64*(2), 286–307. doi:10.1111/apps.12014

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., . . . Lazarević, L. B., et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. doi:10.1177/2515245918810225

Lang, P., Bradley, M., & Cuthbert, B. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual.* University of Florida.

Lee, E. C., Whitehead, A. L., Jacques, R. M., & Julious, S. A. (2014). The statistical interpretation of pilot trials: Should significance thresholds be reconsidered? *BMC Medical Research Methodology*, *14*, 1–8. doi:10.1186/1471-2288-14-41

Leslie, K. J., George, L., & Drazen, P. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. doi:10.1177/0956797611430953

Lindebaum, D. (2013). Pathologizing the healthy but ineffective: Some ethical reflections on using neuroscience in leadership research. *Journal of Management Inquiry*, *22*(3), 295–305. doi:10.1177/1056492612462766

Lindebaum, D. (2016). Critical essay: Building new management theories on sound data? the case of neuroscience. *Human Relations*, *69*(3), 537–550. doi:10.1177/0018726715599831

Lindebaum, D. & Zundel, M. (2013). Not quite a revolution: Scrutinizing organizational neuroscience in leadership studies. *Human Relations*, *66*(6), 857–877. doi:10.1177/0018726713482151

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. doi:10.1037//1082-989X.7.1.19

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–63. doi:10.1146/annurev.psych.59.103006.093735

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245. doi:10.1080/00031305.2018.1527253

Merton, R. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

Molenberghs, P., Cunnington, R., & Mattingley, J. (2012). Brain regions with mirror properties: A meta-analysis of 125 human fmri studies. *Neuroscience & Biobehavioral Reviews*, *36*(1), 341–349. doi:10.1016/j.neubiorev.2011.07.004

Molenberghs, P., Prochilo, G., Steffens, N. K., Zacher, H., & Haslam, S. A. (2017). The neuroscience of inspirational leadership: The importance of collective-oriented language and shared group membership. *Journal of Management*, *43*(7), 2168–2194. doi:10.1177/0149206314565242

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(0021), 1–9. doi:10.1038/s41562-016-0021

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*(1), 511–534. doi:10.1146/annurev-psych-122216-011836

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., . . . Yeo, B. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, *20*(3), 299–303. doi:10.1038/nn.4500

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., . . . Yeo, B. T. T. (2016). Best practices in data analysis and sharing in neuroimaging using MRI. *bioRxiv*, 1–71. Retrieved from https://www.biorxiv.org/node/14915.abstract

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. doi:I0.1037//1082-989X.S.2.241

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107. doi:10.1038/nn.2886

Niven, K. & Boorman, L. (2016). Assumptions beyond the science: Encouraging cautious conclusions about functional magnetic resonance imaging research on organizational behavior. *Journal of Organizational Behavior*, *37*(8), 1150–1177. doi:10.1002/job.2097

Nofal, A. M., Nicolaou, N., Symeonidou, N., & Shane, S. (2017). Biology and management: A review, critique, and research agenda. *Journal of Management*, *44*(1), 7–31. doi:10.1177/0149206317720723

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. doi:10.1073/pnas.1708274114

Nosek, B. A. & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. doi:10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-1-8. doi:10.1126/science.aac4716

Penny, W. D. & Holmes, A. J. (2007). Random effects analysis. In K. Friston, J. Ashburner, S. J. Kiebel, T. E. Nichols, & W. D. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images*. Jordan Hill, UNITED KINGDOM: Elsevier Science Technology. doi:10.1016/B978-0-12-372560-8.X5000-1

Pernet, C., Wilcox, R., & Rousselet, G. (2013). Robust correlation analyses: False positive and power validation using a new open source Matlab toolbox. *Frontiers in Psychology*, *3*(606), 1–18. doi:10.3389/fpsyg.2012.00606

Peters, G. Y. & Crutzen, R. (2017). Knowing exactly how effective an intervention, treatment, or manipulation is and ensuring that a study replicates: Accuracy in parameter estimation as a partial solution to the replication crisis. *Preprint; PsyArXiv*, 1–31. Retrieved from https://doi.org/10.31234/osf.io/cjsk2

Peters, G. Y., Verboon, P., & Green, J. (2018). Userfriendlyscience: Quantitative analysis made accessible. R package version 0.7.2. Computer Program. Retrieved from https://cran.r-project.org/package=userfriendlyscience

Peterson, S. J., Balthazard, P. A., Waldman, D. A., & Thatcher, R. W. (2008). Neuroscientific implications of psychological capital: Are the brains of optimistic, hopeful, confident, and resilient leaders different? *Organizational Dynamics*, *37*(4), 342–353. doi:10.1016/j.orgdyn.2008.07.007

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. doi:10.1016/j.tics.2005.12.004

Popper, K. (1962). *Conjectures and refutations: The growth of scientific knowledge*. Basic Books.

Rich, B. L., Lepine, J. A., & Crawford, E. R. (2010). Job engagement: Antecedents and effects on job performance. *Academy of Management Journal*, *53*(3), 617–635. Retrieved from http://amj.aom.org/content/53/3/617.abstract

Robbins, T. W. (1996). Dissociating executive functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1346), 1463–71. doi:10.1098/rstb.1996.0131

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*(1), 127–41. doi:10.1002/sim.2331

Senn, S. (2005). Dichotomania: An obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. In *International Statistical Istitute 55th Session*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–66. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–47. doi:10.1037/a0033242

Urigüen, J. A. & Garcia-Zapirain, B. (2015). EEG artifact removal—state-of-the-art and guidelines. *Journal of Neural Engineering*, *12*(3), 1–43. doi:10.1088/1741-2560/12/3/031001

Vanhove, J. (2018). cannonball: Tools for teaching statistics. R package version 0.0.0.9000. Computer Program. Retrieved from http://janhove.github.io/teaching/2018/09/26/cannonball

Veazie, P. J. (2006). When to combine hypotheses and adjust for multiple tests. *Health Services Research*, *41*(3 Pt 1), 804–818. doi:10.1111/j.1475-6773.2006.00512.x

Waldman, D. A., Balthazard, P., & Peterson, S. (2011a). Leadership and neuroscience: Can we revolutionize the way that inspirational leaders are identified and developed? *The Academy of Management Perspectives*, *25*(1), 60–74. doi:10.5465/amp.25.1.60

Waldman, D. A., Balthazard, P., & Peterson, S. (2011b). Social cognitive neuroscience and leadership. *The Leadership Quarterly*, 1092–1106. doi:10.1016/j.leaqua.2011.09.005

Waldman, D. A. & Balthazard, P. A. (2015). *Organizational neuroscience*. Monographs in Leadership and Management. Emerald Group Publishing Limited. doi:10.1108/S1479-357120150000007017

Waldman, D. A., Stikic, M., Wang, D., Korszen, S., & Berka, C. (2015). Neuroscience and team processes. In *Organizational neuroscience* (Chap. 12, Vol. 7, pp. 277–294). Monographs in Leadership and Management. Emerald Group Publishing Limited. doi:10.1108/S1479-357120150000007012

Waldman, D. A., Wang, D., & Fenters, V. (2016). The added value of neuroscience methods in organizational research. *Organizational Research Methods*, 1–27. doi:10.1177/1094428116642013

Waldman, D. A., Wang, D., Stikic, M., Berka, C., Balthazard, P. A., Richardson, T., ... Maak, T. (2013a). Emergent leadership and team engagement: An application of neuroscience technology and methods. In *Academy of management annual meeting proceedings* (Vol. 2013, pp. 632–637). Academy of Management. doi:10.5465/AMBPP.2013.63

Waldman, D. A., Wang, D., Stikic, M., Berka, C., Balthazard, P. A., Richardson, T., ... Maak, T. (2013b). Emergent leadership and team engagement: An application of neuroscience technology and methods. *Preprint; ResearchGate*, 1–33. Retrieved from https://www.researchgate.net/publication/259678311_Emergent_Leadership_and_Team_Engagement_An_Application_of_Neuroscience_Technology_and_Methods

Waldman, D. A., Ward, M., & Becker, W. J. (2017). Neuroscience in organizational behavior. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 425–444. doi:10.1146/annurev-orgpsych-032516-113316

Ward, M. K., Volk, S., & Becker, W. J. (2015). An overview of organizational neuroscience. In D. Waldman & P. Balthazard (Eds.), *Organizational neuroscience* (Chap. 2, Vol. 7, pp. 17–50). Emerald

Group Publishing Limited. doi:10.1108/S1479-357120150000007001

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, *7*, 1832–1832. doi:10.3389/fpsyg.2016.01832

Woodson, M. I. C. E. (1969). Parameter estimation vs. hypothesis testing. *Philosophy of Science*, *36*(2), 203–204. doi:10.1086/288247

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—commentary on vul et al. (2009). *Perspectives on Psychological Science*, *4*(3), 294–298. Retrieved from http://pps.sagepub.com/content/4/3/294.abstract

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. doi:10.1037/1082-989x.12.4.399