Pooling resources to enhance rigour in psychophysiological research

Saunders, Blair; Inzlicht, Michael

Published in:
International Journal of Psychophysiology

DOI:
10.1016/j.ijpsycho.2021.01.018

Publication date:
2021

Pooling resources to enhance rigour in psychophysiological research: Insights from open science approaches to meta-analysis

Blair Saunders[1] and Michael Inzlicht[2,3]

[1] School of Social Sciences, University of Dundee, UK

[2] Department of Psychology, University of Toronto

[3] Rotman School of Management, Canada

Corresponding Author:
Blair Saunders,
Scrymgeour Building,
University of Dundee,
Dundee, DD1 4HN
Email: b.z.saunders@dundee.ac.uk

*****accepted for publication at International Journal of Psychophysiology. Final version subject to copy editing**

7,709 words (excluding abstract)

Recent years have witnessed calls for increased rigour and credibility in the cognitive and behavioural sciences, including psychophysiology. Many procedures exist to increase rigour, and among the most important is the need to increase statistical power. Achieving sufficient statistical power, however, is a considerable challenge for resource intensive methodologies, particularly for between-subjects designs. Meta-analysis is one potential solution; yet, the validity of such quantitative review is limited by potential bias in both the primary literature and in meta-analysis itself. Here, we provide a non-technical overview and evaluation of open science methods that could be adopted to increase the transparency of novel meta-analyses. We also contrast *post hoc* statistical procedures that can be used to correct for publication bias in the primary literature. We suggest that traditional meta-analyses, as applied in ERP research, are exploratory in nature, providing a range of plausible effect sizes without necessarily having the ability to confirm (or disconfirm) existing hypotheses. To complement traditional approaches, we detail how prospective meta-analyses, combined with multisite collaboration, could be used to conduct statistically powerful, confirmatory ERP research.

Keywords: open-science; meta-analysis; statistical power; ERPs; cognitive neuroscience;

Pooling resources to enhance rigour in psychophysiological research: Insights from open science approaches to meta-analysis

Recent years have witnessed a call for increased rigour, credibility, and transparency in the methods used to create, synthesize, and communicate science. This credibility revolution (Vazire, 2018) has been motivated in part by findings that results are often unreliable because they are published selectively (Ferguson & Brannick, 2012), derived from underpowered statistical analyses (Rossi, 1990; Stanley, Carter, & Doucouliagos, 2018), and because questionable research practices (QRPs) drive unacceptably high false-positive rates (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). Multiple open science methods have been proposed to enhance rigour, including preregistration, increasing statistical power, encouraging replication, the free sharing materials and data, and new publishing formats that accept articles based on the soundness of their question and methods prior to data collection and analysis (i.e., Registered Reports: Chambers, 2013; Nosek & Lakens, 2014).

Recent reviews indicate that low statistical power is a particularly acute problem in cognitive neuroscience (Button et al., 2013; Clayson, Carbine, Baldwin, & Larson, 2019; Clayson, Carbine, & Larson, 2020; Szucs & Ioannidis, 2017). Szucs and Ioannidis (2017) estimated statistical power for over 25,000 statistical tests in 3,801 cognitive neuroscience and psychology papers, and found that studies achieve 12%, 44%, and 73% statistical power for small, medium, and large effect sizes, respectively (see Clayson et al., 2019 for similar results for ERP research). Another review estimated that statistical power was as low as 20% in ERP studies of feedback processing in depression (Clayson et al., 2020)—this finding indicates that even if the underlying hypothesis was true, that 8/10 studies testing this hypothesis should return null results (i.e.,

false negatives). Equally troubling, though less appreciated, is that low power can also increase the false discovery rate (i.e., false positives), that is the rate of significant findings that are in fact false (Krzywinski & Altman, 2013). Together, these findings suggest that the credibility of cognitive neuroscience as a discipline might critically depend on increasing statistical power.

Particularly strong barriers to increased statistical power exist in fields that are resource intensive, such as cognitive neuroscience. Conducting a .9 powered study for a small-to-medium sized difference between means (i.e., Cohen's *d* = .4), for example, would require 68 participants for a within-subjects test, and 266 participants for a between subjects test. We suspect that these sample sizes, particularly between-subjects, are unachievable for all but the best resourced neuroscience laboratories. Even well-resourced laboratories would necessarily reduce their research output if they were to power their studies to this level. This issue is likely intensified for the recruitment of harder to reach populations (e.g., patients, ethnic/racial minorities, and infants), meaning that calls to increase sample sizes might unintentionally restrict breadth and generalizability by shifting research towards easy to access populations and scalable methodologies (Lakens et al., 2018). Despite these legitimate concerns, the nature of null hypothesis significance testing, as well as the need to quantify effect sizes with precision, means that increasing sample sizes is a valuable goal. Here, we evaluate meta-analytical methods as a solution to increasing statistical power by pooling data across multiple laboratories. We focus specifically on ERP research; however, much of our analysis would likely apply equally to other resource intensive cognitive neuroscience methods (Elliott et al., 2020) or to other measures of peripheral psychophysiology.

Prior reviews have detailed how meta-analyses could facilitate rigorous individual difference ERP studies (Moran et al., 2017). However, the validity of conclusions drawn from meta-analyses are susceptible to multiple forms of bias, both in the primary literature and in the production of the meta-analyses themselves. Such bias requires close consideration before concluding that meta-analyses enhance rigour. First, as an initial defence against this bias, we summarize open science methods that can make meta-analyses transparent from conception. Subsequently, we review *post hoc* statistical techniques to detect and correct for publication bias arising from the primary research. Finally, we detail how prospective meta-analyses, in combination with a collective, multi-site approach to gathering psychophysiological data, could drive stronger, confirmatory inferences in ERP research.

Meta-analysis in ERP research

Meta-analyses combine effects from multiple studies testing the same theoretical question, resulting in a meta-analytic effect size that represents a weighted average of included studies (Rosenthal & DiMatteo, 2002). Meta-analyses are intended to facilitate cumulative science by providing an objective measure of consistency across studies (i.e., the meta-analytical effect size), while diminishing nonspecific error between smaller studies (Borenstein, Hedges, Higgins, & Rotherstein, 2011). In contrast to smaller individual studies that often produce wide confidence intervals around an effect size, appropriately conducted meta-analyses can draw on the power of their large data-sets to more precisely estimate the underlying effect size. This apparent power to summarize what is known means that meta-analyses are often given considerable weight when developing new studies, grants, or public policy (e.g., Hunter & Schmidt, 1996).

Meta-analysis has a clear appeal in the context of ERP research where individual laboratories are often limited in their ability to collect large data-sets. There are already six meta-analyses focusing on the relationship between trait anxiety and error-related ERPs (i.e., the error-related negativity, ERN; Cavanagh & Shackman, 2015; Moser, Moran, Kneip, Schroder, & Larson, 2016; Moser, Moran, Schroder, Donnellan, & Yeung, 2013; Pasion & Barbosa, 2019; Riesel, 2019; Saunders & Inzlicht, 2020), and other meta-analyses have focused on the P300 and schizophrenia (Jeon & Polich, 2001); and the face-related N170 and autism (Kang et al., 2018). Here, meta-analyses are particularly useful for between-subject's contrasts that are notoriously noisy in EEG research (Luck, 2014). In addition to confirming established empirical effects, many analyses use meta-regression to test novel hypotheses (e.g., gender differences; Moser et al., 2016), further highlighting the power of meta-analyses to reveal effects otherwise hidden in small, individual studies.

It cannot be taken as given that any meta-analysis is necessarily rigorous. Meta-analyses involve highly multi-dimensional data sets, requiring many decisions in their production. Consequently, meta-analytic reviews are susceptible to many sources of publication bias and questionable research practices (Lakens, Hilgard, & Staaks, 2016). Selectively reporting significant or large results, while omitting small and non-significant results, for example, can give a false impression that the meta-analytical effect is large and robust. Other sources of bias include selectively reporting moderators based on their statistical significance, including dependent effect sizes to increase the sample-size of a meta-analysis, or not accounting for the inflationary influence of publication bias on the meta-analytic effect size (Sterne, Egger, & Smith, 2001; Thornton & Lee, 2000; Williamson, Gamble, Altman, & Hutton, 2005).

Left unchecked, the combined influence of questionable research practices and publication bias mean that meta-analyses will often provide effect sizes that are unreliably inflated (Pereira & Ioannidis, 2011). If a field has even a modest sized file-drawer of studies that do not support an established hypothesis, any meta-analysis will be blind to these studies, necessarily inflating the average effect size. In addition to bias in the primary literature, the potential of the meta-analyst to steer the review towards specific outcomes have led some researchers to seriously question if meta-analysis can ever truly resolve disputes between opposing ideological positions (Ferguson, 2014). This charge is in stark contrast to the occasional valorisation of meta-analyses as tools to find truth among seemingly contradictory findings (Hunter & Schmidt, 1996).

Providing definitive, irrefutable evidence about the base truth of a prediction is an unrealistically difficult test for any methodology, not least because it will never be possible to satisfy every critic—even the best conducted review is susceptible to acrimonious and/or ad hominem counter arguments (Ferguson, 2015). Furthermore, central to any scientific discipline is the need to make inferences based on the generation of cumulative knowledge, and scientists will continue to do this with or without meta-analyses. Thus, it would be a non-solution, to dismiss meta-analysis entirely due to challenges to validity. Rather than viewing meta-analyses as either credible or not, we take the stance that it is more fruitful to accept that there are a range factors that influence the credibility of a meta-analysis that should be considered when conducting a new meta-analytic review, or when consuming a published meta-analysis.

In the following, we use the relationship between anxiety and the error-related negativity (ERN) to illustrate how open science methods can be used to enhance meta-analyses in ERP research. The ERN is a negative-going deflection in the response locked ERP that peaks at frontocentral electrodes within 100 ms after mistakes and is putatively generated by the anterior midcingulate cortex (Gehring, Liu, Orr, & Carp, 2012). Multiple studies have indicated that this component is increased in anxious samples (Hajcak, 2012), with increased reactivity to mistakes suggested as a potential biomarker for anxious psychopathology (Meyer, 2017; Weinberg, Dieterich, & Riesel, 2015). The anxiety-ERN relationship is a useful case-study for several reasons. Foremost, six meta-analytic reviews have already been conducted on this hypothesis in the past 7 years, indicating the prominence of this hypothesis in the field. Taking a concrete example also allows us to compare and contrast the influence of different methodological decisions on meta-analytic conclusions in ERP research.

Open science methods to enhance novel meta-analyses

Many steps that can improve the credibility and rigour of meta-analyses can be taken during the production of the meta-analyses itself, by publicly declaring a protocol for the production of the review in advance, and by publishing the meta-analysis in a transparent manner that facilitates the complete understanding, verification, and re-use of the meta-analytic data.

Preregistration

The many decisions made to produce a meta-analysis means that there is not one inevitable analysis that emerges from the literature, but, instead, the process of searching,

coding, and analysing data leads the researcher to construct only one of many potential meta-analytic reviews from a set of studies. There are justifiable reasons why two meta-analyses on the same topic might differ (e.g., excluding vs. including clinical samples; Moser et al., 2013; Cavanagh & Shackman, 2015). Other differences emerge through decision points that are less germane to the specific research question (e.g., selecting among possible effect sizes, coding moderators, exclusion criteria). From the outside, it is often impossible to verify if these such decisions were taken with or without knowledge of their impact on the outcome.

Preregistration provides a solution to this garden-of-forking paths for two related reasons (Quintana, 2015). First, a sound meta-analysis relies on the precise formulation of a research question. Preregistration ensures that the author starts upfront with a well-formulated research question that can constrain subsequent methodological steps, while ensuring that the hypotheses do not shift after the results are known, itself a questionable research practice (Kerr, 1998). Second, the preregistration should provide an analysis plan that at least makes a transparent distinction between *a priori* confirmatory analyses and *post hoc* exploratory analyses, avoiding the potential for analytical flexibility and the cherry-picking of results. Here, preregistration does not aim to eliminate novel analyses or exploratory findings. Instead, the aim is to distinguish between confirmatory, *a priori* hypotheses and exploratory analyses to avoid potentially questionable research practices, such as hypothesizing after the results are known (i.e., HARKing; Kerr, 1998).

Preregistration was not referenced in any of the six meta-analyses on the relationship between anxiety and performance monitoring. Considering the range of decisions necessary to define this research question, both variables (i.e., anxiety and performance monitoring) could

be defined either broadly or narrowly. Performance monitoring could refer to multiple ERPs with comparable neural generators and functional significance (e.g., ERN, N2, feedback-related negativity; cf., Yeung, Botvinick, & Cohen, 2004), or, as is more often the case, performance monitoring could focus exclusively on the ERN. Anxiety could cover a broad range of clinical and non-clinical diagnoses and traits, or could focus more exclusively on specific diagnoses (e.g., generalised anxiety disorder, obsessive compulsive disorder). Without preregistration it is impossible to know if these meta-analyses started with a broader research question that was narrowed based on the results, or if moderators (e.g., comparing clinical and non-clinical anxiety) were hypothesized *a priori*, or were included in the final report after results were known. Here, we do not wish challenge the validity of existing meta-analyses on the anxiety-ERN relationship, but merely highlight that the *a priori* nature of the hypothesis tests cannot be verified.

Preregistering a meta-analysis may at first seem like an unwieldy task. However, extensive evidence-based reporting standards have been developed for systematic reviews and meta-analyses, such as the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA; Moher, Liberati, Tetzlaff, & Altman, 2009), or the Meta-Analysis Reporting Standards (MARS; Cooper, 2010). Checklists are freely available online for both sets of standards, and deciding as many of these as steps as possible *a prioiri* can guide the comprehensive preregistration of a novel meta-analysis. Criteria that can be decided in advance include rules for study inclusion and exclusion, search strategy, plans for extracting and collating effect sizes, and specifics about summary statistics. Transparency can be enhanced further by preregistering a formal analysis plan, ideally posting analysis syntax alongside the study preregistration. In

psychology and cognitive neuroscience, it is common to post these registrations to an online repository such as the Open Science Framework (OSF). In addition to the OSF, PROSPERO is an international data-base in the health sciences that provides a template to prospectively register protocols for systematic reviews along multiple dimensions, including the title, research question, population/domain of interest, search criteria, outcome measures, and strategy for data synthesis (Booth et al., 2012). While deciding this array of criteria up-front might seem daunting, it is important to note that each step is eventually required for a successful meta-analysis or systematic review, meaning that in many cases preregistration largely shifts the timeline of work that needs to be done anyway.

Increased credibility is but one benefit of preregistration. The *a priori* construction of a protocol, for example, means that many difficult questions are considered before the labour intensive work of the meta-analysis commences. A concrete registered protocol should increase the efficiency of the review by properly constraining the search and analysis in advance, while building reporting standards into the meta-analysis from the start ensures that authors do not omit important steps, facilitating publication. Lastly, public registration can establish primacy over other meta-analyses, encourage collaboration between similar projects, and/or help to uncover unpublished work to incorporate into the meta-analysis. Indeed, a primary objective of the PROSPERO registration database is to allow researchers to assess if a similar review question is already in progress to avoid unplanned duplication of systematic reviews or meta-analyses (Booth, Clarke, Ghersi, Moher, Petticrew, & Stewart, 2011; Booth et al., 2012). Together, these considerations suggest that preregistration can form an intrinsically valuable component of meta-analysis.

Finally, one specific challenge for preregistering a meta-analysis in ERP research is that even single ERPs can be operationalized in multiple different ways depending on choice of referencing system, electrode site, ERP quantification, number of trials included in averaged ERPs, or baseline selection, and so on. Indeed, many psychometric investigations have explored the influence of these decisions on the reliability and validity of specific ERP components (e.g., for the ERN: Fischer, Klien, & Ullsperger, 2017; Meyer, Riesel, & Hajcak, 2013; Riesel, Weinberg, Endrass, Meyer, & Hajcak, 2013; Sandre, Bancia, Riesel, Flake, Klawohn, & Weinberg, 2020). In the formulation of a plan to extract data for a meta-analysis, it can be useful to review available literature on your ERP of choice to determine a best-practice or gold-standard quantification and use this information to generate principled approaches for selecting statistics to include in your meta-analysis. In many cases, you might be able to implement homogenous selection criteria, for example, if studies report sufficient information to select a specific effect size, or if you can obtain original data for re-analysis through correspondence with an author. In practice, however, there is often a large degree of heterogeneity in quantification between studies, and it would be counterproductive to exclude large amounts of data due to overly narrow criteria for defining an ERP. As such, preregistered criteria might need to strike a programmatic balance, stating the ideal ERP measures that would be extracted wherever possible, while also defining a universe of acceptable and unacceptable analysis protocols based on psychometric evidence to form data inclusion and exclusion criteria, respectively.

Increasing transparency beyond (or without) pre-registration

Preregistration is not the only open science method that facilitates rigorous meta-analysis. In addition, the various analytical steps that make up a meta-analysis are often neither

transparently reported nor reproducible (Lakens et al., 2016; Lakens et al., 2017; Polanin, Hennessy, & Tsjui, 2020). One investigation examined 150 published meta-analyses and reported that just over half of these (55%) included sufficient information for replication, and it was particularly rare to include effect size and moderator information for each study, and rarer still to include analyses scripts to reproduce the meta-analysis from raw data (Polanin et al., 2020). Irreproducibility can also arise through relatively common errors in statistical transformations applied to convert effect sizes onto a common scale (Gøtzsche, Hróbjartsson, Marić, & Tendal, 2007), and in other cases, QRPs have been revealed in registered meta-analyses themselves, including switching outcome measures between *a priori* protocols and the final review (Kirkham, Altman, & Williamson, 2010). This latter finding indicates that preregistration, in and of itself, does not guarantee that QRPs will not occur. Nevertheless, changes in outcome measurement would be entirely concealed without preregistration.

The reproducibility of published meta-analyses is seriously limited by the rarity with which data and analysis syntax are made available openly with the publication. One review estimated that only 1% of meta-analyses shared analysis code with their publication (Polanin et al., 2020). An obvious benefit of sharing syntax and data-sets is that peers can independently reproduce and verify the meta-analysis. Given the prevalence of various sources of error that have been identified in already published meta-analyses (Gøtzsche et al., 2007), the ability to check and correct statistical analyses is essential to allow quality control. Syntax also provides an unambiguous record of the analysis, meaning that readers and reviewers can use this code to aid their comprehension of the methodology of a given meta-analysis. As such, sharing code enhances transparency as well as reproducibility. One challenge when collating data for meta-

analyses is that the empirical reports may contain insufficient data, meaning that some data points in the meta-analysis are obtained through communication with authors. Here, it can be useful to maintain a record of this communication to document the provenance of the data, as well as getting verification from the authors that it is permissible to share this data openly.

Meta-analyses and systematic reviews become out-of-date rapidly, sometimes even before publication (Beller, Chen, Wang, & Glasziou, 2013; Créquit, Trinquart, Yavchitz, & Ravaud, 2016). This might occur because of new studies entering the primary literature, or because there is often a long delay between the end of a literature research and the publication of the meta-analysis, or due to the development of novel statistical techniques to conduct a meta-analysis and correct for bias. Adopting transparent reporting standards facilitates cumulative science by allowing future researchers to add newer studies to open meta-analytic data. To ensure datasets are maximally useful for future scientists, Lakens et al. (2016) recommend meta-analysts share effect sizes, confidence intervals, sample sizes, means, standard deviations, test statistics, and the type of design for each study included in the meta-analysis. To facilitate the transparent reporting of meta-analyses in ERP research specifically, authors should include as a minimum study-level information about analysis electrodes and ERP operationalisation (e.g., peak, mean amplitude), as well as cataloguing other information that might reasonably contribute to the heterogeneity of ERP results (e.g., hardware, electrode numbers, referencing system). Where possible, meta-analysts should aim to minimize between-study heterogeneity—sometimes by contacting primary authors for statistics that more closely match the criteria for the review. We are aware of no meta-analysis that reported this range of parameters for the anxiety-ERN relationship. Thus, there appears to be considerable room for

improvement in reporting standards for meta-anlayses in ERP research, at least as indicated by the anxiety-ERN relationship.

Sharing data and analysis scripts alone, however, is insufficient to allow existing meta-data to be used in future research. For example, inclusion and exclusion criteria for a meta-analysis contain some degree of subjectivity (Lakens et al., 2016). Consequently, it can be difficult for independent researchers to update already published meta-analyses if it is uncertain that their ongoing procedures closely mirror those used to construct the pre-existing data-set. Furthermore, for a paper that ostensibly meets the exclusion criteria for a given meta-analysis, there might be multiple effect sizes that could feasibly be included in the analysis. Original articles exploring the anxiety-ERN relationship, for example, commonly present the same statistic from multiple electrodes (e.g., Fz, FCz, & Cz), and several approaches might be justified to end up with only one effect size per data set (e.g., always using FCz, using the electrode emphasized by the authors, or pooling across electrodes). Supplementary text that unambiguously identifies the selected effect size from a given paper (e.g., including quoted text and page numbers to identify which effect size that was selected among the many in a paper) can be used to document specifically how authors extracted effect sizes based on their more subjective inclusion/exclusion criteria. In addition to providing supplementary text and analysis scripts, one basic step that should be taken in the publication of meta-analyses is to follow established minimum standards for reporting, such as PRISMA or MARS. Most studies that have included a meta-analysis of the anxiety-ERN relationship have included a statement and flow chart indicating that they followed the PRIMSA guidelines (cf., Cavanagh & Shackman, 2015; Moser et al., 2013, 2016; Riesel, 2019).

Post hoc methods to identify and adjust for publication bias in existing meta-analyses

The validity of even the most open and transparently conducted meta-analysis depends on the credibility of the primary literature. A fully preregistered and maximally reproducible meta-analysis will nevertheless provide a biased estimate of the underlying effect size if publication bias and QRPs are present in the summarized literature. Without any formal attempt at accounting for these sources of bias, a meta-analytical effect size will likely over-estimate the size of a hypothesized effect. This can occur for several reasons. If studies are selected based on statistical significance or their large effect sizes (i.e., publication bias; Rothstein et al., 2006), then multiple null results will be omitted from the meta-analytical estimate. Effect sizes are further inflated when publication bias is combined with low statistical power as only very large effects will reach conventional significance thresholds of $p < .05$ (Sterne, Gavaghan, & Egger, 2000). These small study effects do not necessarily mean that the true effect is not different from zero, as even true but small effects would be inflated by publication bias. Consequently, the validity of a meta-analysis also depends on employing some methods to assess the extent of this publication bias, and estimating the size of the 'true' underlying effect size in the absence of small study effects.

Emptying the file-drawer by finding unpublished effect sizes

Publication bias arises when studies with significant results and/or large effects are more likely to be published than non-significant results (Rothstein et al., 2006). Even if statistical power is .8 in a field and hypothesis is correct, 20% of the tests of should return non-significant results. Non-significant results should become more prevalent when statistical power is low, as

is likely true for the average ERP study (Clayson et al., 2019; Clayson et al., 2020; Szucz & Ioannidis, 2017). The inflationary effects of publication bias can be partially mitigated by seeking out unpublished effect sizes to include in their meta-analysis (Pigot & Polanin, 2020). Unpublished effects can be sought through multiple means, including student dissertations, contacting authors identified from literature reviews, seeking results from registered studies that were not published, or posting data requests—perhaps including links to your registration—to mailing lists and/or discussion forums of topic-relevant academic societies. This approach works against the so-called *file-drawer* problem by uncovering real studies that were suppressed due to publication bias. If publication bias exists, unpublished studies will likely have smaller effect sizes that are not statistically significant (Polanin, Tannin-Smith, & Hennessey, 2016). We recently sought unpublished studies while conducting a meta-analysis on the anxiety-ERN relationship; while published studies were associated with a small, significant effect (r = -.22, N = 2942), no significant effect was observed for unpublished studies (r = -.03, N = 877; Saunders & Inzlicht, 2020). Thus, unpublished effect sizes were not only smaller than published ones, but, in fact, suggested no significant relationship between anxiety and the ERN.

Conclusions derived from unpublished effect sizes should be interpreted with caution. Factors that might contribute to null results sometimes give authors good reason to avoid publication (e.g., non-specific error, data quality, low statistical power) meaning it is possible that unpublished studies have lower quality data. Equally, however, it should be noted that small published studies with large effects might also have lower quality data, but that noise moved the effect in a predicted direction that was advantageous for publication. In this sense, even finding lower quality studies in the opposite direction of the predicted effect might still

help to get a more balanced picture of the field overall. One further limitation of this method is that it relies on cooperation from other researchers to locate, and sometimes re-analyse, unpublished data. These factors often mean that attempts to find unpublished studies returns a low yield (Polanin, Espelage, et al., 2020). Indeed, we (Saunders & Inzlicht, 2020) only uncovered 7 unpublished effect sizes—two of which were from our own laboratory. Consequently, even if the unpublished data is of high quality, seeking hidden data sets will unlikely uncover sufficient information to markedly change the effect of publication bias that exists in the literature.

Statistical methods to correct for publication bias

Beyond seeking unpublished effect sizes, meta-analysts can statistically detect and correct for small publication bias using an ever increasing number of distinct techniques (cf., Carter et al., 2019; Duval & Tweedie, 2000; Iyengar & Greenhouse, 1988; McShane, Böckenholt, & Hansen, 2016; Stanley & Doucouliagos, 2014). Recent studies have indicated that some correction methods are more or less valid than others, particularly for the type of meta-data common in cognitive science.

Inappropriate bias-correction methods. Two popular methods related to publication bias that are insufficient are Fail-Safe N (Rosenthal, 1979) and trim-and-fill (Duval & Tweedie, 2000). Fail-Safe N attempts to determine the number of non-significant results that would need to be discovered to render the uncorrected meta-analytical effect non-significant. As such, Fail-Safe N attempts to estimate the tolerance of a meta-analytical effect sizes to the addition of undisclosed null-results, rather than estimating and correcting for publication bias itself. This

intuitively appealing method too often overestimates how robust meta-analytic effect sizes are to the inclusion of null results (Becker, 2005). For example, two meta-analyses on the anxiety-ERN relationship suggested that more than 1,000 null results would have to be uncovered in order for the meta-analytical effect to be rendered non-significant (Cavanagh & Shackman, 2015; Riesel, 2019)—these estimates are based on meta-analyses that each contain fewer than 40 studies. Large Fail-Safe N does not necessarily indicate low publication bias. Most problematically, the number of hidden null studies estimated by Fail-Safe N increases with each significant study that is added to the analysis, and increases rapidly when significant studies in the opposite direction of the hypothesized effect are omitted from a meta-analysis (Becker, 2005; Hilgard, 2016). As such, publication bias itself can inflate Fail-Safe N.

Trim-and-fill aims to quantify publication bias, and, unlike Fail-Safe N, it provides a bias-corrected estimate of the meta-analytical effect size (Duval & Tweedie, 2000). Trim-and-fill centres on detecting asymmetry in funnel plots—scatterplots showing the association between study effect sizes and their standard errors. In the absence of publication bias, more precise studies (e.g., those with larger samples, lower error) are assumed to provide the best estimate of the true underlying effect size. Additional nonspecific error in less precise studies would have more erratic effect sizes, causing them to fall equally in either direction around the stronger estimates. Such an unbiased literature creates the symmetrical, pyramid-like distribution on a scatterplot of effect sizes against standard error that gives the funnel plot its name (see Figure 1, left chart). However, publication bias results in the omission of effect sizes in the opposite direction of the established hypothetical effect, resulting in funnel-plot asymmetry and an

inflated estimate of the true effect size when the biased population of studies is aggregated in a meta-analysis (see figure 1, right panel).

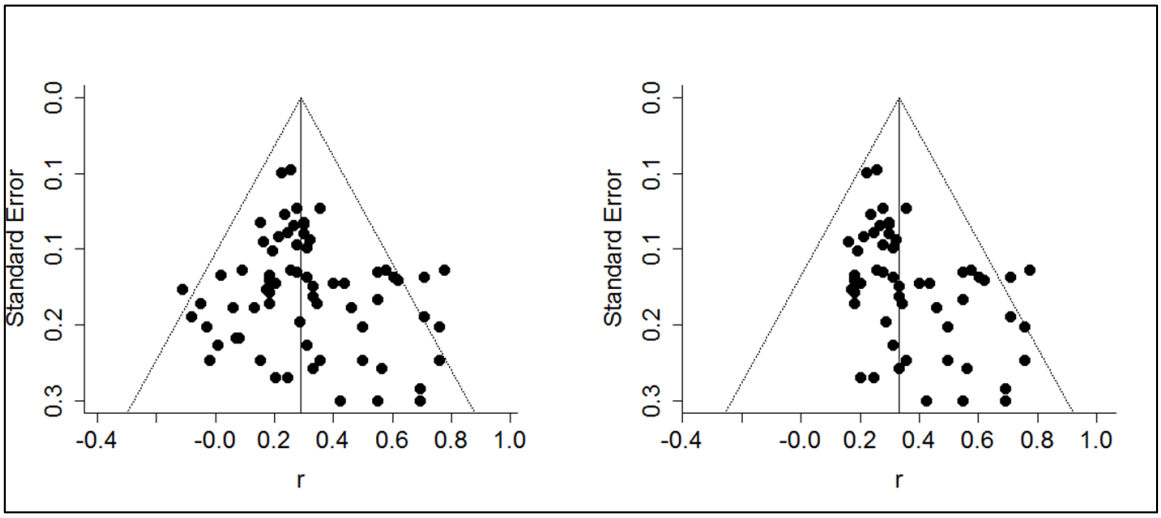

Figure 1: Left chart depicts an idealised funnel plot for a meta-analytical effect size of r = .29, k = 68, 95% CIs [.25,.32] based on simulated data. Right panel shows an asymmetric funnel plot based on the same data set, but excluding the 23 studies with small, presumably non-significant, effects (i.e., r < .2), that would likely not find its way into the published literature. This bias results in a rightward skew of the funnel plot and an inflated meta-analytical estimate, r = .37, k = 45, 95% CIs [.32,.41].

Some meta-analysts have assessed publication bias using visual inspection of scatterplots. However, as has been noted elsewhere (Ioannidis, 2008), this method lacks objectivity. Trim-and-fill attempts to reinstate symmetry in the funnel plot first by 'trimming' studies to achieve symmetry, and subsequently imputing (i.e., 'filling') values to restore symmetry when the trimmed values are re-instated. A corrected effect size can then be estimated my meta-analysing over the original and imputed values. Methods based around funnel plot asymmetry, in addition to Fail-Safe N, are the most frequently used test of bias in the anxiety-ERN relationship. Trim-and-fill resulted in little correction in our recent meta-analysis of the anxiety-

ERN relationship (Saunders & Inzlicht, 2020), while one other meta-analytic review detected no funnel plot asymmetry (Pasion & Barboa, 2019). Two other meta-analyses reported related methods of visual inspection of funnel plot asymmetry (Cavanagh & Shackman, 2015; Riesel, 2019). In addition to popularity in these meta-analyses, trim-and-fill was recommended in a recent tutorial on meta-analyses in ERP research (Moran et al., 2017).

Despite its apparent popularity, trim-and-fill has been criticised for failing to adequately adjust for publication bias, resulting in unacceptably high false-positive rates. Simulation studies have indicated that trim-and-fill performs poorly when there is anything more than mild heterogeneity (Carter et al., 2019; Jin, Zhou, & He, 2015). Furthermore, trim-and-fill also shows unacceptably high levels of false-positives when medium levels of publication bias exist, even under conditions where effect size heterogeneity is low (Carter et al., 2019). Heterogeneity levels are typically moderate-to-high in psychology (Cafri, Kromrey, & Brannick, 2010), and were moderate in three of the four meta-analyses that reported heterogeneity statistics in the anxiety-ERN relationship (Pasion & Barbosa, 2019; Riesel, 2019; Saunders & Inzlicht, 2020). As such, trim-and-fill seems unlikely to provide an appropriately conservative bias adjustment is psychology and neuroscience.

More appropriate tests of publication bias. Other methods based on regression appear to provide more conservative and appropriate adjustments for publication bias: the Precision Effect Test (PET) and Precision-Effect Estimate with Standard Error (PEESE; Stanley & Doucouliagos, 2014). Both methods follow a similar logic where more precise studies (i.e., those with less measurement error) are assumed to give a closer estimate of the "true" underlying effect size than studies with more error. In cases with publication bias or other small study

effects, an artifactual gradient emerges where effect sizes decrease as study precision increases. Regression-based bias-detection tools can estimate this artifactual gradient and correct for it when estimating meta-analytic effect sizes. PET involves a linear regression predicting effect sizes from their standard errors, weighted by the inverse of the standard error squared. In contrast, PEESE follows a similar logic but with a quadratic relationship estimated between effect sizes and their standard errors. The rationale for the quadratic term is that, if there is a true underlying effect, smaller, less precise studies will likely only become publishable if they severely overestimate the effect size, while larger, more precise studies, will be publishable (e.g., achieve conventional levels of statistical significance) even for smaller effect sizes. In both cases, the intercept for the model is taken to be the most precise study possible, and, therefore be the bias-corrected effect size.

While PET and PEESE might be used independently, a conditional logic (PET-PEESE) has been suggested. If the PET intercept is statistically significant (i.e., the corrected effect size is non-zero), it is suggested to use PEEESE to get a better estimate of the overall effect size. Alternatively, if PET is not significant, then the analyst should conclude that the meta-analytical effect size is not different from zero. Recent simulation studies have suggested that PET-PEESE works adequately well in relatively realistic circumstances, including cases with moderate heterogeneity, so long as there are sufficient studies in the meta-analysis (~ $k \geq 30$) to achieve sufficient statistical power (cf., Carter et al., 2019). Our recent meta-analysis of the anxiety-ERN relationship included PET-PEESE, indicating a significant meta-analytical effect for PEESE (r = -.12) but not PET (r = -.05). The conditional logic of PET-PEESE would therefore put forward the conclusion that the anxiety-ERN relationship is, overall, not significantly different from zero.

This conclusion is starkly different from one of little bias suggested by trim-and-fill, but is nevertheless consistent with the effect size from our summary of uncovered unpublished effect sizes.

One final class of correction procedures are selection methods. Here, we focus on a three parameter model developed by Iyengar & Greenshouse (1988) that has shown favourable results in recent simulation studies (Carter et al., 2019; McShane et al., 2016). The three-parameter selection method has two parameters that attempt to describe the data: an effect size parameter for the population effect size, and a second parameter that reflects the heterogeneity of the effect sizes in the meta-analyses. The third selection parameter is a weight parameter that provides the probability that a non-significant effect will enter the literature (cf., Iyengar & Greenhouse, 1988). This selection model can be implemented using the *weightr* (Coburn & Vevea, 2017) package in R that reports the adjusted effect size and the likelihood ratio test, which provides a χ^2 statistic comparing the unadjusted and adjusted effect-size estimates. We also included the three-parameter selection model in our assessment of publication bias in the anxiety-ERN relationship, with this analysis suggesting a small but significant bias-corrected effect size (r = -.14, Saunders & Inzlicht, 2020).

Returning mixed results across multiple corrections for publication bias, as in the case with our recent investigation of the anxiety-ERN relationship, is unsatisfactory, as it leaves confusion about the true effect size. Easing this uncertainty somewhat, Carter et al (2019) compared multiple correction methods—including trim-and-fill, PET-PEESE, and selection models—in a simulation that varied parameters to capture the typical state of meta-analyses psychology (i.e., publication bias, QRPs, heterogeneity, effect sizes). Here, PET-PEESE and the

three-parameter selection model faired similarly, and both were better than trim-and-fill, which showed an unacceptable false-positive rate. Carter et al. (2019) suggested using multiple correction methods in a sensitivity analysis to determine how robust the meta-analytic effect size is across a range of correction methods that perform well under different circumstances. It should be noted that, while PET-PEESE and the three-parameter selection model both performed adequately, the three-parameter selection model routinely approximated the true underlying effect size best. The three-parameter selection method, then, might provide the best currently available bias-corrected effect size for the anxiety-ERN relationship as r = -.14. However, it is important to know that each adjusted effect size is an estimate. Future well-powered confirmatory tests are required to assess if the effect sizes predicted by each correction method bear out.

Challenges and future directions: prospective meta-analyses

As can be seen from the prior sections, the straightforward interpretation of meta-analyses as a definitive, statistically powerful estimate of a true underlying effect size is complicated by factors that introduce bias to either the primary literature, the production of the meta-analysis itself, or both. Furthermore, while statistical methods can provide a fair impression of the bias-corrected effect size, the results from these analyses are also not definitive. As illustrated in our example, viable meta-analytic estimates of the anxiety-ERN relationship range between medium sized effects (Moser et al., 2013), to an effect that is not distinguishable from zero bias (PET-PEESE, Saunders & Inzlicht, 2020). Whether each of these reflect the true effect size, or whether the truth lies somewhere in the middle, cannot be determined from retrospective meta-analyses and correction methods. This impasse is

particularly disappointing considering that recent meta-analyses integrated data from thousands of participants. These statistics point to a real inefficiency in the verification of relatively straightforward hypotheses in ERP research, and suggest that steps should be taken to conduct high quality confirmatory tests.

As much of the uncertainty in meta-analytical effect sizes come from bias in the primary literature itself, future ERP research could be made more confirmatory by increasing the use of preregistration and registered reports. While large meta-analyses often fail to settle debates or provide definitive conclusions (Ferguson, 2014), estimates from meta-analyses can be used as the basis for power analyses for ongoing confirmatory studies. These power analyses should be based on bias-corrected estimates unless the author can be confident that publication bias did not influence the uncorrected meta-analytic effect size. However, actually running a study that is sufficiently powered to find this effect confers a considerable cost on the researcher. A well powered study with appropriate parameters (one-tailed, $\alpha = .05$, Power = .9) would require N=430 to detect $r = -.14$. As mentioned earlier in this manuscript, most labs would be unable or unwilling to collect this quantity of data. What this suggests is that other, more prospective approaches are also needed.

Traditional meta-analysis is retrospective, meaning that authors often make decisions about meta-analytical protocols (e.g., selection criteria, search terms, moderators) based on their expert knowledge of the research area subjected to quantitative review. These factors often mean that meta-analyses are largely exploratory rather than confirmatory, as different patterns of decisions by experts in the field can result in meta-analyses with diverging conclusions about the same topic (Watt & Kennedy, 2017). In prospective meta-analyses, on

the other hand, the meta-analysis is preregistered following established reporting standards, but, rather than integrating already existing results (a *retrospective meta-analysis*), only data that is collected after the registration is included (a *prospective meta-analysis*; Ghersi, Berlin, & Askie, 1999; Reade et al., 2010). This approach allows for truly confirmatory meta-analyses by ensuring that analytic decisions cannot be based on prior knowledge of existing results.

While prospective meta-analysis has potential benefits over *post hoc* analyses, some practical limitations exist. First, as it is difficult to anticipate the design of future studies, the registered meta-analysis plan might need to be adapted to account for new studies with unanticipated design quirks (Watt & Kennedy, 2017). In this sense, prospective meta-analysis progresses somewhat iteratively as do retrospective analysis (Lakens et al., 2016; Moher et al., 2009). Second, it is impossible to know if and when new studies will emerge on your prospective research question, with the production of new data dependent on the popularity of the question. Indeed, it is not unusual for prospective meta-analyses to take more than a decade to complete (Reade et al., 2010). Third, while a prospective meta-analysis can eliminate bias arising from the meta-analysis itself, the validity of meta-analyses is also challenged by bias in the primary literature. One remedy is to base the prospective meta-analysis solely on studies with minimal bias (e.g., a meta-analysis of Registered Reports; Gronau, Van Erp, Heck, Cesario, Jonas, & Wagenmakers, 2017). However, as registered reports only make up a very small minority of published studies, prospective meta-analysis based on the ad-hoc publication of registered reports would be data poor, and would involve ignoring the majority of studies on a topic. Fourth, prospective meta-analysis alone depend on data appearing through traditional

means, meaning that prospective meta-analysis alone would not counter the inability of ERP researchers to collect large data sets.

Multisite collaboration

One fruitful approach to resolve these issues it to combine prospective meta-analysis with multi-site projects that coordinate data collection across independent laboratories (Watt & Kennedy, 2017; Simons, Holcombe, & Spellman, 2014). This method allows for confirmatory meta-analysis based on distributed data-collection, while also giving the consortium of researchers control over the rate and quantity of data collection. This collective approach requires considerable organisation compared to each laboratory working independently, however, there are numerous examples of such approaches from psychology (Moshontz et al., 2018; O'Donnell et al., 2018; Wagenmakers et al., 2016). For example, Registered Replication Reports (cf., Simons et al., 2014) aim to replicate established and influential effects in psychology by having distributed laboratories run a study with an identical preregistered protocol that are combined in a meta-analysis with minimal bias. Interestingly, examples of this format include cases where retrospective meta-analyses have provided conflicting results, such as the effect of so-called ego depletion on self-control (Carter & McCullough, 2014; Hagger, Wood, Stiff, & Chatzisarantis, 2010; Hagger et al., 2016; Inzlicht, Gervais, & Berkman, 2015). In addition to Registered Replication Reports, other initiatives, such as the Psychological Science Accelerator (Moshontz et al., 2018) have attempted to build networks of researchers interested in conducting multisite research that are then in a state of readiness to join collective data collection efforts once studies are accepted by the network.

While methods exist to conduct multisite prospective meta-analyses, challenges arise when conducting such a process in EEG research. First, past examples have required each replicating lab to conduct a study that was sufficiently powered to detect the effect size of interest (Open Science Collaboration, 2015). This method would be inappropriate for ERP research as it would mandate each lab recruiting hundreds of participants for individual difference studies. Instead, a collective effort in which many laboratories collect more modest samples of data and integrate this data into a later meta-analysis would likely be more practical. Second, hardware differs across labs, including amplifiers, electrode sets (active vs. passive; number of electrodes; electrode placement), electrical shielding (e.g., labs may or may not record inside a faraday cage), and other apparatus common to wider psychological experimentation (e.g., response boxes, audio equipment, monitor refresh rates). As a minimum, multisite collaborations should provide supplemental materials that catalogue differences between labs and ensure that some essential similarities are maintained across sites (e.g., common references, analysis electrodes). Third, analysis of ERP data is incredibly heterogeneous even when authors ostensibly extract the same ERP component (see Fischer et al., 2017 for discussion regarding the ERN). One method to ensure consistency would be to decide all analysis steps in advance through communications within the participating authors. Alternatively, analysis of the EEG data could be centralized to one lab, or authors might conduct a multiverse analysis in which the robustness of a given finding is checked by attempting to conduct every justifiable analysis of a given data set to test how much the conclusions are consistent across a range of justifiable analytic choices (cf., Steegen, Tuerlinickx, Gelman, & Vanpaemel, 2016).

Finally, prospective multisite investigations present logistic difficulties. Multisite studies sometimes occur in a relatively *ad hoc* manner with a collection of researchers interested in a specific hypothesis (Nieuwland et al., 2018), while other approaches have been journal led, such as the RRR initiative in *Perspectives in Psychological Science* (O'Donnell et al., 2018; Simons et al., 2014; Wagenmakers et al., 2016). A benefit of the journal-led approach is that in-principle acceptance can be granted before commencing an undertaking that potentially involves thousands of research participants in tens of laboratories. This advance commitment to publishing the study results not only helps to recruit participating labs, but also means that the protocols are both submitted to advanced scrutiny meaning that the design can be improved based on reviewer and editor feedback, but also that the prospective protocol can be decided in advance and locked within the journals systems to protect against experimenter degrees of freedom.

Summary and conclusions

Meta-analysis can be a powerful tool to integrate smaller pools of data to make powerful statistical inferences. This benefit might be particularly salient in resource intensive fields, such as ERP research, where individual laboratories will likely struggle to achieve large enough samples to precisely estimate effect sizes associated with a given hypothesis, especially for research involving between-subject designs such as individual difference research. However, the potential strengths of meta-analysis is limited by multiple sources of bias. Despite 6 meta-analyses existing on the anxiety-ERN relationship, for example, estimates of the true underlying effect range from small-to-medium uncorrected effects, to their potentially being no real relationship between anxiety and the ERN after correcting for publication bias. Furthermore,

while a hierarchy emerges of better and worse performing statistical methods that can be applied to correct for publication bias in retrospective meta-analyses. The *post hoc* nature of these methods, in addition to the range of values returned by different correction methods, makes them more suitable as a sensitivity analysis to determine plausible meta-analytical effect sizes, rather than providing a confirmatory test of the underlying hypothesis.

Adopting a range of open science practices can help meta-analyses to realise their potential as a method to facilitate cumulative scientific inferences. Novel meta-analyses can be improved with *a priori* preregistration and transparent reporting practices that both help consumers of meta-analytic reviews to understand and evaluate their claims. One further benefit of increased transparency—particularly the sharing of data—is that it allows future researchers to update meta-analyses as and when newer studies emerge testing the same hypotheses. While these open science practices will increase the credibility of meta-analysis, any retrospective meta-analysis is limited by multiple sources of bias that, as mentioned, cannot be completely resolved through statistical methods that correct for publication bias. Prospective meta-analyses, ideally based on studies that are preregistered in order to minimize bias—have the potential to conduct truly confirmatory hypothesis testing while relying on smaller pools of data collected in distributed laboratories.

In sum, we advocate for a more team-science approach to the study of neurophysiology. Team-science—in which multiple laboratories collaborate to collect sufficient data to test a hypothesis of mutual interest—can facilitate the timely completion of prospective meta-analyses, and ERP research could follow established models from psychology (e.g., Registered Replication Reports; Psychological Science Accelerator) in order to achieve this goal. In

conjunction with modern meta-analytic techniques, team science might allow for a truly cumulative science that makes fewer errors and expedites the uncovering of truths.

References

Beller, E. M., Chen, J. K., Wang, U. L., & Glasziou, P. P. (2013). Are systematic reviews up-to-date at the time of publication? *Systematic Reviews*, *2*, 36.

Becker, B. J. (2005). Failsafe N or file-drawer number. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 111-125.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). Introduction to meta-analysis. *John Wiley & Sons.*

Booth, A., Clarke, M., Dooley, G., Ghersi, D., Moher, D., Petticrew, M., & Stewart, L. (2012). The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Systematic Reviews*, 1, 1-9.

Booth A, Clarke M, Ghersi D, Moher D, Petticrew M, Stewart L: An international registry of systematic review protocols. *Lancet*. 2011, 377: 108-109. 10.1016/S0140-6736(10)60903-8.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365-376.

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, *45*, 239-270.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated?. *Frontiers in Psychology*, *5*, 823.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*, 115-144.

Cavanagh, J. F., & Shackman, A. J. (2015). Frontal midline theta reflects anxiety and cognitive control: meta-analytic evidence. *Journal of Physiology-Paris*, *109*, 3-15.

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*, 609-610.

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*, e13437.

Clayson, P. E., Carbine, K. A., & Larson, M. J. (2020). A registered report of error-related negativity and reward positivity as biomarkers of depression: P-Curving the evidence. *International Journal of Psychophysiology*, *150*, 50-72.

Cooper, H (2010). Research synthesis and meta-analysis: A step-by-step approach (4th ed., Applied Social Research Methods Series, Vol. 2). Thousand Oaks, CA: Sage.

Créquit, P., Trinquart, L., Yavchitz, A., & Ravaud, P. (2016). Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: the example of lung cancer. BMC Medicine, *14*, 8.

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89-98.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*, 792–806. https://doi.org/10.1177/0956797620916786

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. Psychological Methods, *17*, *120-128*. doi:10.1037/a0024445

Ferguson, C. J. (2014). Comment: Why Meta-Analyses Rarely Resolve Ideological Debates. *Emotion Review*, *6*, 251–252. https://doi.org/10.1177/1754073914523046

Ferguson, C. J. (2015). Pay no attention to that data behind the curtain: On angry birds, happy children, scholarly squabbles, publication bias, and why betas rule metas. *Perspectives on Psychological Science*, *10*, 683-691.

Fischer, A. G., Klein, T. A., & Ullsperger, M. (2017). Comparing the error-related negativity across groups: The impact of error-and trial-number differences. *Psychophysiology*, *54*, 998-1009.

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The error-related negativity (ERN/Ne). In S. J. Luck, & E. Kappenman (eds.), *Oxford handbook of event-related potential components (pp. 231-291).* New York: Oxford University Press.

Ghersi, D., Berlin, J., & Askie, L. (2011). Cochrane prospective meta-analysis Methods Group. COCHRANE METHODS, 35.

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *JAMA*, *298*, 430-437.

Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E. J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*, 123-138.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... & Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546-573.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological Bulletin*, *136*, 495.

Hajcak, G. (2012). What we've learned from mistakes: Insights from error-related brain activity. *Current Directions in Psychological Science*, *21*, 101-106.

Hilgard, J. (2016, July 19). The Failure of Fail-safe N. Retireved from http://crystalprisonzone.blogspot.com/2016/07/the-failure-of-fail-safe-n.html

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law*, *2*, 324–347. https://doi.org/10.1037/1076-8971.2.2.324

Ioannidis, J. P. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice*, *14*, 951-957.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 109-117.

Jeon, Y. W., & Polich, J. (2001). P300 asymmetry in schizophrenia: a meta-analysis. *Psychiatry Research*, *104*, 61-74.

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

Kang, E., Keifer, C. M., Levy, E. J., Foss-Feig, J. H., McPartland, J. C., & Lerner, M. D. (2018). Atypicality of the N170 event-related potential in autism spectrum disorder: A meta-analysis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*, 657-666.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*, 196–217.

Kirkham, J. J., Altman, D. G., & Williamson, P. R. (2010). Bias due to changes in specified outcomes during the systematic review process. *PloS ONE*, *5*, e9810.

Krzywinski, M., & Altman, N. (2013). Power and sample size. *Nature Methods*, *10*, 1139–1140.

Inzlicht, M., Gervais, W., & Berkman, E. (2015). Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015. Kofler, Forster, & McCullough.

Jin, Z. C., Zhou, X. H., & He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, *34*, 343-360.

Lakens, D., Adolfi, F., Albers, C., Anvari, F., Apps, M., Argamon, S., Baguley, T., Becker, R., Benning, S., Bradford, D., Buchanan, E., Caldwell, A., Van, C. B., Carlsson, R., Chen, S., Chung, B., Colling, L., Collins, G., Crook, Z., Cross, E., Daniels, S., Danielsson, H., Debruine, L., Dunleavy, D., Earp, B., Feist, M., Ferrell, J., Field, J., Fox, N., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J., Grieve, A., Guggenberger, R., Grist, J., Van, H. A., Hasselman, F., Hochard, K., Hoffarth, M., Holmes, N., Ingre, M., Isager, P., Isotalus, H., Johansson, C., Juszczyk, K., Kenny, D., Khalil, A., Konat, B., Lao, J., Larsen, E., Lodder, G., Lukavský, J., Madan, C., Manheim, D., Martin, S., Martin, A., Mayo, D., McCarthy, R., McConway, K., McFarland, C., Nio, A., Nilsonne, G., De, O. C., De, X. J., Parsons, S., Pfuhl, G., Quinn, K., Sakon, J., Saribay, S., Schneider, I., Selvaraju, M., Sjoerds, Z., Smith, S., Smits, T., Spies, J., Sreekumar, V., Steltenpohl, C., Stenhouse, N., Swiatkowski, W., Vadillo, M., Van, A. M., Williams, M., Williams, S., Williams, D., Yarkoni, T., Ziano, I., Zwaan, R., (2018), Justify your alpha, *Nature Human Behaviour*, *2*, 168-171.

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*, 24.

Lakens, D., LeBel, E. P., Page-Gould, E., van Assen, M. A. L. M., Spellman, B., Schönbrodt, F. D., … Hertogs, R. (2017, July 9). Examining the Reproducibility of Meta-Analyses in Psychology. Retrieved from osf.io/q23ye

Luck, S. J. (2014). *An introduction to the event-related potential technique.* MIT press.

Meyer, A. (2017). A biomarker of anxiety in children and adolescents: A review focusing on the error-related negativity (ERN) and anxiety across development. *Developmental Cognitive Neuroscience*, *27*, 58-68.

Meyer, A., Riesel, A., & Proudfit, G. H. (2013). Reliability of the ERN across multiple tasks as a function of increasing errors. *Psychophysiology*, *50*, 1220-1225.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med 6*: e1000097.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730-749.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a

Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, *1*, 501–515. https://doi.org/10.1177/2515245918797607

Moser, J. S., Moran, T. P., Kneip, C., Schroder, H. S., & Larson, M. J. (2016). Sex moderates the association between symptoms of anxiety, but not obsessive compulsive disorder, and error-monitoring brain activity: A meta-analytic review. *Psychophysiology*, *53*, 21-29.

Moser, J., Moran, T., Schroder, H., Donnellan, B., & Yeung, N. (2013). On the relationship between anxiety and error monitoring: a meta-analysis and conceptual framework. *Frontiers in Human Neuroscience*, *7*, 466.

Moran, T. P., Schroder, H. S., Kneip, C., & Moser, J. S. (2017). Meta-analysis and psychophysiology: A tutorial using depression and action-monitoring event-related potentials. *International Journal of Psychophysiology*, *111*, 17-32.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... & Mézière, D. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, 7, e33468.

Nosek, B.A., & Lakens, D. (2014). Registered Reports: A method to increase the credibility of published results. *Social Psychology, 45,* 137-141.

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., ... & Balatekin, N. (2018). Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13*, 268-294.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: high-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*, 24-46.

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

Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*, 1745691620906416.

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*, 207-236.

Polich, J., Pollock, V. E., & Bloom, F. E. (1994). Meta-analysis of P300 amplitude from males at risk for alcoholism. *Psychological Bulletin, 115*, 55.

Pasion, R., & Barbosa, F. (2019). ERN as a transdiagnostic marker of the internalizing-externalizing spectrum: A dissociable meta-analytic effect. *Neuroscience & Biobehavioral Reviews*, *103*, 133-149.

Pereira, T. V., & Ioannidis, J. P. (2011). Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology*, *64*, 1060-1069.

Quintana D. S. (2015). From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology*, *6*, 1549. https://doi.org/10.3389/fpsyg.2015.01549

Reade, M.C., Delaney, A., Bailey, M.J. et al. Prospective meta-analysis using individual patient data in intensive care medicine. Intensive Care Med 36, 11–21 (2010). https://doi.org/10.1007/s00134-009-1650-x

Riesel, A. (2019). The erring brain: Error-related negativity as an endophenotype for OCD—A review and meta-analysis. *Psychophysiology*, *56*, e13348.

Riesel, A., Weinberg, A., Endrass, T., Meyer, A., & Hajcak, G. (2013). The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks. *Biological Psychology*, *93*, 377–385. https://doi.org/10.1016/j.biopsycho.2013.04.007

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646–656. https://doi.org/10.1037/0022-006X.58.5.646

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638-64.

Rosenthal, R., & DiMatteo, M. R. (2002). Meta-analysis. *Stevens' handbook of experimental psychology.*

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). Publication bias in meta-analysis: Prevention, assessment and adjustments. John Wiley & Sons.

Sandre, A., Banica, I., Riesel, A., Flake, J., Klawohn, J., & Weinberg, A. (2020). Comparing the effects of different methodological decisions on the error-related negativity and its association with behaviour and gender. *International Journal of Psychophysiology*, *156*, 18-39.

Saunders, B., & Inzlicht, M. (2020). Assessing and adjusting for publication bias in the relationship between anxiety and the error-related negativity. *International Journal of Psychophysiology.*

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,* 1359-1366.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science, 9(5), 552-555.*

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin,* 144(12), 1325-1346. http://dx.doi.org/10.1037/bul0000169

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

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702-712.

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Investigating and dealing with publication and other biases in meta-analysis. *BMJ, 323,* 101-105.

Sterne, J.A.C., Gavaghan, D., &, Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J. Clin. Epidemiol. 53,* 1119–1129

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology, 15,* e2000797.

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

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology*, *53*, 207-216.

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, *13*, 411–417. https://doi.org/10.1177/1745691617751884

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Bulnes, L. C. (2016). Registered replication report: strack, martin, & stepper (1988). *Perspectives on Psychological Science*, *11*, 917-928.

Watt, C. A., & Kennedy, J. E. (2017). Options for prospective meta-analysis and introduction of registration-based prospective meta-analysis. *Frontiers in Psychology*, 7, 2030.

Weinberg, A., Dieterich, R., & Riesel, A. (2015). Error-related brain activity in the age of RDoC: A review of the literature. *International Journal of Psychophysiology*, *98*, 276-299.

Williamson, P. R., Gamble, C., Altman, D. G., & Hutton, J. L. (2005). Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research*, *14*, 515-524.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*, 931.