International Journal of Psychophysiology 111 (2017) 17–32

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Meta-analysis and psychophysiology: A tutorial using depression and action-monitoring event-related potentials

Tim P. Moran [a,*], Hans S. Schroder [b], Chelsea Kneip [b], Jason S. Moser [b]

[a] School of Psychology, Georgia Institute of Technology, United States
[b] Department of Psychology, Michigan State University, United States

ARTICLE INFO

Article history:
Received 15 February 2016
Received in revised form 9 June 2016
Accepted 1 July 2016
Available online 1 July 2016

Keywords:
Meta-analysis
Psychophysiology
Depression
Error-related negativity
Feedback negativity

ABSTRACT

Meta-analyses are regularly used to quantitatively integrate the findings of a field, assess the consistency of an effect and make decisions based on extant research. The current article presents an overview and step-by-step tutorial of meta-analysis aimed at psychophysiological researchers. We also describe best-practices and steps that researchers can take to facilitate future meta-analysis in their sub-discipline. Lastly, we illustrate each of the steps by presenting a novel meta-analysis on the relationship between depression and action-monitoring event-related potentials – the error-related negativity (ERN) and the feedback negativity (FN). This meta-analysis found that the literature on depression and the ERN is contaminated by publication bias. With respect to the FN, the meta-analysis found that depression does predict the magnitude of the FN; however, this effect was dependent on the type of task used by the study.

1. Introduction

Most scientific questions are addressed by multiple studies conducted by independent research teams using a diverse range of methods rather than by a single study. Researchers understand and accept that the results of these studies will often vary and, in some cases, may directly contradict each other. Yet researchers also want to be able to use these varied and conflicting findings to come to a consensus regarding a body of work – for example, it is often desirable to determine whether the predictions of a theory have been supported or whether a finding has practical applications. For the greater part of the previous century, researchers from a number of fields including physics, psychology, ecology, zoology, archaeology, astronomy and medicine (Birge, 1929, 1932; Haidich, 2010; Petticrew, 2001) have relied on meta-analysis to quantitatively summarize a body of work and draw conclusions. "Meta-analysis" refers to a set of procedures that statistically analyze the results of primary studies (i.e. the original research) in order to synthesize the findings (Glass, 1976).

The purpose of this article is to provide a broad overview of what meta-analysis is as well as a practical tutorial aimed at psychophysiologists. This article is organized around a series of steps that nearly all meta-analyses will follow (adapted from Cooper, 2010; Cumming, 2012): formulating the problem, conducting the literature search, coding studies and extracting data, synthesizing effect sizes and assessing for heterogeneity, and assessing for threats to validity. Each of these sections will present tips, strategies and best-practices for conducting a meta-analysis. Each of the five sections will end with an illustrative example from a novel meta-analysis we performed on the relationship between depression and action-monitoring event-related potentials (ERPs), namely the error-related negativity (ERN) and the feedback negativity (FN). We conclude by identifying challenges to conducting robust meta-analyses, and offer some possible solutions for psychophysiologists to take up in planning, executing, and reporting on future studies.

2. Step 1: formulate the problem

Conducting a meta-analysis can take a great deal of time and effort. For example, one of the authors recently completed a meta-analysis which required approximately 16 months (Moran, in press). Given the work involved, one could legitimately wonder if summarizing the existing literature with a meta-analysis is a better use of one's time than trying to address an existing question with a new primary study or by summarizing the literature with a narrative review. The type of study that one conducts should, of course, be dependent on one's goals. The following are goals that meta-analysis is particularly well-suited, or uniquely-suited, to addressing: 1) Determine if an effect is "real." Psychologists of all stripes must deal with findings that

* Corresponding author at: Georgia Institute of Technology, 654 Cherry St., Atlanta, GA 30332, United States.
E-mail address: timothy.moran@psych.gatech.edu (T.P. Moran).

http://dx.doi.org/10.1016/j.ijpsycho.2016.07.001
0167-8760/

occasionally fail to replicate – whether due to false positives or simply sampling error and low power. A meta-analysis can help test an effect, often with far greater power than any single study (see Section 6.6 for an example). 2) Determine the consistency of an effect. Psychologists are accustomed to the fact that some effects may be dependent on a particular population/setting/design etc. – that is, an effect might be moderated by some other variable. Given that even two studies testing the same hypotheses can differ markedly in these variables, a highly inclusive meta-analysis is well-suited to testing the role of moderating variables – and with much greater power than any single study. Additionally, meta-analyses can test moderators that no single study is capable of addressing. For example, a researcher may wish to determine if the effects in a given field are shrinking over time – the so-called "decline effect" (Schooler, 2011). Obviously, no single study is capable of addressing this; but a meta-analysis involves summarizing the results of many studies published over several years. 3) Increase the precision of an estimate. In some cases, a psychologist may wish to go beyond statistical significance and compute a highly precise estimate of the magnitude of an effect. This is likely to be most important in applied settings where precise estimates of a measure's predictive validity are highly desired. 4) Assess the literature for publication bias. Publication bias, described below, occurs when the published literature systematically differs from the population of all studies conducted on a topic. Thus, publication bias affects an entire literature, not an individual study. A meta-analysis may be able to determine whether a literature is contaminated by publication bias whereas a primary study is not.

That having been said, meta-analyses are not appropriate for all situations. In particular, a meta-analysis cannot fix a "broken" literature. For example, if a researcher believes that a given literature is full of poorly conducted studies – e.g. invalid instruments, poor experimental designs, etc. – a meta-analysis will not be able to produce a meaningful summary of that literature. Meta-analyses obey the "law of conservation of garbage" – i.e. garbage in, garbage out. If a meta-analysis includes low-quality studies with questionable findings, the results of the meta-analysis will also be questionable. In these situations, a meta-analysis might be able to directly compare effects from low- and high-quality studies (although, the low/high distinction must be determined by the meta-analyst), but it will not be able to correct for poor design. When a literature is contaminated with poorly conducted studies, a new primary study which address the limitations of the literature may be a more prudent use of time.

Once it has been determined that a meta-analysis is the appropriate design for a given question, the researcher must carefully formulate the problem they wish to address. This may be as simple as wanting to know if one variable predicts another or as complicated as testing the predictions of a theoretical model. The meta-analyst must also carefully define the scope of their investigation. For example, a psychophysiologist who wants to study the N2 ERP component must decide whether to include the N2a, the N2b, the N2pc etc. This is not a strictly linear process. It is possible that the scope of a meta-analysis may need to be refined as research reports are located. Given that the process of formulating a research question is likely to be familiar to most researchers, we will not discuss these issues in depth.

In the following section, we formulate the research problem regarding the relationship between depression and action-monitoring ERPs – i.e., the ERN and the FN. We briefly review the relevant literature and describe why a meta-analysis is a useful way to proceed.

2.1. Example using depression and the ERN/FN

Depression is among the most common psychiatric conditions and is associated with a high rate of recurrence and significant personal and societal cost (Greden, 2001; Lai, 2011; Lopez et al., 2006; World Health Organization, Switzerland, 2011). For example, depression is associated with increased healthcare costs and service utilization, missed work, impaired academic and social functioning, recurrent depressive episodes and increased risk for suicide. Given these findings, the last few years have seen increased effort to identify "biomarkers" which indicate risk for the development of depression thereby facilitating early diagnosis and preventative care.

For depression, two event-related potentials have shown promise as candidate biomarkers: the error-related negativity (ERN) and the feedback negativity (FN). The ERN is a negative deflection in the human event-related potential (ERP) that occurs within 100 ms of the commission of an error during forced-choice reaction time tasks – e.g. the Flankers task (see Fig. 1). In the Flankers task, participants must identify a central stimulus that is surrounded by several flanking distracters (e.g. <<<<< or <<><<) as quickly as possible. In tasks such as this, participants often make quick incorrect responses due to lapses in attention or the incongruency between the central and flanking stimuli. The ERN is generated in the anterior cingulate cortex and surrounding motor areas and is often considered an error detection/correction (Carter and van Veen, 2007; Gehring et al., 2012; Holroyd and Coles, 2002; van Veen and Carter, 2002) or response-conflict (Yeung et al., 2004) signal. With respect to psychopathology, it has been hypothesized that the ERN may serve as a biomarker for all internalizing disorders including depression (Olvet and Hajcak, 2008). In support of this proposal, an enlarged ERN has been observed in individuals with major depression (e.g. Holmes and Pizzagalli, 2008, 2010) as well as undergraduates high in self-reported sadness (e.g. Dywan et al., 2008).

Unlike the ERN, which is elicited by an internal monitoring process, the FN is elicited by external feedback indicating an unfavorable response. The FN is most often elicited in a gambling/guessing task in which the participant must make a choice (e.g. determine which door has a prize behind it) and is then rewarded for a "correct" choice (e.g. money might be awarded for a correct guess and taken away for an incorrect guess). The FN is thought to originate in the ACC (Gehring et al., 2012; Holroyd and Coles, 2002) or the striatum (Foti et al., 2011); theorizing on the FN suggests that it signals that an event was worse than expected or that it signals a desired event (Holroyd and Coles, 2002; Foti et al., 2011). With respect to depression, Hajcak and colleagues have conducted a number of studies demonstrating that depression is related to an attenuated FN in undergraduates (Foti and Hajcak, 2009), children (Bress et al., 2012) and patients suffering from major depression (Foti et al., 2014).

The findings reviewed above suggest that depression is characterized by both an enhanced ERN and a blunted FN. However, this set of findings has proven somewhat difficult to replicate. For example, with respect to the ERN, a number of studies have found no difference between depressed individuals and controls in both adults (Olvet et al., 2010; Weinberg et al., 2012) and children (Bress et al., 2015) whereas others have found that depression is associated with a reduced ERN (Ladouceur et al., 2012; Ruchsow et al., 2004, 2006; Schrijvers et al., 2009). Similarly, although several studies have found evidence for a reduced FN in depression, other work has found evidence for an enlarged FN in depression (e.g. Mies et al., 2011; Mueller et al., 2015). Before the ERN/FN can be applied to clinical settings, it must be determined whether, and how strongly, they are associated with depression.

There are a number of possibilities that can potentially explain these disparate findings. For example, the differences in findings may be attributable to some untested moderator(s). Studies assessing the association between depression and the ERN/FN have employed a variety of different types of samples and tasks. Some have studied depressed undergraduates whereas others have studied patients suffering from major depressive disorder. Among the studies of major depressive disorder, some have focused on untreated patients only whereas others have included patients receiving medication. Additionally, a number of different tasks (e.g. Flanker, Stroop, gambling etc.) have been used to elicit the ERN/FN. Each of these factors may have a role in explaining these disparate findings. It is also possible that publication bias – the tendency for smaller studies to be published only if they produce positive findings and for larger studies to be published regardless of their

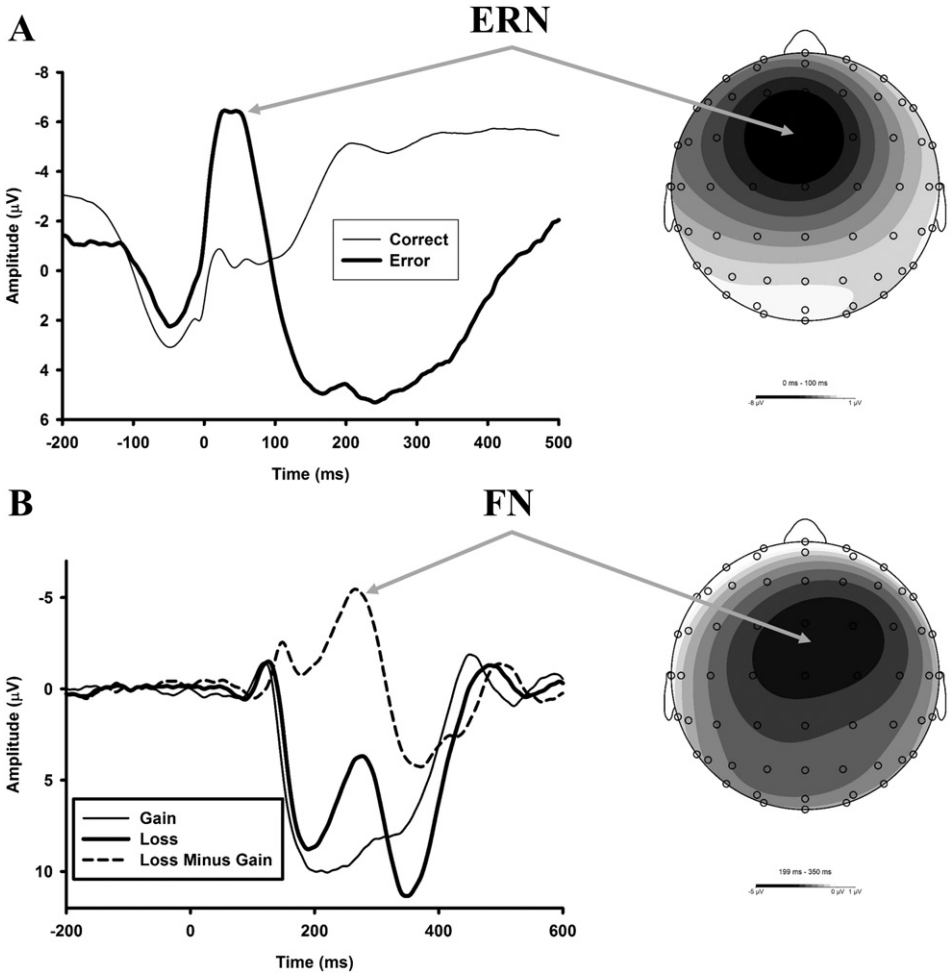

Fig. 1. A: The error-related negativity (ERN) and its scalp distribution. B: The feedback negativity (FN) and its scalp distribution.

findings – may play a role as well. Lastly, it may be that these disparate findings can simply be attributed to a combination of sampling error and small sample sizes.

In other words, we want to know 1) if these effects are "real", 2) whether there are any moderators that might affect the effect size, 3) whether the literature base is contaminated by publication bias and 4) given the potential clinical applications, we want to determine the most precise estimate of the effect size possible. A meta-analysis is well suited to evaluating these possibilities. Thus, in the present study we aimed to meta-analytically review the literature on depression and the ERN/FN as well to evaluate the stability of this effect across different population (e.g. students vs patients) and procedural (e.g. type of task) moderators. Finally, the present analysis aimed to evaluate the role of publication bias in the association between depression and the ERN.

3. Step 2: conduct the literature search

Searching the literature for relevant studies is among the simplest steps in conducting a meta-analysis but it is also among the most important. The literature search determines which studies are eligible for inclusion in the meta-analysis and, therefore, the scope, validity and generality of the meta-analysis. Having a set of clear, well-formulated questions is critical for conducting a literature search as they will guide the inclusion/exclusion of primary studies.

In order to locate as many relevant citations as possible, a wide range of scientific databases should be searched. Most of these databases will be well-known to readers, for example: EMBASE, Google Scholar, MedLine, PsychInfo, PubMed, Scopus, and Web of Science. The Cochrane Collaboration also includes many journals not indexed by other search engines. To avoid publication bias, it is also desirable to expand the literature search beyond published articles. In recent years, several online repositories have begun accepting researchers' unpublished work as well as replication attempts. "Psych file drawer" (http://www.psychfiledrawer.org/) is among the most well-known but there are likely to be many depending on the discipline. Unpublished master's theses and doctoral dissertations are another source of data. Many of these can be accessed using the ProQuest database (http://www.proquest.com/). When searching for older articles in journals not indexed by search engines, meta-analysts can supplement electronic searches with hand searches through databases such as Index Medicus and Excerpta Medica. Lastly, the reference sections of primary studies and existing reviews can also serve as sources of new studies. When conducting a meta-analysis, we recommend using as many of these sources of data as is appropriate.

We generally believe that a meta-analyst should cast as a wide a net as possible when searching for relevant studies. This will help increase the generality of the findings across populations, instruments and settings, increase the power of statistical tests, and increase the precision of the estimates. However, there may be situations in which the meta-analyst wishes to exclude certain studies. For example, primary studies may include invalid instruments or fail to include the appropriate control group. As noted earlier, if a meta-analysis includes low-quality studies with questionable findings, the results of the meta-analysis will also be questionable. Thus, low-quality primary studies can threaten the validity of a meta-analysis.

In addition to differences in data quality, there may be qualitative differences between studies that may make them inappropriate to combine. For example, is it appropriate to combine effects for studies that measure the same dependent variable using different tasks (e.g. measuring the conflict N2 using the Stroop task and the Flanker task)? Should randomized controlled trials, quasi-experimental designs and observational studies be combined in a meta-analysis? The degree to which two studies can differ is not a statistical question; meta-analytic methods will produce a summary effect regardless of the source of the data. It is a substantive issue that will depend largely on the scope of one's meta-analysis. If, for example, a meta-analyst wants to produce a very precise estimate of the effect of conflict on the N2 in the Stroop task, then limiting the meta-analysis to studies that examined the Stroop task is an appropriate course of action. However, one of the primary advantages of meta-analyses is that they are able to address bigger-picture issues than primary studies. In this example, researchers have used many tasks which putatively induce conflict to study the N2 and, thus, many tasks have contributed to our understanding of conflict and the N2. This diversity should be represented in any meta-analysis that purports to summarize the field. Of course, there are circumstances in which the task (or setting/population/recording methods etc.) does systematically impact the results of a study. Another benefit of meta-analysis is that this heterogeneity between studies can be explicitly modeled (described below). A very narrowly-focused meta-analysis will be unable to determine which moderating variables are relevant to the outcome of interest. A meta-analysis that includes a wide range of studies, on the other hand, will be better able to detect moderators and direct future research.

There is no universally-accepted list of specific inclusion/exclusion criteria for meta-analytic work. Indeed, such a list would be unfeasible given that the meta-analytic method can be applied to a wide range of research questions. However, to aid interested psychophysiologists, we recommend the following broad criteria: 1) The meta-analyst should be able to determine the population under investigation and whether that population is relevant to the meta-analysis. 2) The meta-analyst should be able to determine which instruments/tasks were used and whether those instruments/tasks are appropriate to the research question. 3) The meta-analyst should be able to determine how the study was designed and whether the design is appropriate for the meta-analysis (e.g. whether the appropriate control group was used etc.). 4) The article should present enough information to allow for the computation of an effect size. When this information is not readily available, the meta-analyst should attempt to contact the corresponding author of the article. This will help ensure that otherwise acceptable studies are not ignored due to lack of information. Other than these basic criteria, meta-analysts should set inclusion criteria that are appropriate to the research question(s), that are in line with the standards of one's sub-discipline, and that are acceptable to a skeptical scientific audience.

Regardless of the specific criteria, the literature search should be handled in a systematic, principled, consistent and explicit manner. Prior to beginning the literature search, the meta-analyst should determine a set of inclusion/exclusion criteria and develop a checklist/procedure that coders consistently apply throughout the literature search. Additionally, the inclusion/exclusion decision should be made independently of other coding decisions and data extraction (see below) in order to reduce the chance of bias. That is, because the literature search determines which effect sizes do and do not get included, inconsistently applied criteria may result in the systematic exclusion of certain effect sizes (e.g. small effect sizes). Explicit criteria and blind coders will help reduce the possibility of bias. Lastly, the criteria should be made explicit in the manuscript.

During the screening process, the meta-analyst should keep a detailed list of which studies were accepted, which were rejected and the reason(s) why a given study was rejected. This information should be presented as part of the methods. Ideally, the literature search should be carried out by at least two, independent coders. This will allow the meta-analyst to compute inter-coder reliability for their inclusion criteria. If coders cannot reach an adequate level of reliability, it may be necessary to retrain coders or amend the checklist/procedure. Additionally, should a coder miss a relevant study – as is likely to happen when conducting a large literature search – it is possible that the second coder will locate it. Lastly, the methods should describe how consensus was reached when coders disagreed (e.g. discussion, deference to the senior coder, tie-breaking vote, etc.).

3.1. Example using depression and the ERN/FN

Articles were obtained via a search of EMBASE, ProQuest, PubMed, PsycInfo, Scopus, Web of Science, Google Scholar and Psych Filedrawer databases. We crossed the following search terms – depression, major depressive disorder, depressive episode, and depress* – with error-related negativity, ERN, feedback negativity, FN, action-monitoring and response-monitoring. The references sections of all empirical and review articles were systematically searched for additional articles.

3.2. Study selection

Studies were included in the meta-analysis if: 1) The study was available through one of the search databases and printed in the English language. 2) The study included either participants who received a diagnosis of depression using the DSM or ICD classification systems or participants who completed a self-reported measure of depression. 3) Either the response-locked ERN or stimulus-locked FN was recorded. And 4) the study reported enough information to allow for the computation of an effect size. The study selection process is depicted in Fig. 2. These criteria resulted in a total of 28 effect sizes ($N = 1757$ individuals) for the ERN and 18 effect sizes ($N = 878$ individuals) for the FN.

4. Step 3: code studies and extract data

Once the meta-analyst has settled on a list of studies, coders must extract the relevant information including effect sizes and important study characteristics such as population, setting, task, recording parameters (psychophysiological recording system, filter settings etc.) and any other information that is relevant given the research question. This information can be used both to simply describe the studies under investigation as well as to test for potential moderating variables.

The specific variables that a meta-analyst codes, as well as the definition of that variable (e.g. how is a given population defined?) will depend on the substantive research question and the sub-discipline. Here we provide general guidelines for coding studies. 1) A checklist/procedure should be established that provides a list of relevant variables as well as explicit definitions for those variables. Variables and definitions should be presented in the finished meta-analysis. 2) At least two researchers should code each study in order to establish reliability or, if that is not possible, a second researcher should code at least a subset of the studies. 3) If an adequate level of inter-coder reliability cannot be attained, the coders should be retrained or the procedure should be refined. Inter-coder reliability should be presented in the finished meta-analysis. 4) Coding study characteristics should be done prior to or independently from computing effect sizes to ensure that coding decisions are not influenced by study results. Lastly, 5) the variables, definitions and coding procedures should satisfy a skeptical scientific audience.

Once studies have been coded, the results must be transformed to a common metric. When all studies use the same units and those units are easily interpretable, then the meta-analyst may simply synthesize the raw data. For example, a meta-analyst may wish to estimate the extent to which earning a bachelor's degree influences one's income; in a situation like this, it is possible that all studies will report their effects in terms of a common currency. In most cases, however, a meta-analyst will be required to combine data from studies reporting on a variety of dependent measures. In these cases, it is useful to convert the results

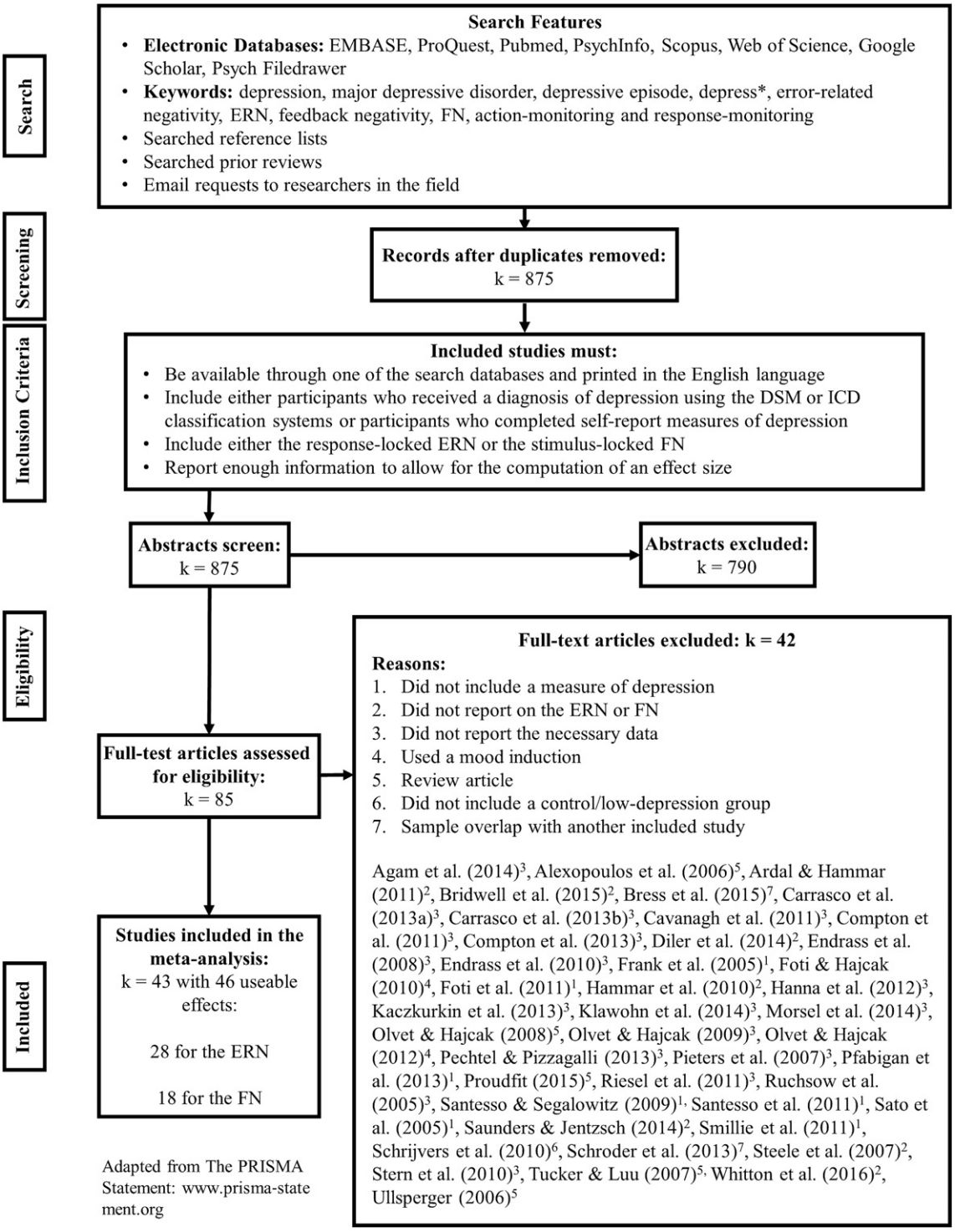

Fig. 2. A flowchart of the literature search and study selection process for the meta-analysis.

to a standardized effect size. The right effect size for a given meta-analysis will be dependent on the substantive question as well as the standards of one's sub-discipline. Given that many psychophysiologists are interested in differences between experimental conditions, naturally existing groups, and in correlations between a psychophysiological measure and some external measure, we will focus on Cohen's *d* and Pearson's r.

In some cases, extracting these effect sizes will be quite straightforward; for example, many researchers will simply present them in the primary study. However, many researchers will not and effect sizes will have to be computed from other, sometimes incomplete, information. In general, a meta-analyst will need enough information to be able to compute both an effect size for each study and a measure of variance for that effect size. To compute these, the meta-analyst will need some/all of the following information:

1) The direction of the effect (i.e. is the effect in the expected direction or in the opposite direction).
2) The sample size.

3) Means and a measure of variance (i.e. standard deviation, standard error etc.). Note that this is only sufficient for between-subjects designs. Means and standard deviations are not sufficient to compute effect sizes for within-subjects designs because of the dependence between means. The correlation between the within-subjects conditions is also needed (see Supplemental Materials).[1]
4) A t, F, r or d value. Note that, when two groups/conditions are being compared, $t = \sqrt{F}$
5) A p value.

In order to aid researchers in computing standardized effect sizes from heterogeneous primary studies that may not all provide the same information, the Supplemental Materials present common methods for computing Cohen's d and Pearson's r.

4.1. Computing effect sizes from p values

Occasionally, rather than presenting the full results of a statistical test, researchers may only present a p value. In such cases it is possible to work backwards and compute the test statistic from the p value. This can be accomplished relatively easily by using the inverse t distribution function available in many statistical software packages (e.g. Microsoft Excel). However, this issue is complicated when the authors of primary studies present summary non-significant effects. For example, a researcher may report: "All other correlations with the N2 were non-significant ($ps > 0.2$)." In this case, we know the effect is non-significant but do not have enough information to compute an exact effect size. When this occurs, it may be necessary to exclude the effect from the meta-analysis; however, doing so may result in non-significant (i.e. smaller) effects being systematically excluded from the meta-analysis. In such cases, the first course of action should be to contact the corresponding author and request the missing data. For cases in which 1) the effect is reported as non-significant, 2) the direction, but not the magnitude, of the effect can be determined and 3) contacting the corresponding author is unsuccessful, we recommend assuming that $p = 0.50$ (see Cooper and Hedges, 1994 for a discussion of this issue). While not ideal, this will help ensure that non-significant effects are represented in the meta-analysis. As an example, let us assume that the author of a primary study wrote the following: "The correlation between anxiety and the N2 was positive but not significant ($N = 25$, $p > 0.2$)." In order to compute the correlation, we will assume $p = 0.50$. Using the inverse t distribution function in Microsoft Excel with 23 degrees of freedom, we can compute the associated t value as 0.685. Then, using the formulae presented in Supplemental Materials, we can convert the t value to a correlation: $r = 0.141$.

4.2. A note on Cohen's d

Cohen's d, as presented above, is a biased estimate of the population effect size. That is, it reliably overestimates the magnitude of the group difference – particularly when the sample size is small. Hedges (1981) first noted this tendency and developed a method to correct for it – known as "Hedges' g". Interested readers may see Hedges (1981) for full computational details; however, Hedges' g can be very closely approximated as follows:

$$g = d \times \left(1 - \frac{3}{4 \times df - 1}\right)$$

where d = Cohen's d (uncorrected), and df = the degrees of freedom.

[1] Combining Cohen's d across between- and within-subjects comparisons is only advisable under certain conditions. We recommend reading Morris and DeShon (2002) for a detailed discussion of this issue.

4.3. An example using depression and the ERN/FN

For all eligible articles, we coded the following information:
1) The ERP component recorded in the study: response-locked ERN or stimulus-locked FN.
2) The type of task that was used to elicit the ERP component.
3) The participant population: patients vs individuals who completed self-report measures of depression.
4) The type of stimuli: whether the stimuli included affective material (e.g. angry faces) or not.
5) Medication status: whether the patient group included individuals receiving medication.
6) Comorbid anxiety: whether the patient group included individuals with comorbid anxiety. Given the high comorbidity rates, if this information could not be determined, it was assumed that the depressed group did include comorbid anxiety.
7) The age of the sample: whether the sample consisted of adults (18 and older) or children (17 and younger).
8) The scoring method: whether the ERN/FN was computed from error/loss trials alone or as the difference between error/loss and correct/gain.

The characteristics of each of the included studies are presented in Table 1 for the ERN and Table 2 for the FN.

4.4. The following decisions were made when coding the primary studies

1) When multiple time-points were collected (e.g. pre- and post-treatment), only the first time-point was included in the meta-analysis.
2) When studies involved a comparison of ERNs/FNs for affectively-neutral material with ERNs/FNs for affective material, only the neutral condition was included.
3) When results from multiple fronto-central electrodes were reported, we computed an effect size that represented the average of the effect sizes for those electrodes.
4) When an effect was reported as non-significant and the information required to compute an effect size could not be obtained, we estimated an effect size assuming $p = 0.5$ in order to ensure non-significant results were not ignored (Cooper and Hedges, 1994).

One author conducted a full extraction of all data; two authors independently verified accuracy. The average inter-coder agreement was high (M Cohen's $\kappa = 0.92$; Range: 0.67–1.0). Disagreement among coders was resolved by discussion.

4.5. Computing effect sizes

The focal effect size for this analysis was Hedges' g. Effect sizes were coded such that a positive value indicates that depressed individuals were characterized by a greater (i.e. more negative) ERP deflection and negative values indicate that depressed individuals were characterized by a reduced (i.e. less negative) deflection. Details regarding the information extracted from each study and the resulting effect sizes can be found in Tables 1 and 2.

5. Step 4: synthesize effect sizes/assess heterogeneity

Once effect sizes have been computed for each study, the individual effects must be synthesized. There are many statistical models with which effects can be synthesized; the current article will focus on two of the most widely used: fixed effects and random effects (FE and RE, respectively) models. When combining effect sizes, the meta-analyst will note that, just like the results of individual participants, the results of individual studies are variable – often substantially so. The major difference between FE and RE models concerns the source of this variability. The FE model assumes that there is a single effect size underlying the observed effects sizes in each study and that the *only* source of

Table 1
Characteristics and effect sizes for studies reporting on the ERN.

Study	Age	Med.	Comorbid	Population	Task	Stimuli	Measure	Data presented	*g*
Aarts et al. (2013)	Adult	Yes	Yes	Clinical	Go/no-go	Neutral	ΔERN	HC: M = 0.7, SEM = 0.9, *N* = 20 Dep: M = 2.09, SEM = 0.8, *N* = 20	0.36
Alexopoulos et al. (2007)	Adult	Yes	Yes	Clinical	Go/no-go	Affective	ERN	HC: M(Fcz) = −3.27, SD(Fcz) = 2.7; M(Cz) = −2.27, SD(Cz) = 2.5; *N* = 6 Dep: M(Fcz) = −5.22, SD(Fcz) = 1.3; M(Cz) = −4.93, SD(Cz) = 1.5; *N* = 6	1.02
Bress et al. (2015)	Child	NA	NA	SR	Flanker	Neutral	ERN	*r* = 0.26; *N* = 25	0.52
Chang et al. (2010)	Adult	NA	NA	SR	Flanker	Neutral	ERN	Peak: *r* = −0.26 Peak-to-peak: *r* = 0.20, *N* = 32	0.46
Chiu and Deldin (2007)	Adult	Yes	Yes	Clinical	Flanker	Neutral	ERN	HC: *N* = 17 Dep: *N* = 18 *F* = 4.6	0.71
Compton et al. (2008)	Adult	NA	NA	SR	Stroop	Both	ERN	*F* < 1, *N* = 34 Assumed *p* = 0.5	0.24
Dywan et al. (2008)	Adult	NA	NA	SR	Source monitoring	Neutral	ERN	r(Fz) = 0.62 r(FCz) = 0.68 r(Cz) = 0.66 *N* = 32	1.68
Endrass et al. (2014)	Adult	Yes	Yes	SR	Flanker	Neutral	ERN	*r* = −0.40, *N* = 72	0.86
Georgiadi et al. (2011)	Adult	Yes	No	Clinical	Go/no-go	Neutral	ERN	Acute vs. HC comparison not presented. *N* = 34, assumed *p* = 0.5	−0.23
Grundler et al. (2009)	Adult	NA	NA	SR	Flanker	Neutral	ΔERN	*r* = 0.32, *N* = 36	0.66
Holmes and Pizzagalli (2008)	Adult	No	Yes	Clinical	Stroop	Affective	ERN	HC: M = −1.42 SD = 1.69 *N* = 18 Dep: M = −2.74 SD = 1.83 *N* = 18	0.73
Holmes and Pizzagalli (2010)	Adult	No	Yes	Clinical	Stroop	Affective	ERN	HC: M = −2.94 SD = 1.77 *N* = 17 Dep: M = −4.65 SD = 2.86 *N* = 13	0.72
Kalayam and Alexopoulos (2003)	Adults	Yes	Yes	Clinical	Stroop	Neutral	ERN	HC: M(left) = −4.84 SD(left) = 1.84 M(right) = −6.58 SD(right) = 2.91 *N* = 13 Dep: M(left) = −8.12 SD(left) = 4.01 M(right) = −4.90 SD(right) = 2.51 *N* = 9	0.25
Ladouceur et al. (2012)	Child	No	Yes	Clinical	Flanker	Neutral	ΔERN	HC: *N* = 14 Dep: *N* = 24 F = 10.76	−1.08
Moran et al. (2012)	Adult	NA	NA	SR	Flanker	Neutral	ERN	*r* = 0.09, *N* = 146	0.18
Olvet et al. (2010)	Adult	No	Yes	Clinical	Flanker	Neutral	ERN	HC: M = −2.14 SD = 1.65 *N* = 22 Dep: M = −2.31	0.17

(continued on next page)

Table 1 (*continued*)

Study	Age	Med.	Comorbid	Population	Task	Stimuli	Measure	Data presented	*g*
								SD = 1.93 *N* = 22	
Ruchsow et al. (2004)	Adult	Yes	No	Clinical	Flanker	Affective	ΔERN	Not provided; assumed *p* = 0.5	−0.24
Ruchsow et al. (2006)	Adult	Yes	No	Clinical	Flanker	Neutral	ΔERN	Not provided; assumed *p* = 0.5	−0.30
Schoenberg (2014)	Adult	Yes	Yes	Clinical	Go/no-go	Neutral	ERN	HC: M(male) = −13.3 SD(male) = 7.2 M(female) = −7.9 SD(female) = 5.5 *N* = 26; Dep: M(male) = −6.8 SD(male) = 5.7 M(female) = −9.4 SD(female) = 7.8 *N* = 55	−0.25
Schrijvers et al. (2008)	Adult	Yes	Yes	Clinical	Flanker	Neutral	ERN	HC: *N* = 25 Dep: *N* = 26 F = 1.18	−0.30
Schrijvers et al. (2009)	Adult	Yes	Yes	Clinical	Flanker	Neutral	ERN	HC: M(Fz) = −9.7 SD(Fz) = 5.5 M(Cz) = −8.5 SD(Cz) = 3.9 *N* = 15 Dep: M(Fz) = −8.6 SD(Fz) = 5.0 M(Cz) = −7.8 SD(Cz) = 4.7 *N* = 15	−0.18
Tang et al. (2013)	Adult	Yes	Yes	Clinical	Flanker	Neutral	ERN	HC: *N* = 24 Dep: *N* = 19 F = 5.68	0.72
Weinberg et al. (2016b)	Adult	NA	No	Clinical	Flanker	Neutral	ERN	HC: M = −0.24 SD = 6.42 *N* = 55 Melancholic: M = 3.02 SD = 10.53 *N* = 17 Dep: M = 1.21 SD = 9.96 *N* = 33	Melan.: −0.43 Dep: −0.18
Weinberg et al. (2010)	Adult	Yes	NA	NA	Flanker	Neutral	ERN	*r* = −0.52, *N* = 35	1.19
Weinberg et al. (2012)	Adult	No	Yes	Clinical	Flanker	Neutral	ERN	HC: M = 4.18 SD = 7.74 *N* = 36 Dep: M = 3.32 SD = 5.72 *N* = 23	0.12
Weinberg et al. (2016a)	Child	NA	NA	SR	Flanker	Neutral	ERN	*r* = 0.10, *N* = 515	−0.20
Weinberg et al. (2015)	Adult	Unknown	Yes	Clinical	Flanker	Neutral	ERN	HC: M = 0.38 SD = 5.43 *N* = 56 Dep: M = −0.66 SD = 5.57 *N* = 62	0.19

Key: Adult: Studies including participants age 18 and over; Child: Studies including participants below 18; SR: Studies using self-report measures of depression; Clinical: Studies using diagnoses of depression; Neutral: Studies using neutral stimuli; Affective: Studies using affective stimuli; ERN: Studies reporting on brain activity recorded following errors; ΔERN: Studies reporting on the difference wave between error and correct trials.

variability between studies is sampling error. The RE model, on the other hand, assumes that observed studies are being drawn from a population of possible studies and the meta-analysis aims to estimate the mean of this population.

The choice of model is not trivial. Using the FE model under conditions in which the RE model would be more appropriate has two primary consequences. First, the summary effect will be incorrect – specifically, it will be unduly influenced by large studies. Since the FE

Table 2
Characteristics and effect sizes for studies reporting on the FN.

Study	Age	Population	Task	Data presented	*g*
Bress et al. (2012)	Child	SR	Gambling	$r = 0.38$, $N = 64$	−0.81
Bress et al. (2013)	Child	Clinical	Gambling	HC: $N = 52$ Dep: $N = 16$ $t = 2.04$	−0.58
Bress et al. (2015)	Child	SR	Gambling	$r = 0.54$, $N = 25$	−1.24
Foti and Hajcak (2009)	Adult	SR	Gambling	$r = 0.23$, $N = 85$	−0.53
Foti et al. (2014)	Adult	Clinical	Gambling	HC: M = −4.90 SD = 3.43 $N = 42$ Dep: M = −2.69 SD = 4.39 $N = 34$	−0.56
Grundler et al. (2009)	Adult	SR	Prob. Learning	Study 1: $r = -0.39$, $N = 39$ Study 2: $r = -0.07$, $N = 30$	Study 1: −0.83 Study 2: −0.14
Liu et al. (2014)	Adult	Clinical	Gambling	HC: M = −7.89 SD = 4.91 $N = 27$ Dep: M = −0.66 SD = 4.67 $N = 27$	−1.49
Mies et al. (2011)	Adult	Clinical	Time estimation	HC: $N = 28$ Dep: $N = 15$ F = 7.94	0.94
Mueller et al. (2015)	Adult	Clinical	Reinforce. Learning	HC: M = 0.57 SEM = 0.41 $N = 15$ Dep: M = 1.75 SEM = 0.24 $N = 14$	0.88
Padrao et al. (2013)	Adult	SR	Gambling	HC: $N = 22$ Dep: $N = 21$ $t = 1.44$	0.43
Peng et al. (2015)	Adult	SR	Time estimation	Not provided assumed $p = 0.5$	0.22
Ruchsow et al. (2004)	Adult	Clinical	Flanker	Not provided; assumed $p = 0.5$	0.24
Ruchsow et al. (2006)	Adult	Clinical	Flanker	Not provided; assumed $p = 0.5$	−0.30
Santesso et al. (2008)	Adult	Clinical	Reinforce. Learning	HC: $N = 15$ Dep: $N = 12$ F = 7.5	1.03
Tucker et al. (2003)	Adult	Clinical	Spatial compatibility	HC: $N = 27$ Dep: $N = 20$ F = 3.68	0.86
Weinberg and Shankman (in press)	Adult	Clinical	Gambling	HC: M = −4.28 SD = 3.69 $N = 81$ Melancholic: M = −2.01 SD = 3.35 $N = 29$ Dep: M = −4.61 SD = 3.87 $N = 56$	Melancholic: −0.62 Dep: 0.09

Key: Adult: Studies including participants age 18 and over; Child: Studies including participants below 18; SR: Studies using self-report measures of depression; Clinical: Studies using diagnoses of depression.

model assumes that the only source of variability is sampling error, the FE model assigns very high weights to studies with large sample sizes. Thus, a few large studies can dominate the summary effect whereas smaller studies are virtually ignored. In the RE model, on the other hand, smaller studies are assigned higher weights (relative to the FE model) and receive more representation in the meta-analytic mean. Second, the standard error will be underestimated. Within the FE model, the standard error of the mean depends on the within-study variance. In contrast, the standard error in the RE model depends on both the within-study variance and the between-study variance. Thus, unless there is very little variability between studies, the RE standard error will be larger than the FE standard error. Underestimating the standard

error means 1) that significance tests are more likely to produce significant results – even spuriously – and 2) that confidence intervals will be too narrow and will yield a false sense of precision in the estimate.

How should a meta-analyst choose between FE and RE models? This decision should be made on both statistical and theoretical grounds. Statistically, the FE model assumes that all variability between studies is due to sampling error. Thus, if measures of heterogeneity (described below) indicate substantial variability – especially variability beyond what is predicted by sampling error – then it is likely that the FE model is inappropriate. However; measures of heterogeneity should not be the sole criterion on which the decision is based; some measures of homogeneity are significance tests that may be underpowered to detect between-study variability. Thus the choice of a model should also be justified theoretically – that is, does the model match our understanding of the underlying effect? If all studies were conducted using the same task, types of stimuli, participant population and psychophysiological recording equipment and procedures, then the FE model may be appropriate. If studies differ on these (or other) key dimensions, then it is likely that the RE model is more appropriate. When in doubt, we recommend the use of the RE model. As noted before, the FE model assumes that a single population effect underlies the individual effects whereas the RE model allows the effect to vary (but does not assume that it must vary). Mathematically, the RE model reduces to the FE model as the between-study variability approaches zero. Thus, there is no cost to using the RE model; when variability is high, it will be more accurate than the FE model and when variability is low it will be nearly identical to the FE model.

In addition to determining the summary effect; researchers are often interested in estimating how stable an effect is across different studies (i.e. across populations, tasks, etc.). Thus, it is often useful to present measures of heterogeneity. While there are several measures of heterogeneity, this article will focus on three measures which index between-study variability above-and-beyond what would be expected by sampling error: *Q*, meta-regression and I^2.

Q is a standardized measure of the variability between studies (see Borenstein et al., 2009). Q typically has two general uses: first, it can be used to determine if the between-study variability for the entire sample of studies is greater than what sampling error would predict. When used this way, Q follows a chi-square distribution with k − 1 degrees of freedom, where k is the number of studies, and the meta-analyst would treat this like a standard significance test. Second, the Q test can be used to test the null hypothesis that effect sizes are the same across some hypothesized moderating variable. For example, suppose a researcher was interested in determining whether an effect was similar for men and women. The Q test could be used to determine whether the difference in effect sizes for men and women was greater than would be predicted by sampling error. When used in this way, Q still follows a chi-square distribution with k − 1 degrees of freedom; but k is now the number of groups being compared (k = 2 in the this example). When the potential moderating variable is continuous, rather than categorical, the significance of the moderator can be assessed using a "meta-regression" – essentially a weighted regression analysis. For example, the "decline-effect" can be evaluated by using the year of publication to predict effect sizes.

In addition to testing whether the between-studies variability is statistically significant, a researcher may want an index of the overall size of the variability. In such cases, I^2 is a useful measure. I^2 is the percentage of "real" variation between effects – that is, the percentage of between-study variance that is not attributable to sampling error. As a rule of thumb, Higgins et al. (2003) proposed that I^2 values of 25, 50 and 75 should be used as cutoffs for low, medium and high variability, respectively.

5.1. Example using depression and the ERN/FN

The included studies involved a number of different participant populations, depression assessment instruments, cognitive tasks and EEG recording hardware and parameters. Additionally, effect sizes for both the ERN ($I^2 = 67.41$; $Q(27) = 82.85$, $p < 0.001$) and FN ($I^2 = 81.05$; $Q(17) = 89.70$, $p < 0.001$) were very heterogeneous. Thus, effect sizes were pooled using the random-effects model to allow for heterogeneity between studies. All computations were conducted using Comprehensive Meta-Analysis Software and G*Power Software.

5.2. ERN – summary effect

Overall, the association between depression and the ERN was statistically significant and relatively small ($g = 0.21$; $k = 28$; $p = 0.03$; 95% CI: 0.02; 0.40). The ERNs of depressed individuals was approximately a fifth of a standard deviation larger (i.e. more negative) than the ERNs of control participants.

5.3. ERN – moderator analyses

Moderator analyses for the ERN are presented in Table 3. Most of the moderator analyses did not reach significance. However, it is noteworthy that effect sizes did differ markedly across many of the moderators. For example, although not significant, effects were largest for student samples (relative to patient samples), the Stroop task (relative to the Flanker and go/no-go tasks), and studies that reported the ERN (relative to the difference wave). Interestingly, there was a significant effect of age ($p = 0.049$). Effect sizes for adults and children were of roughly equal magnitude but in opposite directions. Lastly, we used the year of publication to predict Hedges' *g* to assess for the "decline effect" as well as to demonstrate a meta-regression. This regression produced an intercept of 32.92 and a slope of −0.02; the regression was not significant ($Q(1) = 0.45$, $p = 0.50$). Thus, there was no evidence for a decline effect in these data.

5.4. FN – summary effect

Overall, FNs were smaller (i.e. less negative) in depression. However, the effect was small and not significant ($g = -0.14$; $k = 18$; $p = 0.39$; 95% CI: −0.47; 0.18).

5.5. FN – moderator analyses

Given the small number of studies, the only moderators we tested for the FN were Population, Type of Task and Age. These are shown in Table 4. The type of task that was used was a significant moderator for the FN. Studies using a version of the gambling/guessing task found that depression predicted a smaller (less negative) FN whereas studies using other tasks (e.g. probabilistic learning tasks) found that

Table 3
Moderator analyses for the error-related negativity.

Moderator	*g*	95% CI	k	N	*p*	Q (df)	*p*
Population						1.97 (1)	0.16
Patient	0.07	−0.16; 0.29	19	830	0.57	–	–
Student	0.37	0.01; 0.73	7	820	0.04	–	–
Task						1.91 (2)	0.39
Flanker	0.13	−0.08; 0.34	19	1436	0.22	–	–
Go/no-go	0.09	−0.39; 0.57	4	167	0.71	–	–
Stroop	0.50	0.01; 0.99	4	122	0.05	–	–
Affect						1.09 (1)	0.30
Affective	0.44	−0.03; 0.92	5	144	0.07	–	–
Neutral	0.17	−0.03; 0.37	23	1613	0.10	–	–
Medication						0.76 (1)	0.78
On	0.18	−0.15; 0.52	12	472	0.28	–	–
Off	0.27	−0.20; 0.74	6	242	0.25	–	–
Scoring						2.11 (1)	0.15
ERN	0.27	0.07; 0.48	23	1591	0.01	–	–
ΔERN	−0.11	−0.57; 0.36	5	166	0.66	–	–
Age						3.87 (1)	0.05
Adult	0.27	0.08; 0.46	25	1179	0.01	–	–
Child	−0.28	−0.80; 0.24	3	578	0.29	–	–

Table 4
Moderator analyses for the feedback negativity.

Moderator	*g*	95% CI	k	N	*p*	Q (df)	*p*
Population						1.37 (1)	0.24
Patient	0.02	−0.41; 0.44	11	554	0.94	–	–
Student	−0.39	−0.92; 0.14	7	324	0.19	–	–
Task						10.27 (1)	0.001
Gambling	−0.56	−0.93; −0.19	9	581	0.003	–	–
Other	0.34	−0.07; 0.73	9	297	0.10	–	–
Age						0.95 (1)	0.33
Adult	−0.44	−0.87; −0.01	6	424	0.04	–	–
Child	−0.83	−1.49; −0.17	3	157	0.01	–	–

depression non-significantly predicted a larger (more negative) FN. Population was not a significant moderator but the effect was notably larger for studies including non-clinical samples. Lastly, we initially aimed to test the effects of age across the entire sample. However, age was partially confounded with task – studies with children used the gambling task. Given that other tasks produced effects in the opposite direction, the moderator test for age focused only on studies using the gambling task. This analysis did not find evidence for moderation by age.

6. Step 5: assess for threats to validity

As with any statistical procedure, the results of a meta-analysis are only as good as the quality of the original data and the methodology employed by the researcher. Meta-analysis can be a powerful tool for summarizing a literature and for answering broader questions than any primary study; however, problems with the available data and the meta-analytic techniques can result in severely distorted findings. Thus, just as manipulation checks are necessary in many primary studies, assessing for threats to validity are needed in meta-analysis. Many of these threats have already been implicitly addressed. For example, Step 2 described excluding poorly conducted studies and Step 4 described choosing the correct model for the data. This section focuses on two additional issues: statistical power and publication bias.

6.1. Statistical power

Power refers to the probability that a significance test will detect a significant effect when that effect is actually present. The procedure for conducting power analyses and the use of a power analysis are very similar across primary studies and meta-analyses. First, it is often useful to conduct a power analysis prior to beginning data collection or searching the literature as part of the study-planning phase (see Larson and Carbine, in press). To do so, the meta-analyst will need to first determine a target effect size (i.e. determine an effect size that they expect to find in the literature). There are a number of ways to do this. For example, a meta-analyst could compute a mean effect size for a small subset of studies in order to estimate what the meta-analytic effect will be. Alternately, meta-analysts could use rules of thumb for effect size magnitudes (e.g. small/medium/large values for Cohen's *d*; Cohen, 1988) and estimate the necessary number of studies to achieve high power for typical effect size magnitudes in their field.

Second, non-significant effects in meta-analysis, as in primary studies, need to be interpreted in the light of the study's power. That is, if a small meta-analysis fails to find an effect, it could reasonably be argued that the meta-analysis was simply underpowered. In situations such as these, researchers will often compute the power to detect the achieved effect size. However, demonstrating that a non-significant test had low power to detect the achieved effect is simply a restatement of the significance test. A more useful procedure would be to determine what effect sizes the meta-analysis was sufficiently powered to detect and whether effect sizes smaller than that are theoretically or practically important.

So far, the discussion of power analysis as focused only on detecting main effects. It is also possible to conduct a power analysis for tests of heterogeneity and moderation. We direct interested readers to the following texts that address this issue in more detail: Hedges and Pigott (2001, 2004).

6.2. Publication bias

Publication bias occurs when the published literature on a given topic systematically differs from the population of all studies conducted on that topic. For example, psychologists are becoming increasingly aware that studies with small effects (or effects in the "wrong" direction) and, therefore, non-significant findings, are less likely to be published. The net result of publication bias is that the published literature will tend to overestimate the population effect size, sometimes dramatically. There is no universally-recognized gold standard for assessing publication bias. Therefore, this section will describe several commonly-used metrics for determining whether publication bias is present and how to address publication bias when conducting a meta-analysis.

6.3. Funnel plot

Many modern methods for detecting publication bias involve examining a funnel plot. Fig. 3 displays example funnel plots created using fictional data. The effect size for each individual study is plotted on the X-axis and the standard error for that effect size is plotted on the Y-axis. Note that the Y-axis for funnel plots, like the Y-axis for many ERP studies, is inverted. Large N studies (i.e. those with small standard errors) cluster near the top of the graph whereas smaller studies spread across the bottom of the graph due to larger sampling variability. The meta-analytic mean, $g = 0.35$, is marked by the vertical line. When no publication bias in present, effect sizes should be dispersed symmetrically around the mean. This is because sampling error is expected to have a mean of zero. When a meta-analysis is contaminated with publication bias, however, this is not the case. Effects sizes are expected to be symmetrical only for large N studies because these studies may be published even if the finding is non-significant (e.g. the authors may have shown that their statistical power was more than sufficient). For smaller studies, large effect sizes are likely to be published whereas smaller effect sizes may be rejected due to insufficient power. Thus, publication bias will often result in an asymmetrical funnel plot such that studies with small sample sizes and small effect sizes will be systematically missing. This is evident in panel A of Fig. 3. While studies with large Ns/small standard errors are roughly symmetrical around the mean, studies with both small Ns and effect sizes are notably absent.

Although the asymmetry in Fig. 3 is obvious, it may be very difficult to detect by visual inspection in real datasets. There are currently two commonly-used objective tests for detecting the asymmetry. Both rely on the fact that the asymmetrical distribution of effect sizes produced by publication bias results in a correlation between the magnitude of an effect and its standard error (see Fig. 3; panel A). First, Kendall's τ (tau) tests this relationship directly and can be interpreted like a correlation. Values differing significantly from 0 indicate the presence of bias. For the dataset presented in Fig. 3, Kendall's $\tau = 0.43$, $p < 0.001$ which suggests that the effect size tends to increase along with the standard error. Second, Egger's test (Egger et al., 1997) produces a regression intercept. Since each effect size is also normalized, the expected intercept is 0; significant deviations from 0 are taken as evidence of bias. For the example dataset, Egger's intercept = 2.07, $p < 0.001$. Meta-analysts should note that Kendall's and Egger's tests are generally low-powered; thus, non-significant findings should be interpreted with caution and in the light of other tests.

In addition to determining whether publication bias is present, researchers may also be interested in determining the extent to which publication bias is influencing the results. Duval and Tweedie's trim

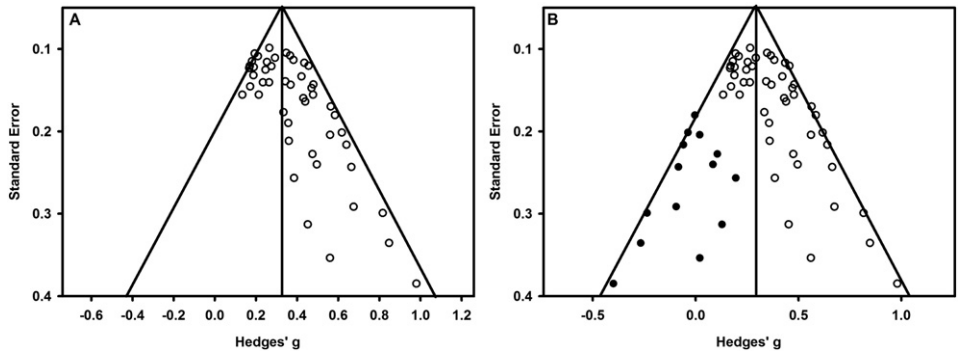

Fig. 3. A: A funnel plot depicting publication bias in the fictional dataset. B: A funnel plot depicting the fictional data set following Duval and Tweedie's trim and fill procedure. Unfilled circles represent observed studies whereas filled circles represent imputed studies. Note that the y-axis of a funnel plot is inverted.

and fill procedure accomplishes this by iteratively removing the studies with the largest effects and smallest Ns and re-computing the mean. This procedure continues until the studies are approximately symmetrical around the mean. The new mean produced by this procedure is expected to be unbiased; however, removing the extreme studies artificially decreases the between-study variance. To deal with this, the procedure also adds the trimmed studies back in and then "fills" in the missing studies by adding in the mirror-image of the original studies across the mean. This is shown in panel B of Fig. 3. The procedure identified 14 studies to be trimmed and filled. The new studies have been imputed and the meta-analytic mean has been reduced slightly ($g = 0.28$). This example also demonstrates that the mere presence of publication bias is not enough to nullify an effect; the change in the meta-analytic mean may be small, as it was in this case.

6.4. Excess significance

Even when an effect is real and reliable, a given study may still fail to produce significant results – often called a "Type II Error". That is, even true effects should fail to replicate some of the time. The probability that a study will successfully detect an effect is dependent on many factors; the most important of which (for the current discussion) is power. Studies should only detect effects at a rate that is consistent with their power to do so.

As an example, imagine a researcher who is interested in examining the size of the ERN across men and women. The population effect size in this example is $d = 0.66$. Data from 19 men and 19 women are collected and a significant difference is found. Before publishing her finding, the researcher decides to replicate the finding in an independent sample of 19 men and 19 women; however, this study fails to reach significance. What are we to make of this failed replication? A simple power analysis tells us that a sample size of 38 results in approximately 50% power to detect a population effect of $d = 0.66$ – that is, whether these studies would detect a significant effect was, essentially, a coin flip. Thus, one significant finding and one non-significant finding is consistent with what we should expect to find given the power of these studies. Now imagine that this researcher decides to consult the published literature and see if other researchers have reported on sex differences in the magnitude of the ERN. She finds 10 studies. All of which report on 19 men and 19 women; since all of these studies have 50% power, we should expect that only approximately 5 of the studies will report significant effects. However, 9 out of the 10 report significant results. In a situation such as this – i.e. when the literature consists of more significant findings than power allows – it might be the case that other, non-significant, studies were conducted but never published.

Ioannidis and Trikalinos (2007) leveraged the fact that studies should occasionally fail to replicate to develop a procedure for detecting excess success. This procedures compares the observed number of significant findings with the expected number of significant findings based on the power. Determining the observed number of significant findings is quite straightforward – the meta-analyst can simply count the number of significant findings in the literature. Determining the expected number of significant findings requires first computing the power for each study. Unless the meta-analyst knows the population effect size, the best estimate of the population effect size is the meta-analytic effect size. The expected number of significant findings can then be computed as the sum of each study's power. Lastly, the excess in significant studies can be tested using a chi-square test or a binomial test (especially when the number of studies is small).

In the example presented above, 9 out of 10 studies produced a significant finding whereas the expected number of significant findings is 5 (i.e. $0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5$). Given there were only 10 studies, we submitted these findings to a binomial test; this indicated that there were more significant findings than one would expect based on power ($z = 2.21$, $p = 0.02$).

6.5. Unpublished studies

Lastly, it may be possible to reduce the impact of publication bias by expanding one's literature search beyond the published literature. As noted earlier, there are a number of repositories hosting unpublished work (e.g. "Psych file drawer") as well as unpublished master's theses and doctoral dissertations (ProQuest). Including unpublished work allows a meta-analyst to directly assess whether published and unpublished work systematically differ. Assuming that the unpublished studies meet the other inclusion criteria and quality-control criteria, we strongly recommend including unpublished work in a meta-analysis whenever possible.

6.6. Example using depression and the ERN/FN

We calculated the power for our meta-analyses as a function of Hedge's g and the ERP component. As shown in Fig. 4, the present meta-analysis had sufficient power (≥80%) to detect significant values of $g \geq 0.26$ for the ERN. With respect to the FN, the present meta-analysis had sufficient power to detect significant values of $g \geq 0.46$. For values of g below 0.46, the present meta-analysis could be considered underpowered. Interested readers will need to determine whether effects below 0.46 are of theoretical and clinical importance – especially considering that primary studies will need to include 76 control and 76 depressed participants to achieve 80% power based on this analysis. However, this finding should be interpreted with caution given the large moderator effect reported in the previous section.

Fig. 4 also demonstrates one of the strengths of meta-analyses: increased power. On average, the primary studies had sufficient power to detect effects of $g > 0.90$ and $g > 0.80$, for the ERN and FN, respectively (compared to 0.26 and 0.46 for the meta-analyses).

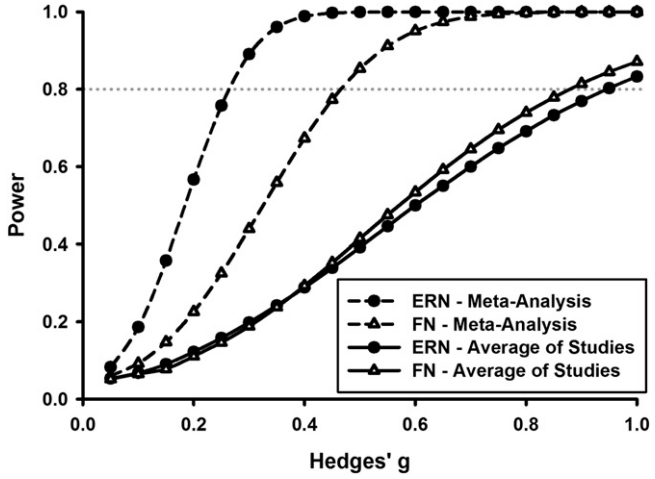

Fig. 4. Power analysis for the meta-analyses and the primary studies. The grey broken line highlights the range of *g* values for which 80% power would be achieved.

6.7. Publication bias

All indices of publication bias agreed that the association between depression and the ERN has been contaminated by publication bias. Both Kendall's tau ($\tau = 0.27$, $p = 0.04$) and Egger's intercept ($B = 1.78$, $p = 0.01$) were significant indicating that effect sizes were generally smaller when studies had greater sample sizes.

Based on the power of each study to detect an effect of $g = 0.21$, we would expect 3.55 studies to reach significance. However, 8 of the effect sizes that we computed reached significance. Both a binomial test ($z = 2.24$, $p = 0.01$) and a chi-square test ($\chi^2(1) = 6.35$, $p = 0.01$) indicated that these 8 findings are more than we would expect based on the power.

Lastly, Duval and Tweedie's trim and fill procedure indicated that 5 studies would be necessary to make the funnel plot symmetrical (see Fig. 5). Once those 5 studies were imputed, the meta-analytic effect was substantially reduced ($g = 0.06$, 95% CI: −0.13; 0.26). Overall then, multiple metrics converge to indicate the presence of publication bias with respect to the ERN. Furthermore, when we attempted to correct for the bias using Duval and Tweedie's trim and fill procedure, the effect was reduced to a trivial magnitude.

Given that the mean effect for the FN was non-significant and moderated by the type of task, we did not conduct publication bias analyses on the FN. That is, studies using different types of tasks are estimating different effect sizes in opposite directions and it is possible that publication bias metrics are insensitive to publication bias in the overall sample. Since the number of studies is quite small, this issue will have to await future research.

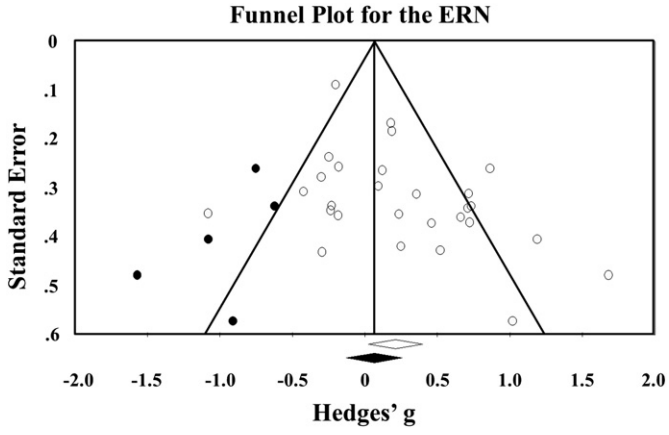

Fig. 5. Funnel plot for the ERN. Observed studies are depicted with unfilled circles; imputed studies are depicted with filled circles. Unfilled circles represent observed studies whereas filled circles represent imputed studies. Note that the y-axis of a funnel plot is inverted.

7. Discussion

Meta-analysis is a powerful statistical tool that can aid researchers in quantitatively summarizing a field, testing hypotheses, directing future research and assessing the field for potential threats to validity such as publication bias. Our goal in writing this tutorial was to point out the potential usefulness of meta-analysis to psychophysiologists unfamiliar with the method as well as to provide practical advice and best-practices to those psychophysiologists interested in conducting their own meta-analyses. We encourage psychophysiologists to use meta-analysis to address substantive questions whenever appropriate.

Meta-analysis is a very broad topic and no article-length tutorial can cover all relevant topics. We recommend the following tutorials to interested readers:

1) Borenstein et al. (2009). Introduction to Meta-Analysis, John Wiley & Sons, Ltd. West Sussex, UK.
2) Cooper (2010). Research Synthesis and Meta-Analysis: A Step-by-Step Approach. Sage Publications Inc. Thousand Oaks, CA.
3) Cumming (2012). Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Taylor & Francis, LLC: New York, NY.

Moreover, we focused only on fixed- and random-effects analyses. While these are among the most commonly-used methods in psychology, there are alternative methods including psychometric meta-analysis and Bayesian meta-analysis. We direct interested readers to the following:

1) Eddy et al. (1992). Meta-analysis by the confidence profile method. The statistical synthesis of evidence. Academic Press. Boston, MA.
2) Hunter and Schmidt (2004). Methods of Meta-Analysis: Correcting Error and Bias in Research Findings (2nd Ed.). Sage Publications Inc. Newbury Park, CA.
3) Schmid and Mengersen (2013). Bayesian Meta-analysis. In Koricheva, J., Gurevitch, J., & Mengersen, K (Eds.), Handbook of Meta-Analysis in Ecology and Evolution. Princeton University Press. Princeton, NJ.
4) Smith et al. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. Statistics in Medicine, 14, 2865–2699.

Finally, Sambrook and Goslin (2014) presented a meta-analytic method that will be of particular interest to ERP researchers. Their "great grand averaging" procedure involves gathering figures from published papers containing grand averaged ERP waveforms – in their example, displaying the FN – and uploading the figures into PlotDigitizer (http://sourceforge.net/projects/plotdigitizer/;as always, only download files from trusted sources), a program that digitizes the waveforms in order to extract their individual data points. They then averaged across these grand averages to create what they call "great grand averages" which allowed them to score the FN using the digitized data points for each study. Since this method can, in principle, be used with any published waveform, this method has the advantage of being less reliant on authors publishing/providing the necessary effect size information. We hope this method will become more widely used, and refined as a robust meta-analytic tool.

7.1. Example using depression and the ERN/FN

In addition to providing an introductory tutorial on meta-analysis, the present article presented a novel meta-analysis between depression and the ERN/FN as an illustrative example. Specifically, both the ERN

and FN are proposed biomarkers for depression. However, previous findings have been inconsistent for both ERPs. Given the potential clinical applications, the goals of the present meta-analysis were to 1) determine the magnitude and direction of the effect across the literature, 2) evaluate the heterogeneity of effects and the role of potential moderators and 3) assess for publication bias in the literature.

7.2. ERN

With respect to the first goal, this meta-analysis found evidence that depression is related to a moderately enhanced ERN. With respect to the second goal, we found that effect sizes were very heterogeneous suggesting that the effects were not entirely consistent across studies (i.e. between-study variability above-and-beyond sampling error). Although the specific moderator analyses generally did not reveal significant moderators, effect sizes varied widely for each moderator. For example, effect sizes were notably larger for students (relative to patients), the Stroop task (relative to the Flanker and go/no-go) and affective stimuli (relative to neutral stimuli). Additionally, effects for adults and children were of approximately equal magnitude but in opposite directions. This degree of heterogeneity suggests that the ERN is unlikely to be related to depression in a very general way and highlights the need to study the relationship between the ERN and depression using specific tasks and sub-populations. For example, a recent report by Weinberg et al. (2016b) found that the ERN is blunted in individuals diagnosed with depression with melancholic features specifically (however, see Aarts et al., 2013). Although several articles in this meta-analysis included individuals with melancholic features, we were not able to compute effect sizes for them separately. Future research may benefit from examining the role of sub-types of depression in determining the depression/ERN relationship.

Lastly, with respect to the third goal, this meta-analysis found evidence of publication bias across multiple metrics. Both Kendall's tau and Egger's intercept were consistent with publication bias and the number of statistically significant findings was larger than power would allow. Additionally, we estimated the unbiased effect size for the ERN using Duval and Tweedie's trim and fill procedure. This resulted in a considerably smaller, and non-significant, relationship between depression and the ERN ($g = 0.06$). This finding, in conjunction with the results discussed above, suggests caution when discussing the clinical utility of the ERN. Specifically, the high heterogeneity and small corrected effect suggest that the evidence for an association between depression and the ERN is weak.

An important note of caution is warranted, however, when considering the publication bias results. This set of findings suggests that depression and the ERN are not strongly related in general. However, as noted in the section of heterogeneity, there may be more specific relationships to uncover (e.g. unmedicated, melancholic patients performing tasks involving affective material). The number of studies was insufficient for examining publication bias across all moderators; however future studies may benefit from examining the specificity of this relationship.

7.3. FN

In addition to examining the ERN, the current study aimed to quantify the relationship between depression and the FN. With respect to the first goal, this meta-analysis found that, overall, depression is not related to the FN. The effect size was quite small ($g = -0.14$) and the confidence interval was extremely wide (−0.47–0.18).

With respect to the second goal, effect sizes were found to be very heterogeneous. Perhaps the most surprising finding of this meta-analysis was the large role that the type of task played in determining the relationship between depression and the FN. Studies using gambling/guessing tasks have found that the FN is over half of a standard deviation smaller in depressed participants whereas studies using other tasks (e.g. reinforcement learning) have found that the FN is a third of a standard deviation larger for depressed participants. At present we can only speculate about the role that task plays. Hajcak and colleagues (Foti and Hajcak, 2009; Foti et al., 2014; Proudfit, 2015) have shown that positive feedback in the gambling task, which is usually associated with a monetary reward, produces a positive deflection – the reward positivity – approximately 300 ms following feedback onset. Negative feedback does not result in a positivity. In other words, Hajcak and colleagues propose that the FN actually results from the absence of a positive component. Importantly, it is the reward positivity that appears to be related to depressive symptoms (Foti and Hajcak, 2009; Foti et al., 2014; Liu et al., 2014; Proudfit, 2015). To the authors' knowledge, the ERP componentry of feedback in other tasks, such as reinforcement learning tasks, has not been systematically studied. Given that these tasks often involve learning in the absence of an explicit reward, it may be that the reward positivity is not elicited by these tasks or, if it is, that it indexes another process.

7.4. Recommendations for authors of primary studies

Finally, even authors and editors who do not wish to conduct a meta-analysis themselves can still contribute to cumulative science. Specifically, by clearly and completely presenting results, authors of primary studies can facilitate future meta-analyses. Journal editors can contribute by enforcing the complete reporting of results. Specifically, we recommend the following:

1) It is standard practice in the ERP literature to present waveforms. Although we agree that waveforms should be regularly presented in published articles, they should not be included in lieu of explicit means and standard deviations. Means and an explicitly-labeled measure of variability should be included for all measures.
2) Non-significant effects should be fully presented. For example, instead of "The main effect of group was not significant ($p > 0.5$)", authors of primary studies should write "The main effect of group was not significant ($t(38) = 0.54$, $p = 0.79$)."
3) For within-subject designs, the mean and standard deviation of the difference scores should be presented or the correlation between dependent measures should be presented. This will allow future meta-analysts to compute the within-subjects effect size.
4) For correlational designs, a correlation table with all measures should be presented in the manuscript.
5) Finally, in many cases, it is simply not possible for the authors of primary studies to include all of the information that might be of interest to future meta-analysts. In such cases, we believe the best course of action is to contact the corresponding author of the primary study and request the necessary information. This will help ensure that otherwise acceptable data are not ignored by meta-analyses; however, this procedure is reliant on the cooperation of those authors. In order to facilitate a cumulative and open science, we recommend that these authors supply the necessary effect sizes whenever possible.

8. Conclusion

Meta-analysis is a powerful tool for quantitatively summarizing an existing literature. The present article aimed to introduce meta-analysis and provide a step-by-step tutorial aimed at psychophysiologists which included: formulating the problem, conducting the literature search, coding studies and extracting data, synthesizing effect sizes and assessing for heterogeneity, and assessing for threats to validity. Each of these steps was accompanied by a substantive example in which we meta-analyzed the relationship between depression and action-monitoring ERPs (i.e. the error-related negativity and the feedback negativity). Recommendations for the authors of primary studies were provided to will help facilitate more robust, replicable, and informative research in psychophysiology.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ijpsycho.2016.07.001.

References

Aarts, K., Vanderhasselt, M.-A., Otte, G., Baeken, C., Pourtois, G., 2013. Electrical brain imaging reveals the expression and timing of altered error monitoring in major depression. J. Abnorm. Psychol. 122, 939–950.

Alexopoulos, G.S., Murphy, C.F., Gunning-Dixon, F.M., Kalayam, B., Katz, R., Kanellopoulos, D., Etwaroo, G.R., Klimstra, S., Foxe, J.J., 2007. Event-related potentials in an emotional go/no-go task and remission if geriatric depression. Cog. Neurosci. Neuropsychol. 18, 217–221.

Birge, R.T., 1929. Probable values of the general physical constants. Physiol. Rev. Suppl. 1, 1–73.

Birge, R.T., 1932. The calculation of errors by the method of least squares. Phys. Rev. 40, 207–227.

Borenstein, M., Hedges, L., Higgins, J., Rothstein, H., 2009. Introduction to Meta-Analysis. John Wiley & Sons, Ltd., West Sussex, UK.

Bress, J.N., Smith, E., Foti, D., Klein, D.N., Hajcak, G., 2012. Neural response to reward and depressive symptoms in late childhood to early adolescence. Biol. Psychol. 89, 156–162.

Bress, J.N., Foti, D., Kotov, R., Klein, D.N., Hajcak, G., 2013. Blunted neural response to rewards prospectively predicts depression in adolescent girls. Psychophysiology 50, 74–81.

Bress, J.N., Meyer, A., Hajcak, G., 2015. Differentiating anxiety and depression in children and adolescents: evidence from event-related potentials. J. Clin. Child Adolesc. Psychol. 44, 238–249.

Carter, C.S., van Veen, V., 2007. Anterior cingulate cortex and conflict detection: an update of theory and data. Cogn. Affect. Behav. Neurosci. 7, 367–379.

Chang, W.-P., Davies, P.L., Gavin, W.J., 2010. Individual differences in error monitoring in healthy adults: psychological symptoms and antisocial personality characteristics. Eur. J. Neurosci. 32, 1388–1396.

Chiu, P.H., Deldin, P.J., 2007. Neural evidence for enhanced error detection in major depressive disorder. Am. J. Psychiatr. 164, 608–616.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. second ed. Lawrence Erlbaum Associates.

Compton, R.J., Lin, M., Vargas, G., Carp, J., Fineman, S.L., Quandt, L.C., 2008. Error detection and posterior behavior in depressed undergraduates. Emotion 8, 58–67.

Cooper, H.M., 2010. Research Synthesis and Meta-Analysis: A Step-by-Step Approach. Sage Publications, Inc., Thousand Oaks, CA.

Cooper, H., Hedges, L.V., 1994. The Handbook of Research Synthesis. Russell Sage Foundation, New York.

Cumming, G., 2012. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Taylor & Francis Group, LLC., New York, NY.

Dywan, J., Wathewson, K.J., Choma, B.L., Rosenfeld, B., Segalowitz, S.J., 2008. Autonomic and electrophysiological correlates of emotional intensity in older and younger adults. Psychophysiology 45, 389–397.

Eddy, D.M., Hasselblad, V., Shachter, R., 1992. Meta-Analysis by the Confidence Profile Method. The Statistical Synthesis of Evidence. Academic Press, Boston, MA.

Egger, M., Smith, G.D., Schneider, M., Minder, S., 1997. Bias in meta-analysis detected by a simple, graphical test. Br. Med. J. 315, 629–634.

Endrass, T., Riesel, A., Kathmann, N., Buhlmann, U., 2014. Performance monitoring in obsessive-compulsive disorder and social anxiety disorder. J. Abnorm. Psychol. 123, 705–714.

Foti, D., Hajcak, G., 2009. Depression and reduced sensitivity to non-rewards versus rewards: evidence from event-related potentials. Biol. Psychol. 81, 1–8.

Foti, D., Weinberg, A., Dien, J., Hajcak, G., 2011. Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: temporospatial principal components analysis and source localization of the feedback negativity. Hum. Brain Mapp. 32, 2207–2216.

Foti, D., Carlson, J.M., Sauder, C.L., Proudfit, G.H., 2014. Reward dysfunction in major depression: multimodal neuroimaging evidence for refining the melancholic phenotype. NeuroImage 101, 50–58.

Gehring, W.J., Liu, Y., Orr, J.M., Carp, J., 2012. The error-related negativity (ERN/Ne). In: Luck, S.J., Kappenman, E. (Eds.), Oxford Handbook of Event-Related Potential Components. Oxford University Press, New York, NY, pp. 231–291.

Georgiadi, E., Liotti, M., Nixon, N.L., Liddle, P.F., 2011. Electrophysiological evidence for abnormal error monitoring in recurrent major depression. Psychophysiology 48, 1192–1202.

Glass, G.V., 1976. Primary, secondary and meta-analysis of research. Educ. Res. 5, 3–8.

Greden, J.F., 2001. The burden of recurrent depression: causes, consequences, andfuture prospects. J. Clin. Psychiatry 62, 5–9.

Grundler, T.O.J., Cavanagh, J.F., Figueroa, C.M., Frank, M.J., Allen, J.J.B., 2009. Task-related dissociation in ERN amplitude as a function of obsessive-compulsive symptoms. Neuropsychologia 47, 1978–1987.

Haidich, A.B., 2010. Meta-analysis in medical research. Hippokratia 14, 29–37.

Hedges, L., 1981. Distribution theory for Glass's estimator of effect size and related estimators. J. Educ. Stat. 6, 107–128.

Hedges, L., Pigott, T.D., 2001. The power of statistical tests in meta-analysis. Psychol. Methods 6, 203–217.

Hedges, L., Pigott, T.D., 2004. The power of statistical tests for moderators in meta-analysis. Psychol. Methods 9, 426–445.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., Altman, D.G., 2003. Measuring inconsistency in meta-analyses. Br. Med. J. 327, 557–560.

Holmes, A.J., Pizzagalli, D.A., 2008. Spatiotemporal dynamics of error processing dysfunctions in major depressive disorder. Arch. Gen. Psychiatry 65, 179–188.

Holmes, A.J., Pizzagalli, D.A., 2010. Effects of task-relevant incentives on the electrophysiological correlates of error processing in major depression. Cogn. Affect. Behav. Neurosci. 10, 119–128.

Holroyd, C.B., Coles, M.G., 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. 109, 679–709.

Hunter, J.E., Schmidt, F.L., 2004. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. second ed. Sage Publications Inc., Newbury Park, CA.

Ioannidis, J.P.A., Trikalinos, T.A., 2007. An exploratory test for an excess of significant findings. Clin. Trials 4, 245–253.

Kalayam, B., Alexopoulos, G.S., 2003. A preliminary study of left frontal region error negativity and symptom improvement in geriatric depression. Am. J. Psychiatr. 160, 2054–2056.

Ladouceur, C.D., Slifka, J.S., Dahl, R.E., Birmaher, B., Axelson, D.A., Ryan, N.D., 2012. Altered error-related brain activity in youth with major depression. Dev. Cog. Neurosci. 2, 351–362.

Lai, C.-H., 2011. Major depressive disorder: gender differences in symptoms, life quality, and sexual function. J. Clin. Psychopharmacol. 31, 39–44.

Liu, W.-H., Wang, L.-Z., Shang, H.-R., Shen, Y., Li, Z., Cheung, E.F.C., Chan, R.C.K., 2014. The influence of anhedonia on feedback negativity in major depressive disorder. Neuropsychologia 53, 213–220.

Larson, M.J. & Carbine, K.A. (in press). Sample size calculations in human electrophysiology (EEG and ERP) studies: A systematic review and recommendations for increased rigor. Int. J. Psychol.

Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T., Murray, C.J., 2006. Global and regional burden of disease and risk factors, 2001: systemic analysis of population health data. Lancet 367, 1747–1757.

Mies, G.W., van der Veen, F.M., Tulen, J.H.M., Birkenhager, T.K., Hengeveld, M.W., van der Molen, M.W., 2011. Drug-free patients with major depression show an increased electrophysiological response to valid and invalid feedback. Psychol. Med. 41, 2515–2525.

Moran, T.P., 2016. Anxiety and working memory capacity: a meta-analysis and narrative review. Psychol. Bull. (in press).

Moran, T.P., Taylor, D., Moser, J.S., 2012. Sex moderates the relationship between worry and performance monitoring brain activity in undergraduates. Int. J. Psychophysiol. 85, 188–194.

Morris, S.B., DeShon, R.P., 2002. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol. Methods 7, 105–125.

Mueller, E.M., Pechtel, P., Cohen, A.L., Douglas, S.R., Pizzagalli, D.A., 2015. Potentiated processing of negative feedback in depression is attenuated by anhedonia. Depress. Anxiety 32, 296–305.

Olvet, D.M., Hajcak, G., 2008. The error-related negativity (ERN) and psychopathology: toward and endophenotypes. Clin. Psychol. Rev. 28, 1343–1354.

Olvet, D.M., Klein, D.N., Hajcak, G., 2010. Depression symptom severity and error-related brain activity. Psychiatry Res. 179, 30–37.

Padrao, G., Mallorqui, A., Cucurell, D., Marco-Pallares, J., Rodriguez-Fornells, A., 2013. Neurophysiological differences in reward processing in anhedonics. Cogn. Affect. Behav. Neurosci. 13, 102–115.

Peng, L., Xinxin, S., Jing, W., Xiaoran, Z., Jiayi, L., Fengtong, L., Zhonghua, H., Xinxin, Z., Hewei, C., Wenmiao, W., Hong, L., Fengyu, C., Roberson, D., 2015. Reduced sensitivity to neutral feedback versus negative feedback in subjects with mild depression: evidence from event-related potentials study. Brain Cogn. 100, 15–20.

Petticrew, M., 2001. Systematic reviews from astronomy to zoology: myths and misconceptions. Br. Med. J. 322, 98–101.

Proudfit, G.H., 2015. The reward positivity: from basic research on reward to a biomarker for depression. Psychophysiology 52, 449–459.

Ruchsow, M., Herrnberger, B., Wiesend, C., Gron, G., Spitzer, M., Kiefer, M., 2004. The effect of erroneous responses on response monitoring in patients with major depressive disorder: a study with event-related potentials. Psychophysiology 41, 833–840.

Ruchsow, M., Herrnberger, B., Beschoner, P., Gron, G., Spitzer, M., Kiefer, M., 2006. Error processing in major depressive disorder: evidence from event-related potentials. J. Psychiatr. Res. 40, 37–46.

Sambrook, T.D., Goslin, J., 2014. A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. Psychol. Bull. 141, 213–235.

Santesso, D.L., Steele, K.T., Bogdan, R., Holmes, A.J., Deveney, C.M., Meites, T.M., Pizzagalli, D.A., 2008. Enhanced negative feedback responses in remitted depression. Cog. Neurosci. Neuropsychol. 19, 1045–1048.

Schmid, C.H., Mengersen, K., 2013. Bayesian Meta-Analysis. In: Koricheva, J., Gurevitch, J., Mengersen, K. (Eds.), Handbook of Meta-Analysis in Ecology and Evolution. Princeton University Press. Princeton, NJ.

Schoenberg, P.L.A., 2014. The error processing system in major depressive disorder: cortical phenotypal marker hypothesis. Biol. Psychol. 99, 100–114.

Schooler, J., 2011. Unpublished results hide the decline effect. Nature 470, 437.

Schrijvers, D., de Bruijn, E.R.A., Maas, Y., De Grave, C., Sabbe, B.G.C., Hulstijn, W., 2008. Action monitoring in major depressive disorder with psychomotor retardation. Cortex 44, 569–579.

Schrijvers, D., de Bruijn, E.R.A., Maas, Y., Vancoille, P., Hulstijn, W., Sabbe, B.G.C., 2009. Action monitoring and depressive symptom reduction in major depressive disorder. Int. J. Psychophysiol. 71, 218–224.

Smith, T.C., Spiegelhalter, D.J., Thomas, A., 1995. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat. Med. 14, 2699–2865.

Tang, Y., Zhang, X., Simmonite, M., Li, H., Zhang, T., Guo, Q., Li, C., Fang, Y., Xu, Y., Wang, J., 2013. Hyperactivity within an extensive cortical distribution associated with excessive sensitivity in error processing in unmedicated depression: a combined event-related potential and sLORETA study. Int. J. Psychophysiol. 90, 282–289.
Tucker, D., Luu, P., Frishkoff, F., Quiring, J., Poulsen, C., 2003. Frontolimbic response to negative feedback in clinical depression. J. Abnorm. Psychol. 112, 667–678.
van Veen, V., Carter, C.S., 2002. The timing of action-monitoring processes in the anterior cingulate cortex. J. Cogn. Neurosci. 14, 593–602.
Weinberg, A., Shankman, S.A., 2016. Blunted reward processing in remitted melancholic depression. Clin. Psychol. Sci. (in press).
Weinberg, A., Olvet, D.M., Hajcak, G., 2010. Increased error-related brain activity in generalized anxiety. Biol. Psychol. 85, 472–480.
Weinberg, A., Klein, D.N., Hajcak, G., 2012. Increased error-related brain activity distinguishes generalized anxiety disorder with and without comorbid major depressive disorder. J. Abnorm. Psychol. 121, 885–896.
Weinberg, A., Kotov, R., Proudfit, G.H., 2015. Neural indicators of error processing in generalized anxiety disorder, obsessive-compulsive disorder, and major depressive disorder. J. Abnorm. Psychol. 124, 172–185.
Weinberg, A., Meyer, A., Hale-Rude, E., Perlman, G., Kotov, R., Klein, D.N., Hajcak, G., 2016a. Error-related negativity (ERN) and sustained threat: conceptual framework and empirical evaluation in an adolescent sample. Psychophysiology 53, 372–385.
Weinberg, A., Liu, H., Shankman, S.A., 2016b. Blunted neural response to errors as a trait marker or melancholic depression. Biol. Psychol. 113, 100–107.
World Health Organization (Switzerland), 2011. Global Burden of Mental Disorders and the Need for a Comprehensive, Coordinated Response from Health and Social Sectors at the Country Level. Who Press, Geneva.
Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111, 931–959.