Selective Reporting, Analytical Flexibility and Multiple Comparisons in ERP Research: The case of Pain Empathy

Michel-Pierre Coll

Department of Experimental Psychology, University of Oxford, New Radcliffe House, Walton St, Oxford, OX26BW, Correspondence: michel-pierre.coll@psy.ox.ac.uk

Abstract

Empathy has received considerable attention from the field of cognitive and social neuroscience. A significant portion of these studies used the event-related potential (ERP) technique to study the mechanisms of empathy for pain in others in different conditions and clinical populations. These show that specific ERP components measured during the observation of pain in others are modulated by several factors and altered in clinical populations. However, issues present in this literature such as analytical flexibility and lack of type 1 error control raise doubts regarding the validity and reliability of these conclusions. The current study compiled the results and methodological characteristics of 36 studies using ERP to study empathy of pain in others. The results of the meta-analysis suggest that the frontal N2 and the centro-parietal P3 and late positive potential component are sensitive to the observation of pain in others. However, the review of the methodological characteristics shows that the presence of selective reporting, analytical flexibility and lack of type 1 error control compromise the interpretation of these results. These issues and potential solutions to improve future investigations of pain empathy using ERP are discussed.

Keywords: Empathy, Pain, ERP, Methods, Multiple comparisons, Flexibility

1. Introduction

Empathy is a complex psychological construct that refers to the ability of individuals to share the experience of others (Batson, 2009; Coll et al., 2017; Cuff et al., 2016). It is of high importance for healthy interactions with others and has been suggested to be altered in several psychiatric conditions (Bird & Viding, 2014; Decety & Moriguchi, 2007). With the hope that understanding the neuronal mechanisms of empathy will bring new insights on this concept, it has been one of the main endeavours of social neuroscience to describe the cerebral processes and computations underlying empathy (de Vignemont & Singer, 2006; Decety & Jackson, 2004; Klimecki & Singer, 2013).

While it is challenging to elicit empathy in a controlled neuroimaging experiment, studies often use cues of nociceptive stimulation in others (i.e. a needle piercing a hand) to study empathy since they are relatively unambiguous, highly salient and easily understood (Vachon-Presseau et al., 2012). In the electroencephalography (EEG) literature, the event-related potential (ERP) technique has mostly been used to study this phenomenon by measuring electrical brain responses to nociceptive cues depicting various levels of pain in others. In a seminal study, Fan and Han, 2008 showed participants real pictures or cartoon depictions of hands in painful or neutral situations and asked participants to either judge the intensity of the pain experienced or to count the number of hands present in the stimuli. The results showed an early effect of pain in the N1 and N2 component that was not influenced by task demands and a later effect of pain in the P3 component that was modulated by task requirements. The authors interpreted these results as the presence of an early automatic response indexing emotional sharing and a late response indexing the cognitive evaluation of others pain (Fan and Han, 2008). Similar paradigms are now

regularly used to study pain empathy in healthy and clinical samples (see Table S1). Although innovative, the study by Fan and Han, 2008 has several limitations. By analysing multiple time-windows at several scalp locations, the authors to perform over 100 statistical tests on ERP data without adjusting the significance threshold for those multiple comparisons. This suggests that some results have a high probability of being false positives (Kilner, 2013).

The issue of multiple comparisons is a common problem in neuroimaging studies due to the large amount of data collected (Luck & Gaspelin, 2017; Poldrack et al., 2008). In the ERP literature, this is often made worse by the traditional use of factorial analyses performed in several time windows and at several scalp locations without clear hypotheses on the main effects and interactions (Luck & Gaspelin, 2017). Furthermore, this large amount of data also allows for considerable analytical flexibility; that is the idea that the same dataset can be analysed in different ways with significant changes in the results and interpretations depending on the analytical pipeline chosen (Carp, 2012). The presence of flexibility in design and analysis choices and ambiguity regarding how to best to make these choices can lead researchers to compare the results of different analytical pipelines and choose the one which gives the most favourable pattern of result (Carp, 2012; Simmons et al., 2011). When considerable analytical variability is present in a particular field without justification, it can raise doubt regarding the validity of the results and their interpretations.

If the study of pain empathy using ERP is to provide results that are appropriate to further our understanding of empathy in different contexts and populations, it seems imperative to assess 1) the reliability of the effect of the observation of pain in others on the ERP response and 2) the amount of variability in the designs employed to investigate this phenomenon. To reach these aims, a review of the methodological practices used in 36 ERP studies investigating pain empathy and a meta-analytical

compilation of their results was performed. The results provide meta-analytical evidence for the association between some ERP components and the observation of pain in others. However, there was considerable variation in the design and analyses and incomplete reporting of results, raising doubts on the validity of these results.

2. Methods

2.1. Study selection and data extraction

The articles included in this review were selected by consulting online databases (*PubMED*, *Web of Science* and *Google Scholar*) for studies that were available online before June 15, 2017 using different combinations of keywords (e.g. EEG, ERP, Pain, Empathy, Vicarious). The reference lists and citations reports of eligible studies were also consulted. To be included in this report, studies had to report scalp ERP data in response to pictures depicting nociceptive stimulations (e.g. Jackson et al. (2005)). Studies using facial expressions stimuli were included only when nociceptive stimulations were visible in the stimuli (e.g. needle piercing the skin of the face). This procedure led to the selection of 36 studies published between 2008 and 2017 in 20 different journals (see Table S1 and asterisks in reference list). All data reported in this review were manually extracted from the text of the published article or accompanying supplementary materials and available in Table S1 (osf.io/h6n4v).

2.2. Methodological review

Several variables concerning the Participants, Materials and procedures, Data collection and preprocessing, ERP measurements, Statistical analyses and the Reporting of results were collected and summarized. When information was not clearly reported, the value was estimated based on the available information or assumed to have a particular value (e.g. when no post-hoc correction was reported, it was

assumed that none were used). When insufficient information was available for a particular variable, it was marked as not reported and treated as a missing value.

In order to assess the exhaustiveness of the hypotheses formulated regarding ERPs in each study, hypotheses were classified in one of four categories: *Complete*, *Partial*, *Alternative* and *None*. Hypotheses were rated as *Complete* if they clearly predicted the specific components that were expected to be influenced by all manipulations as well the direction and the location of this effect. If some predictions were present but were incomplete or unclear, the hypothesis was rated as *Partial*. Hypotheses that were formulated as two alternative outcomes without a clear prediction were labeled *Alternative* and the absence of prediction regarding ERP effects was labeled *None*. This procedure was applied separately for the factorial analysis of variance performed on ERP components and for the correlational analysis of ERP components with other variables.

2.3. Meta-analysis of ERP components

A quantitative meta-analysis was carried out to assess the evidence for a modulation of different ERP components by the observation of pain. However, this was complicated by the fact that most studies reviewed reported only significant results (see Section 3.5 Results reporting) and the general lack of clarity and precision of the results section of many studies. Nevertheless, when possible, F-values for the omnibus repeated measures test comparing the ERP response to pain and neutral stimuli were collected for each study.

When several between and within-subject factors were manipulated, the F-value of the baseline condition was selected when available. (e.g. in healthy controls or following neutral priming). Similarly, when available, the F values for individual electrodes were collected. However, in most cases, the omnibus F value from the

main effect of pain in a multi-factorial analysis was collected and attributed to all electrodes included in the analysis. When t-values were reported, the square of the t-value was taken as the F-value (Brozek & Alexander, 1950). When mean amplitudes and standard deviations or standard errors were reported, the paired sample t-value for the pain effect was calculated assuming a correlation of 0.70 between measurements and converted to an F-value. When only the exact p-value was reported, the corresponding F value was found using the F distribution. When only the p-value threshold was reported (i.e. p ¡ 0.05), the value was not considered in the quantitative meta-analysis. This was done for each component of the most frequently analysed components (N1, P2, N2, P3, LPP, see Section 3.4). When a component was described in the analysis section but not reported in the results, it was assumed that the effect of pain was not significant. When multiple studies were published using the same dataset (Ikezawa et al., 2014; Wexler et al., 2014; Corbera et al., 2014; Fan and Han, 2008; Han et al., 2008), studies were included only if they analysed different components and/or time windows. From the 36 studies reviewed, 1 was fully excluded from the meta-analysis because it used the same dataset and analyses as another study (Han et al., 2008), 1 because it did not report sufficient information (Ikezawa et al., 2012) and 1 because it used non-parametric statistics (Fitzgibbon et al., 2012). Thirty-three studies were therefore included in the meta-analysis (marked with double asterisks in the References). All 36 studies were included in the review of methodological characteristics.

In order to compare the effect sizes across studies, the F-values collected for each electrode and component were converted to partial eta-squared (Bakeman, 2005; Lakens, 2013). The 90% confidence interval around the partial eta-squared for the effect of pain observation was calculated using the *conf.limits.ncf* function from the *MBESS* R package (Kelley, 2017) The proportion of studies reporting a significant

effect of pain for each component was also calculated.

3. Results

3.1. Goals and hypotheses

All studies reviewed aimed at comparing the effect of an experimental manipulation and/or participant characteristics on the ERP to pain stimuli with the goal of furthering the understanding of the mechanisms underlying empathy in the general population and/or in various clinical groups. As shown in Table 1, the majority of studies presented incomplete hypotheses regarding the analysis of ERP components. While some studies provided complete hypothesis for the factorial analysis of ERPs, this was not the case for correlational analyses of ERP and other behavioural or physiological variables, suggesting that most of these analyses were exploratory in nature.

Table 1: Ratings of the the exhaustiveness of hypotheses for the factorial and correlational analyses of ERP data.

Judgement of hypothesis	Factorial analyses on ERP data (% of 36 studies)	Correlations with ERP data (% of 22 studies)
None	11.11	68.18
Partial	50.00	27.27
Complete	22.22	0
Alternative	16.67	4.55

3.2. Participants

In the 36 studies reviewed, 52.78% used a between-subject design and 47.22% used a within-subject design. Among studies employing a between-subjects design,

52.63% compared participants from the general population to participants from a clinical group. These clinical conditions included autism, amputation, bipolar disorder, cyber-addiction, fibromyalgia, juvenile delinquents and schizophrenia. No studies presented an a priori power analysis and only one study provided a justification of sample size based on the expected effect size. The average sample size, sample size per group and participants excluded are shown in Figure 1. In order to asses the power of each study to detect a small (d = 0.2), medium (d = 0.5) or large (d = 0.8) effect size (Cohen, 1992), a power analysis was performed for each study and each of these effect sizes using the sample size per group, a two-sided test and a significance threshold of 0.05. As shown in Figure 1, most studies were only adequately powered to detect a large effect size equal to or higher than d = 0.8. No studies had 80% power to detect a small or medium effect size.

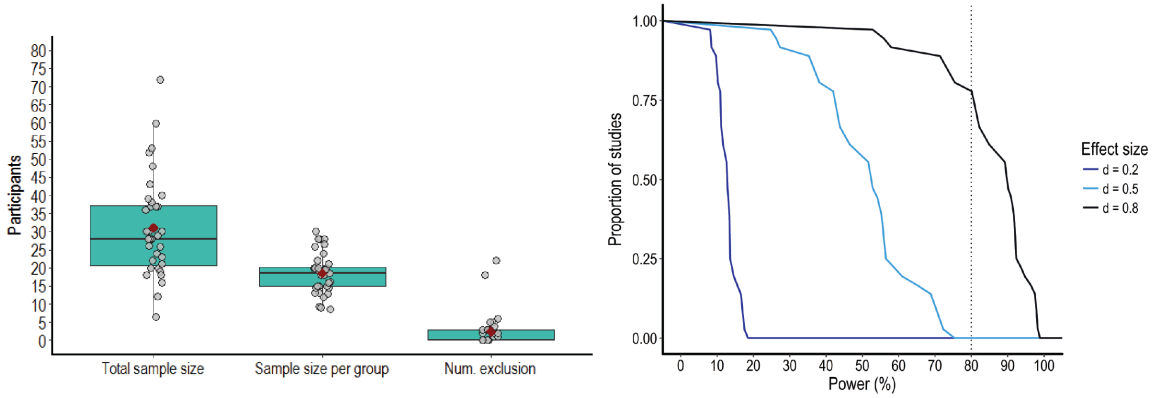

Figure 1: **Left:** Total sample size, sample size per group and number of exclusion. For all box plots in this paper, the boxes show the 25th and 75th percentile, the error bars show 1.5 interquartile interval, the horizontal line shows the median, the red square shows the mean and the gray points show the individual studies. **Right:** Proportion of studies as a function of the level of power to detect a small (d = 0.2), medium (d = 0.5) and large (d = 0.8) effect size.

3.3. Materials and procedure

3.3.1. Visual stimuli

The majority of studies reviewed (83.33%) used static pictures depicting limbs (generally hands or feet) in painful and nonpainful situations similar to those used initially by Jackson et al. (2005). The other types of stimuli included short 3-frames clips of limbs in painful situations (8.33%), static pictures of faces pricked by a needle or touched with a cotton bud (5.56%) and anthropomorphised objects pricked by a needle or touched with a cotton bud (2.78%). The average stimulus duration is shown in Figure 2.

3.3.2. Experimental task

All studies compared ERPs in response to painful and non-painful stimuli. Including this Pain factor, studies had on average 2.06 within-subjects factors (SD = 0.53, range: 1-3) and an average of 4.5 within-subjects conditions (SD = 1.81, range: 2-8). During the ERP measurements, participants were either asked to detect the presence of pain in a forced choice format (58.33% of studies), to assess the intensity of the pain observed using a rating scale (19.44%), to passively observe the pictures (16.67%) or to perform another cognitive task (5.56%). The average number of trials per condition is shown in Figure 2.

3.3.3. Recordings

EEG was collected from 60-64 scalp electrodes in the majority of cases (63.89 %) while the remaining studies used 32 (27.78%), 72 (2.78%), 128 (5.56%) scalp electrodes. 30.56 % of studies did not report the manufacturer of the EEG system used. The majority of studies (30.56 %) used an EEG system manufactured by *Brain Products*. Other manufacturers included *Biosemi* (25%), *EGI Geodesic* (5.56%),

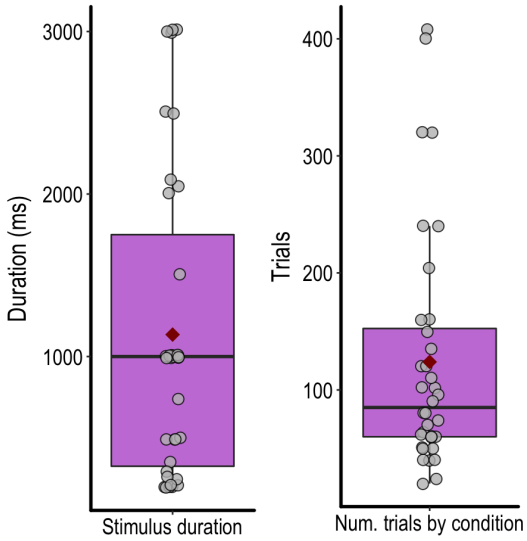

Figure 2: **Stimulus duration and number of trials per condition.**

Neuroscan (5.56%) and *NuAmps* (2.78%).

3.3.4. Preprocessing

EEG data were high-pass filtered in most cases with a cutoff value of 0.1 Hz (52.78%) and the rest of the studies used a value between 0.01 and 1 Hz. For low-pass filters, the most used cutoffs were 100 Hz and 30 Hz used in 33.33% of cases each. Other low pass filters cutoffs included values between 40 and 80 Hz. 11.11% and 6.6% of studies did not report using a high-pass filter or a low-pass filter respectively. 8.33% of studies reported using a notch filter to filter out electrical noise.

The average of the mastoid processes and the average of all scalp electrodes were the most popular reference schemes for EEG analyses (38.89% each). Other studies use the average of the earlobes (5.56%), a single mastoid (13.89%) or did not report the reference used for analysis (2.78%). EEG data were epoched for analyses and in all cases the average pre-stimulus baseline was subtracted from the post-stimulus

epoch. The duration of this baseline was on average 193.06 ms (SD =29.65, range = 100-250). The average post-stimulus epoch duration is shown in Figure 3

All studies reported using at least one method to remove or correct for artifacts. Artifact rejection procedures included rejecting epochs by visual inspection or using a fixed amplitude threshold. Artifact correction procedures included removing components after independent component analysis (ICA) or using various algorithms to remove EOG activity from the data. Some studies reported using additional filters to remove artifacts without providing further details. The percentage of studies using each of the main procedures is shown in Figure 3. Automatic rejection using a fixed threshold was the most used method and the average rejection threshold is shown in Figure 3. When using an artifact rejection procedure, 44.44% of studies reported the average number of epochs removed. On average, 11.93% of trials were removed (SD = 6.41, range = 4.58-29

3.4. ERP analyses

3.4.1. ERP selection and measurement

The average number of components analysed is shown in Figure 4A. In most cases, the choice of ERP components to analyse was justified using previous studies (55.56%) while other studies chose to analyse all grand average peaks (25%) or did not justify their selection of components (11.11%). As shown in Figure 4B, the most widely analysed ERP components were the N1, N2, P2, N3, P3 and the LPP. Note that in some cases, slightly different names were used for these components (e.g. P320 instead of P3). Furthermore, some studies also performed several analyses on the same component (e.g. early and late LPP). See Table S1 for the names and all components analysed in each study.

The average number of locations analysed is shown in Figure 4A and the percent-

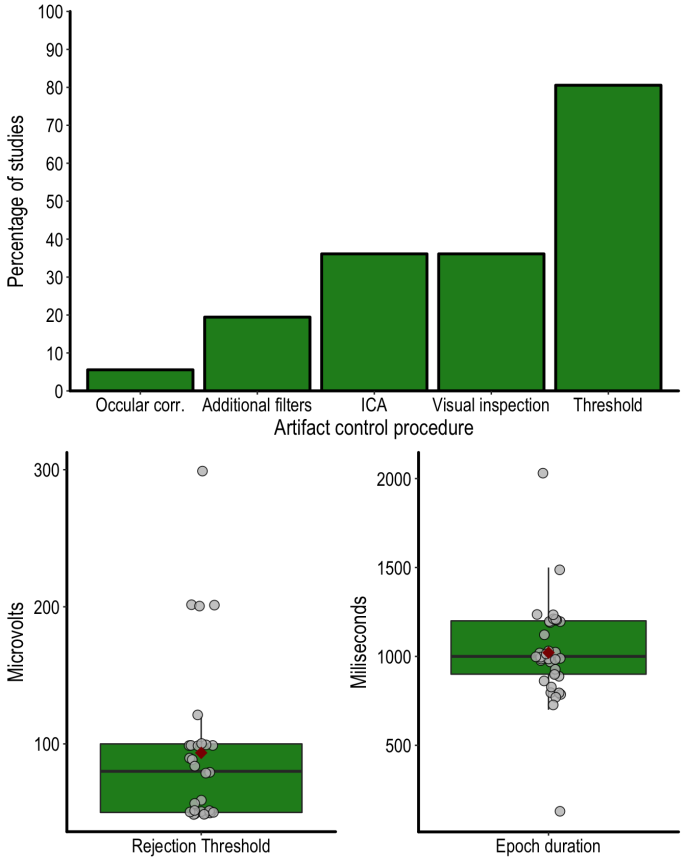

Figure 3: **Percentage of studies reviewed using each type of artifact control procedure identified, value of the rejection threshold if used and duration of epochs..**

age of studies analysing each scalp location is shown on a 64 electrode montage in Figure 4C. A minority of studies provided a justification for the choice of locations to analyse (36.11%) and in most cases this choice was based on previous studies (25%).

Almost all studies quantified the ERP components using the mean amplitude within a time window (80.56%) while other studies used the peak amplitude within a time window (8.33%) or point by point analyses (i.e., performing analyses in small time windows covering the whole ERP epoch; 8.33%). One study used both peak

amplitude and mean amplitude depending on the component (2.78%). The choice of the time window to analyse was either based on visual inspection (50%), previous studies (19.44%), the data itself (i.e circular analyses, 16.67%) or not justified (13.89%). The percentage of studies analysing each time point in the post-stimulus window for each component is shown in Figure 4D.

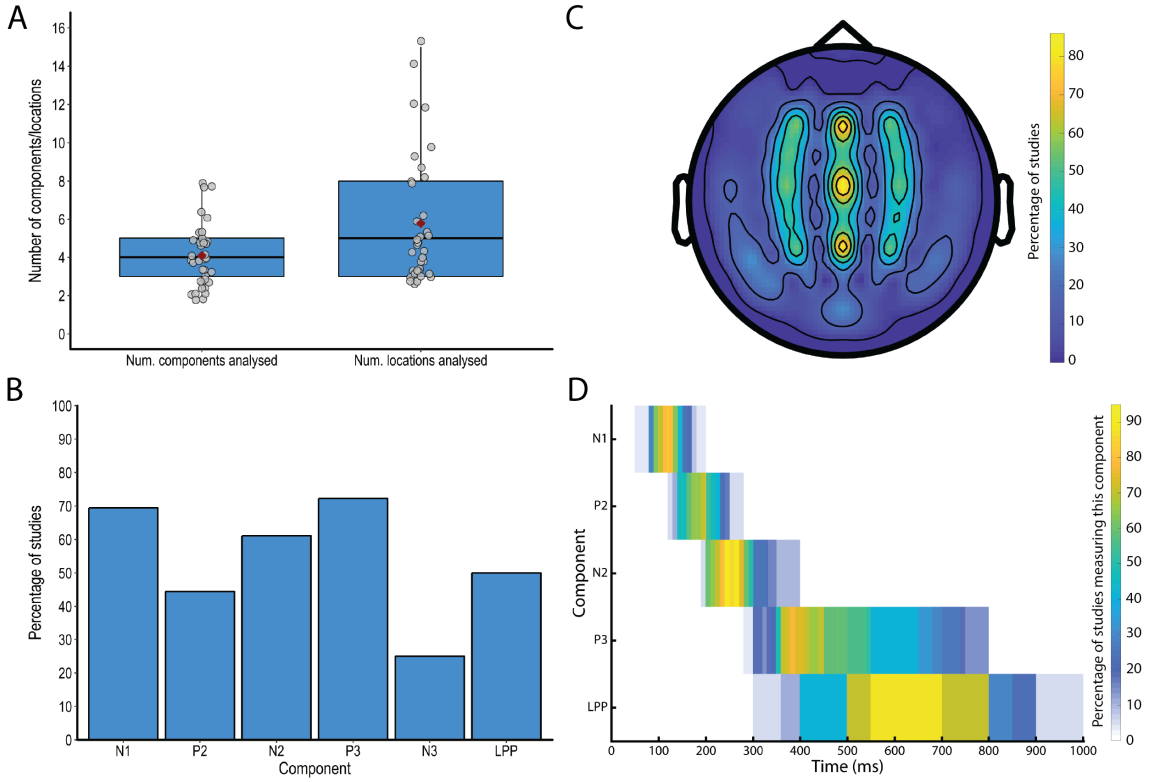

Figure 4: **(A)** Number of components and locations analyzed in each study. **(B)** Percentage of studies analyzing each of the main component. **(C)** Percentage of studies analyzing each scalp location. **(D)** Percentage of studies analyzing each component analyzing each time-point in the post-stimulus window.

3.4.2. ERP statistical analyses

Almost all studies used factorial analyses of variance (ANOVA) to assess the statistical significance of the experimental factors on the ERPs (94.44%). One study used the analysis of covariance (ANCOVA) and another used the non-parametric Kruskal-Wallis test. Since the few studies using point-by-point analyses (8.33%) performed a large number of ANOVAs compared to the rest of the studies (on average 130 ANOVAs), these studies were not considered in the following description of the factorial analyses. Another study performing 54 ANOVAs was also considered an outlier and was not included in the descriptive statistics. The average number of ANOVAs performed in the remaining studies (N = 31) is shown if Figure 5. These ANOVAs had on average 3.16 factors (SD = 0.73, range = 2-5) and 18.39 cells (SD = 14.76, range = 4-60). To assess how many statistical tests these ANOVAs represent, the number of main effects and potential interactions was multiplied by the number of ANOVAs (see Luck & Gaspelin 2017). The total number of tests in ANOVAs is shown in Figure 5. No studies corrected the significance threshold for the total number of ANOVAs performed. However, some studies (55.45%) corrected the significance threshold when performing post-hoc comparisons using the Bonferroni (38.89%), Tukey (11.11%) or Scheffe correction (5.56%). Several studies used the Greenhouse-Geisser correction when performing repeated-measure analyses (52.78%).

61.10% of studies performed correlations between ERP data and other variables in addition to the factorial analyses. These correlations either used the difference between the ERP amplitude in two conditions (54.55%) or the mean amplitude in a particular condition (45.45%) and were often performed in multiple time windows and scalp locations. The average number of correlations per study is shown in Figure 5. No studies corrected the significance threshold to control for the possibility of a

type 1 error in correlation analyses.

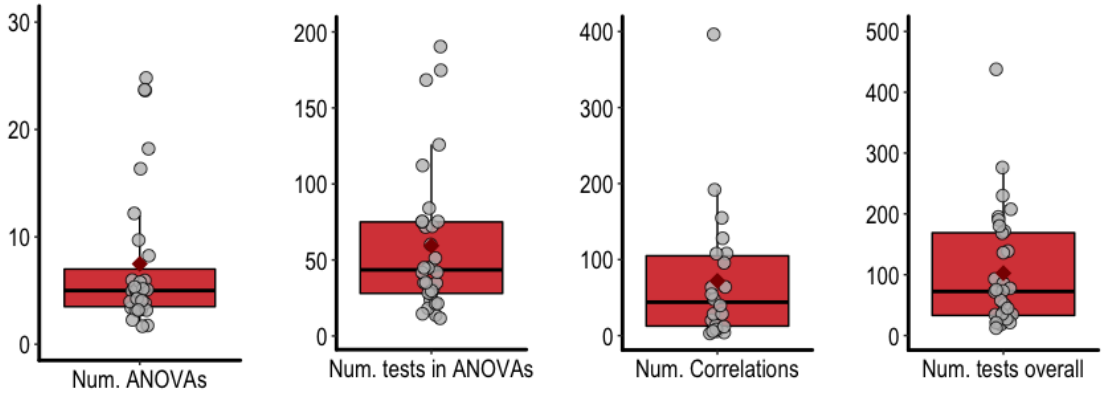

Figure 5: **Number of ANOVAs performed, total number of statistical tests performed using ANOVAs, number of correlations and overall number of statistical tests performed.**

3.5. Results reporting

88.89% of studies reported only significant results. When reporting the results form factorial analyses, most studies did not report any estimate of effect size (55.56%) while the rest reported the partial eta-squared (41.67%) or Cohen's d (5.56%).

All but one study plotted the ERP data. 36.11% plotted only the time-course of the ERP response while 61.11% plotted both the time course and scalp maps at particular time points. In most cases the locations and time points plotted were chosen because they were thought to be representative of the results (55.56%) while other studies plotted all locations analyzed (25.00%) or only the locations showing significant effects (16.67%). On the time course plots, only one study plotted the

error intervals. In addition to the time course and scalp maps, 22.22% of studies also reported the ERP amplitude in a table and 36.11% in bar graphs.

3.6. Meta-analysis of the pain effect on the main ERP components

The results from the meta-analysis of the effect of the observation of pain stimulation on the ERP components that were the most frequently analysed are shown in Figure 6, 7, 8 and 9. Overall, the results from this meta-analysis show that the centro-parietal P3 component seems to be the component that is the most sensitive to the observation of nociceptive stimulation in others and the most influenced by experimental manipulations. While the N2 and LPP component also seem to be sensitive to the effect of pain observation at frontal and centro-parietal sites respectively, these components were rarely found to be modulated by other experimental factors. Finally, it is unclear from these results if the N1 and P2 components are indeed modulated by the observation of pain in others.

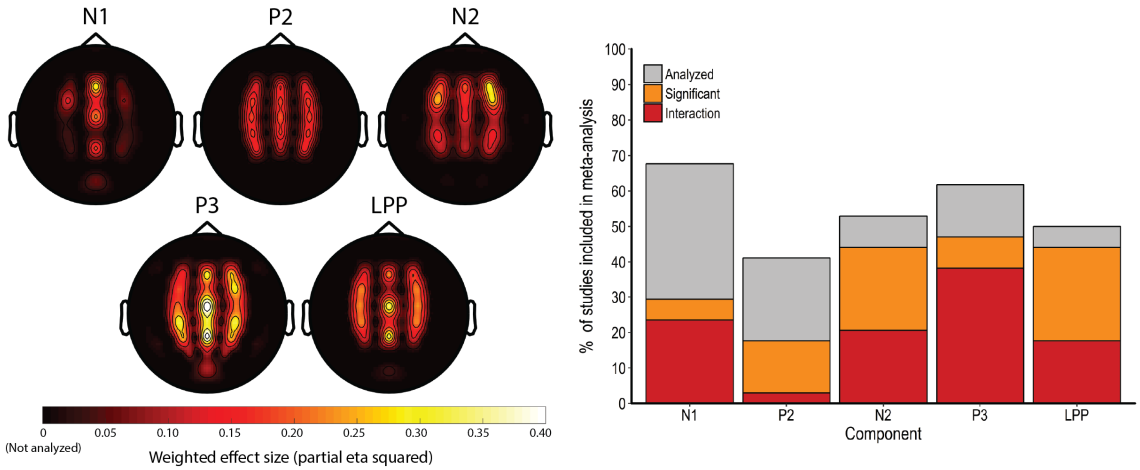

Figure 6: **Left:** Scalp map of the average partial eta-squared weighted by the number of studies reporting results at this electrode for the main effect of pain observation on each electrode and component. **Right:** Bar graph showing the proportion of studies included in the meta-analysis analyzing each component, reporting a significant main effect of pain and a significant interaction between this effect and another factor.

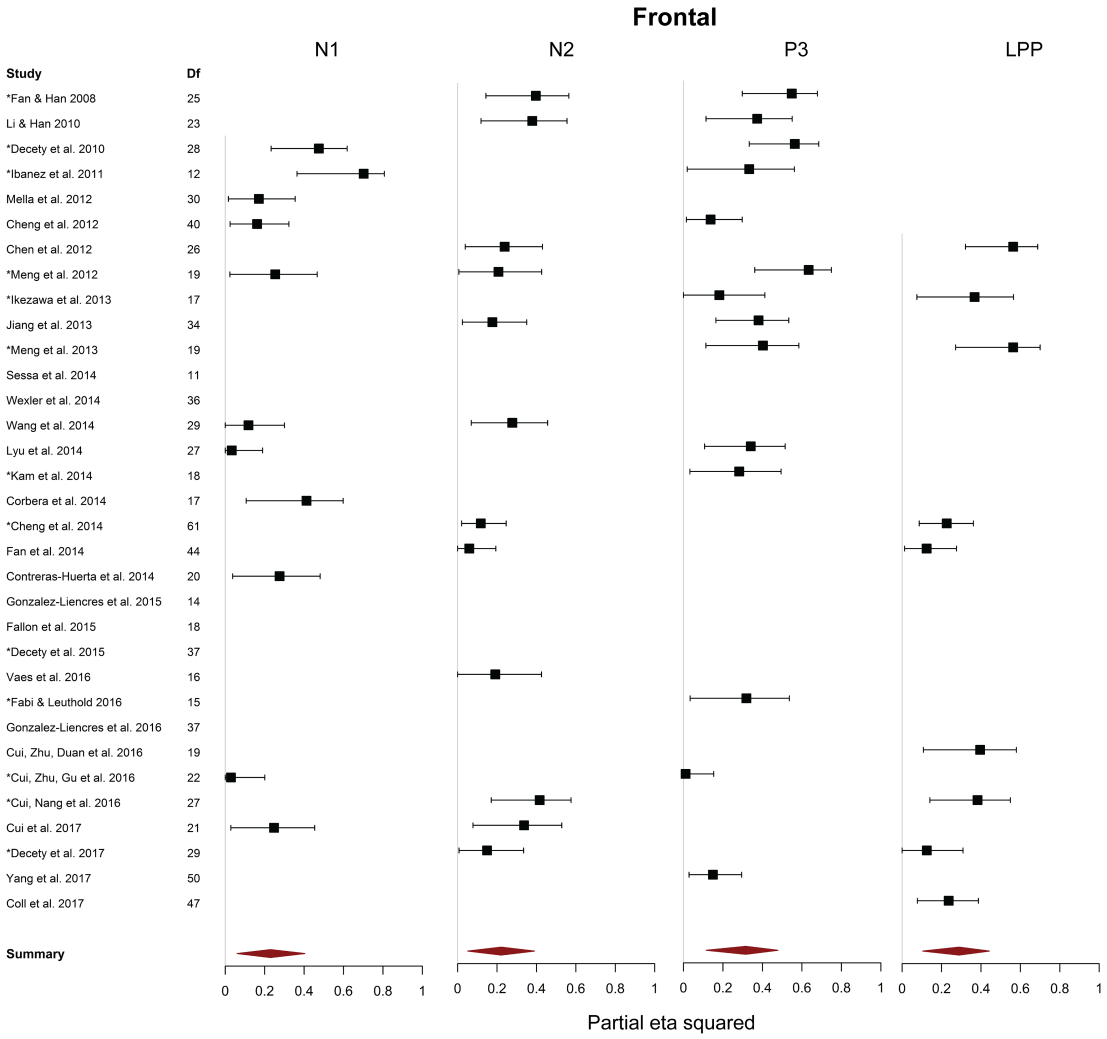

Figure 7: Forest plot showing the effect size and 90% confidence interval for the effect of size of pain observation on the ERP amplitude at the frontal electrodes fro the N1, N2, P3 and LPP components. **Studies for which electrodes form other regions were also included in the effect size for at least one other location.*

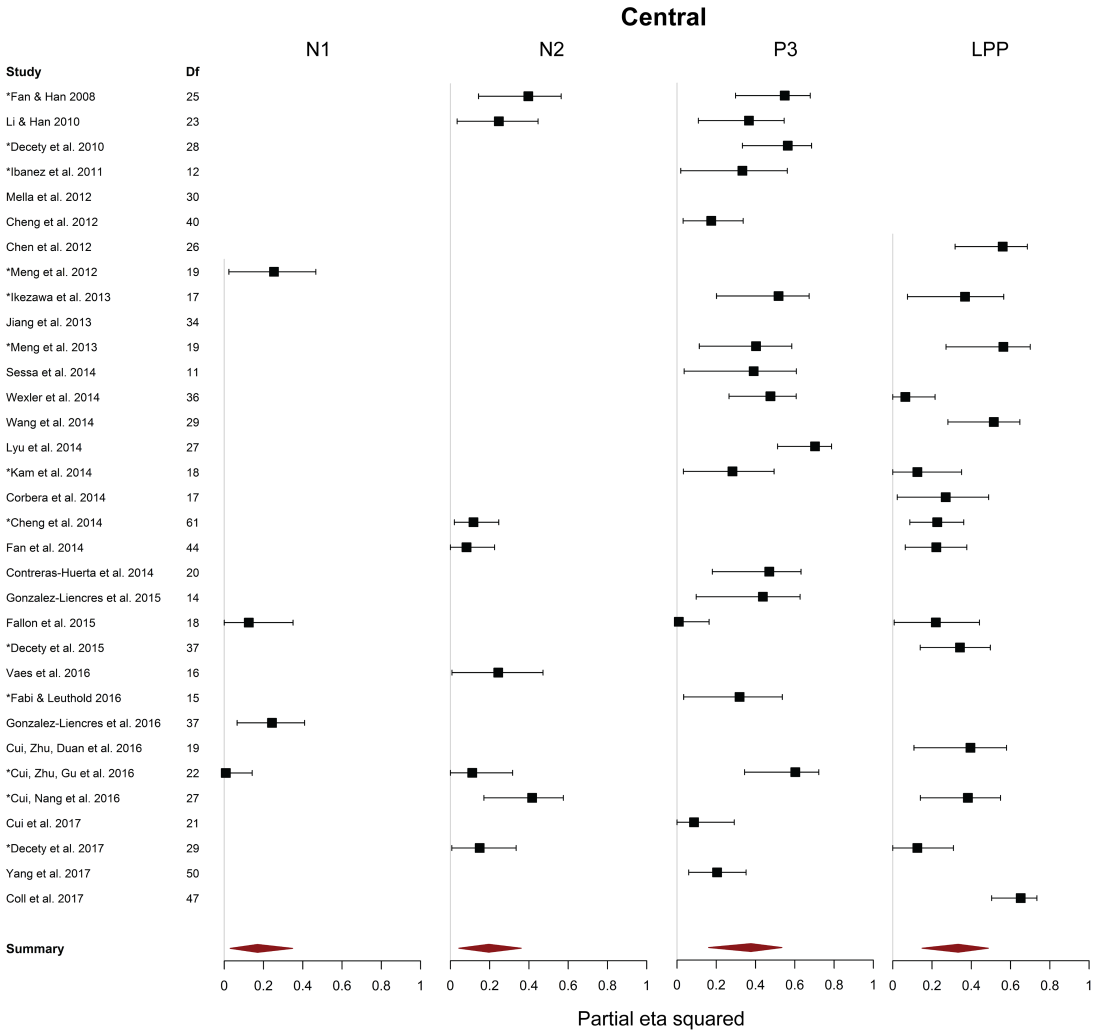

Figure 8: Forest plot showing the effect size and 90% confidence interval for the effect of size of pain observation on the ERP amplitude at the central electrodes fro the N1, N2, P3 and LPP components. **Studies for which electrodes form other regions were also included in the effect size for at least one other location.*

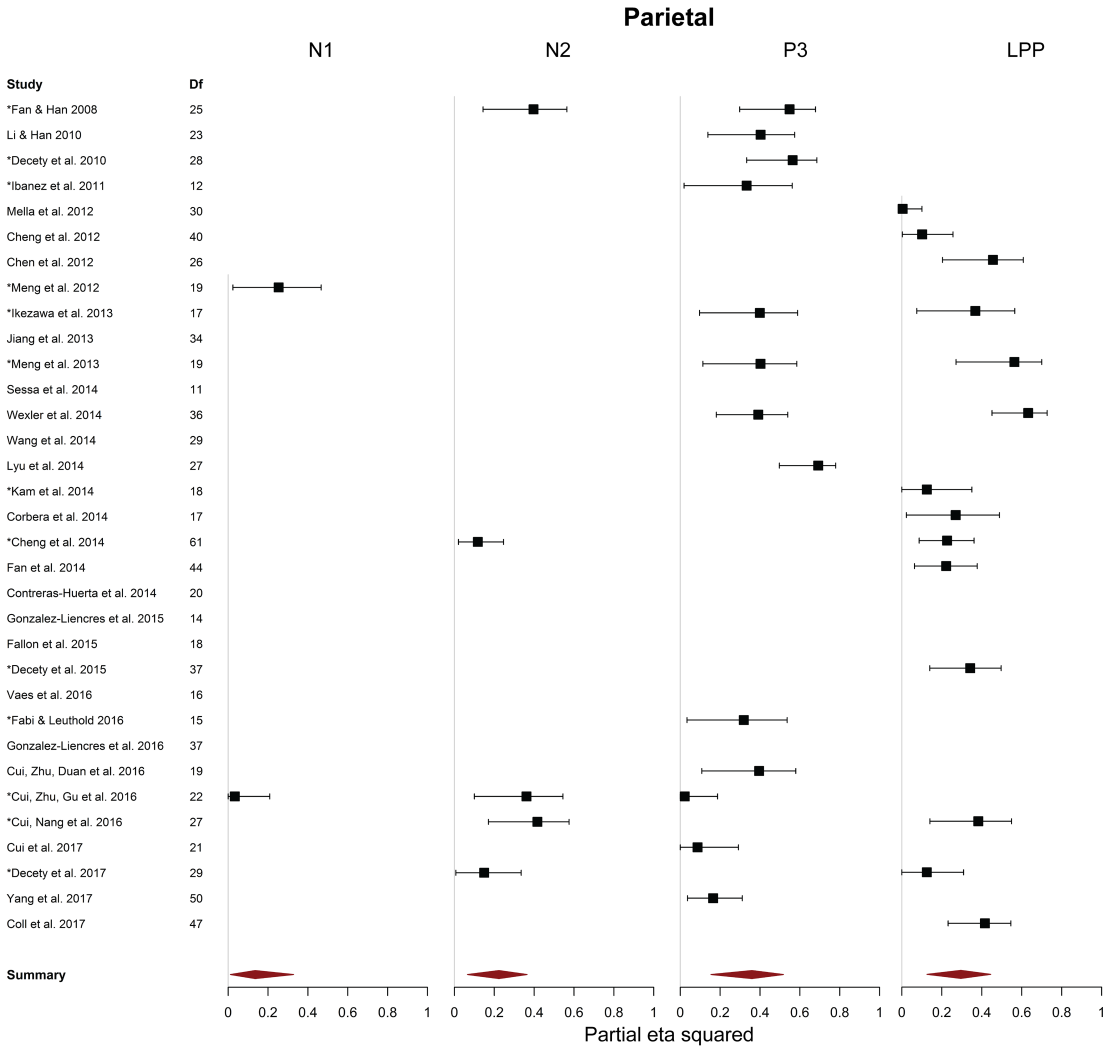

Figure 9: Forest plot showing the effect size and 90% confidence interval for the effect of size of pain observation on the ERP amplitude at the parietal electrodes fro the N1, N2, P3 and LPP components. **Studies for which electrodes form other regions were also included in the effect size for at least one other location.*

4. Discussion

4.1. Meta-analysis of the pain effect on ERP components suggest that some effects seem reliable but is undermined by methodological issues

The quantitative meta-analysis performed suggests that the largest effect of the observation of pain in others is found for the frontal N2 and centro-parietal P3 and LPP components. This is in line with the initial experiment by (Fan and Han, 2008) suggesting that the empathic response to pain in others is reflected in both early and late components of the ERP response. However, the review of the methodological characteristics of the studies included in the meta-analysis highlights several important issues that undermine the confidence that should be placed in these results and interfere with a meaningful meta-analytical compilation of this research. These issues and potential solutions will be discussed here.

4.2. Selective and incomplete reporting of methods and results undermines incremental science

A first striking observation permitted by this review is the lack of comprehensive reporting for the results from the statistical analysis of ERP components. Indeed, it was found that the vast majority of studies only reported significant results form a large number of factorial analyses. Therefore, in addition to a potential publication bias (Rosenthal, 1979), the ERP studies reviewed here also show a within-study reporting bias according to which analyses leading to non-significant results are less likely to be reported. Putting aside the fact that negative results can sometimes be informative if statistical power is high enough (Greenwald, 1975), the main consequence of this practice is that any attempt to meta-analytically summarize the results of such studies will be difficult, inevitably biased and of questionable useful-

ness. Therefore, it should be noted that the effect size calculated in the meta-analysis performed in this report are necessarily inflated.

In several cases, this practice was justified by the necessity to provide a concise and brief report of the results. The short term solution to this issue is to make it mandatory for authors to appropriately report all the results of all statistical analyses performed on ERP components in the text or in supplementary materials. It would also be beneficial for the field of ERP research to adopt a standard reporting procedure that would enable the automatic extraction of results from published articles and facilitate meta-analyses and large-scale automated summaries of all published studies. For example, the field of functional magnetic resonance imaging research has taken advantage of the standard reporting of activation coordinates in tables to produce automated meta-analytical tools (Yarkoni et al., 2011). A more preferable long term solution would be to favorise the sharing of ERP data in online repository which would allow the re-analysis and meta-analysis of large datasets and a quick and efficient assessment of the evidence for specific effects (Poldrack & Gorgolewski, 2014).

A related issue to the incomplete reporting of results was incomplete and unclear reporting of several important methodological details. For example, less than half of the studies reported the number of trials left after artifact rejection procedures, meaning that it was unknown how many trials were included in the analyses in a majority of cases. To avoid omitting to report such crucial information, researchers should refer to published guidelines for reporting of EEG experiments (Keil et al., 2014) and reviewers should enforce these guidelines in all relevant cases.

4.3. Analytical flexibility, multiple comparisons and large amount of data will always produce significant results

The quantification of the analytical approaches used in the 36 ERP studies reviewed showed a relatively standard pre-processing pipeline including in most cases filtering, re-referencing, segmentation, baseline correction and artifact removal. The specific preprocessing parameters varied to some extent but this could be expected given the different EEG systems and experimental design used.

There was, however, considerable variability in the quantification and statistical analysis of the ERP data. Indeed, while most studies used the mean amplitude between fixed-latencies to quantify ERPs, the time-windows and the electrode locations used for this measure often varied considerably across studies without a clear rationale underlying this choice. The problems of analytical flexibility, sometimes called researchers degrees of freedom, have already been discussed elsewhere (Luck & Gaspelin, 2017; Simmons et al., 2011). However, the current study suggests that previous investigations of pain empathy using ERPs might be compromised by this practice. It is worth noting that the components which showed the highest effect size in the meta-analysis were the components for which there was the greatest flexibility in the selection of the time window.

This analytical flexibility was often combined with a shotgun analytical approach in which a high number of statistical analyses were performed on several locations and time-windows and any significant effect was interpreted as meaningful. On average, the studies reviewed here performed 100 statistical tests on ERP data and none corrected the significance threshold to reduce the risk of a false positive finding. This is problematic since such a high number of statistical tests leads to a high probability that several significant results are in fact false positives (Kilner, 2013; Luck & Gaspelin, 2017). To diminish this risk, researchers can use analytical techniques that

can take into account the spatial and temporal distribution of ERP data to reduce the number of comparisons such as mass-univariate analyses (Groppe et al., 2011; Pernet et al., 2015) with adequate type 1 error control.

Alternatively, researchers can restrict their analyses to scalp regions and time windows for which an effect was predicted. However, a worrying result of this study is the fact that in several cases the analysis and interpretation of the data were not constrained by the researchers' predictions since no clear hypothesis were formulated. This suggests that despite using the confirmatory analytical approach of null hypothesis significance testing, most of ERP research in the field of pain empathy is exploratory (Wagenmakers et al., 2012). A solution to this issue would be to require researchers to clearly formulate their research hypotheses and label as exploratory the results of analyses that were not predicted. Ideally, researchers could pre-register their hypotheses online or publish using the pre-registered report format to establish the analysis plan before collecting data (Munafo et al., 2017).

4.4. Caveats

Due to the issues discussed above, the results from the meta-analysis are most likely biased. This means that while they can be used to roughly guide future investigations, the precise value of the effect sizes should not be taken at face value. Furthermore, the use partial eta-squared as a measure of effect size limits the capacity to compare effect sizes across studies since the inclusion of other factors in the design influences this measure (Olejnik & Algina, 2003). Due the paucity of the information available in the results section of the studies reviewed, this measure was the most appropriate in this case but future studies should report generalized measures of effect size such as the generalized eta-squared (Olejnik & Algina, 2003)

The current report did not assess the construct validity of the use of ERPs in the

context of pain empathy. Future reviews should assess to what extent the specific ERP components are indeed valid and specific measures of empathic responses and the extent to which other processes such as attention or arousal could explain these effects.

The methodological issues highlighted in this report are not specific to the domain of pain empathy and the present observations could potentially be generalised to many other fields of research in cognitive neuroscience (e.g. Hobson & Bishop 2017, for similar observations on EEG studies of action observation). Furthermore, while several issues were found to be prevalent in the studies reviewed, the scientific quality and usefulness of all papers cited in this report should be assessed on an individual basis. Finally, the solutions proposed to the issues raised are not exhaustive, nor can they be applied indiscriminately to all ERP research.

5. Conclusion

In conclusion, this study suggests that the investigation of the empathic response to pain in others using ERPs is undermined by several methodological problems that raise serious doubts regarding the reliability, validity and overall usefulness of this research. Researchers in the field should take into account the methodological issues raised here when designing and reviewing ERP experiments. This is of critical importance if this paradigm is to be used to draw conclusions on the socio-emotional functioning of different clinical populations.

Acknowledgements

MP Coll is funded by a fellowship from the Canadian Institutes of Health Research and was funded by a fellowship from the *Fonds de Recherche Quebec - Sante* during this work.

Conflicts of interest

The author declares no conflicts of interest.

References

** : Reference included in the methodological review*

***: Reference included in both the meta-analysis and the methodological review*

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behav Res Methods*, *37*(3), 379–384. DOI: 10.3758/BF03192707.

Batson, C. D. (2009). These things called empathy: eight related but distinct phenomena. In J. Decety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 3–16). The MIT Press. DOI: 10.7551/mitpress/9780262012973.003.0002.

Bird, G. & Viding, E. (2014). The self to other model of empathy: providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neurosci Biobehav Rev*, *47*, 520–532. DOI: 10.1016/j.neubiorev.2014.09.021.

Brozek, J. & Alexander, H. (1950). The formula t2 = f. *Am J Psychol*, *63*(2), 262–269. DOI: 10.2307/1418933.

Carp, J. (2012). The secret lives of experiments: methods reporting in the fmri literature. *Neuroimage*, *63*(1), 289–300. DOI: 10.1016/j.neuroimage.2012.07.004.

**Chen, C., Yang, C.-Y., & Cheng, Y. (2012). Sensorimotor resonance is an outcome but not a platform to anticipating harm to others. *Soc Neurosci*, *7*(6), 578–590. DOI: 10.1080/17470919.2012.686924.

**Cheng, Y., Chen, C., & Decety, J. (2014). An eeg/erp investigation of the development of empathy in early and middle childhood. *Dev Cogn Neurosci*, *10*, 160–169. DOI: 10.1016/j.dcn.2014.08.012.

**Cheng, Y., Hung, A.-Y., & Decety, J. (2012). Dissociation between affective sharing and emotion understanding in juvenile psychopaths. *Dev Psychopathol*, *24*(2), 623–636. DOI: 10.1017/S095457941200020X.

Cohen, J. (1992). A power primer. *Psychol Bull*, *112*(1), 155–159. DOI: 10.1037/0033-2909.112.1.155.

Coll, M.-P., Viding, E., Rtgen, M., Silani, G., Lamm, C., Catmur, C., & Bird, G. (2017). Are we really measuring empathy? proposal for a new measurement framework. *Neurosci Biobehav Rev*. DOI: 10.1016/j.neubiorev.2017.10.009.

**Contreras-Huerta, L. S., Hielscher, E., Sherwell, C. S., Rens, N., & Cunnington, R. (2014). Intergroup relationships do not reduce racial bias in empathic neural responses to pain. *Neuropsychologia*, *64*, 263–270. DOI: 10.1016/j.neuropsychologia.2014.09.045.

**Corbera, S., Ikezawa, S., Bell, M. D., & Wexler, B. E. (2014). Physiological evidence of a deficit to enhance the empathic response in schizophrenia. *Eur Psychiatry*, *29*(8), 463–472. DOI:10.1016/j.eurpsy.2014.01.005.

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion review: journal of the International Society for Research on Emotion*, *8*(2), 144–153. DOI: 10.1177/1754073914558466.

**Cui, F., Ma, N., & Luo, Y.-J. (2016). Moral judgment modulates neural responses to the perception of other's pain: an erp study. *Sci Rep*, *6*(1), 20851. DOI: 10.1038/srep20851.

**Cui, F., Zhu, X., Duan, F., & Luo, Y. (2016). Instructions of cooperation and

competition influence the neural responses to others' pain: An erp study. *Soc Neurosci*, *11*(3), 289–296. DOI: 10.1080/17470919.2015.1078258.

**Cui, F., Zhu, X., Gu, R., & Luo, Y.-J. (2016). When your pain signifies my gain: neural activity while evaluating outcomes based on another person's pain. *Sci Rep*, *6*(0), 26426. DOI: 10.1038/srep26426.

de Vignemont, F. & Singer, T. (2006). The empathic brain: how, when and why? *Trends Cogn Sci (Regul Ed)*, *10*(10), 435–441. DOI: 10.1016/j.tics.2006.08.008.

Decety, J. & Jackson, P. L. (2004). The functional architecture of human empathy. *Behav Cogn Neurosci Rev*, *3*(2), 71–100. DOI: 10.1177/1534582304267187.

**Decety, J., Lewis, K. L., & Cowell, J. M. (2015). Specific electrophysiological components disentangle affective sharing and empathic concern in psychopathy. *J Neurophysiol*, *114*(1), 493–504. DOI: 10.1152/jn.00253.2015.

**Decety, J., Meidenbauer, K. L., & Cowell, J. M. (2017). The development of cognitive empathy and concern in preschool children: A behavioral neuroscience investigation. *Dev Sci*. DOI: 10.1111/desc.12570.

Decety, J. & Moriguchi, Y. (2007). The empathic brain and its dysfunction in psychiatric populations: implications for intervention across different clinical conditions. *Biopsychosoc Med*, *1*, 22. DOI: 10.1186/1751-0759-1-22.

**Decety, J., Yang, C.-Y., & Cheng, Y. (2010). Physicians down-regulate their pain empathy response: an event-related brain potential study. *Neuroimage*, *50*(4), 1676–1682. DOI: 10.1016/j.neuroimage.2010.01.025.

**Fabi, S. & Leuthold, H. (2016). Empathy for pain influences perceptual and motor processing: Evidence from response force, erps, and eeg oscillations. *Soc Neurosci*, 1–16. DOI: 10.1080/17470919.2016.1238009.

**Fallon, N., Li, X., Chiu, Y., Nurmikko, T., & Stancak, A. (2015). Altered cortical processing of observed pain in patients with fibromyalgia syndrome. *J Pain*, *16*(8), 717–726. DOI: 10.1016/j.jpain.2015.04.008.

**Fan, Y. & Han, S. (2008). Temporal dynamic of neural mechanisms involved in empathy for pain: an event-related brain potential study. *Neuropsychologia*, *46*(1), 160–173. DOI: 10.1016/j.neuropsychologia.2007.07.023.

**Fan, Y.-T., Chen, C., Chen, S.-C., Decety, J., & Cheng, Y. (2014). Empathic arousal and social understanding in individuals with autism: evidence from fmri and erp measurements. *Soc Cogn Affect Neurosci*, *9*(8), 1203–1213. DOI: 10.1093/scan/nst101.

*Fitzgibbon, B. M., Enticott, P. G., Giummarra, M. J., Thomson, R. H., Georgiou-Karistianis, N., & Bradshaw, J. L. (2012). Atypical electrophysiological activity during pain observation in amputees who experience synaesthetic pain. *Soc Cogn Affect Neurosci*, *7*(3), 357–368. DOI: 10.1093/scan/nsr016.

**Gonzalez-Liencres, C., Breidenstein, A., Wolf, O. T., & Brne, M. (2016). Sex-dependent effects of stress on brain correlates to empathy for pain. *Int J Psychophysiol*, *105*, 47–56. DOI: 10.1016/j.ijpsycho.2016.04.011.

**Gonzalez-Liencres, C., Brown, E. C., Tas, C., Breidenstein, A., & Brne, M. (2016). Alterations in event-related potential responses to empathy for pain in schizophrenia. *Psychiatry Res*, *241*, 14–21. DOI: 10.1016/j.psychres.2016.04.091.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychol Bull*, *82*(1), 1–20. DOI: 10.1037/h0076157.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields i: a critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. DOI: 10.1111/j.1469-8986.2011.01273.x.

*Han, S., Fan, Y., & Mao, L. (2008). Gender difference in empathy for pain: an electrophysiological investigation. *Brain Res*, *1196*, 85–93. DOI: 10.1016/j.brainres.2007.12.062.

Hobson, H. M. & Bishop, D. V. M. (2017). The interpretation of mu suppression as an index of mirror neuron activity: past, present and future. *R Soc Open Sci*, *4*(3), 160662. DOI: 10.1098/rsos.160662.

**Ibez, A., Hurtado, E., Lobos, A., Escobar, J., Trujillo, N., Baez, S., Huepe, D., Manes, F., & Decety, J. (2011). Subliminal presentation of other faces (but not own face) primes behavioral and evoked cortical processing of empathy for pain. *Brain Res*, *1398*, 72–85. DOI: 10.1016/j.brainres.2011.05.014.

*Ikezawa, S., Corbera, S., Liu, J., & Wexler, B. E. (2012). Empathy in electrodermal responsive and nonresponsive patients with schizophrenia. *Schizophr Res*, *142*(1-3), 71–76. DOI: 10.1016/j.schres.2012.09.011.

**Ikezawa, S., Corbera, S., & Wexler, B. E. (2014). Emotion self-regulation and empathy depend upon longer stimulus exposure. *Soc Cogn Affect Neurosci*, *9*(10), 1561–1568. DOI: 10.1093/scan/nst148.

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005). How do we perceive the pain

of others? a window into the neural processes involved in empathy. *Neuroimage*, *24*(3), 771–779. DOI: 10.1016/j.neuroimage.2004.09.006.

**Jiang, C., Varnum, M. E. W., Hou, Y., & Han, S. (2014). Distinct effects of self-construal priming on empathic neural responses in chinese and westerners. *Soc Neurosci*, *9*(2), 130–138. DOI: 10.1080/17470919.2013.867899.

**Kam, J. W. Y., Xu, J., & Handy, T. C. (2014). I don't feel your pain (as much): the desensitizing effect of mind wandering on the perception of others' discomfort. *Cogn Affect Behav Neurosci*, *14*(1), 286–296. DOI:10.3758/s13415-013-0197-z.

Keil, A., Debener, S., Gratton, G., Junghfer, M., Kappenman, E. S., Luck, S. J., Luu, P., Miller, G. A., & Yee, C. M. (2014). Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*(1), 1–21. DOI: 10.1111/psyp.12147.

Kelley, K. (2017). *MBESS: The MBESS R Package*. R package version 4.3.0.

Kilner, J. (2013). Bias in a common eeg and meg statistical analysis and how to avoid it. *Clin Neurophysiol*, *124*(10), 2062–2063. DOI: 10.1016/j.clinph.2013.03.024.

Klimecki, O. & Singer, T. (2013). Empathy from the perspective of social neuroscience. In *Handbook of human affective neuroscience* (pp. 533–551). Cambridge University Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Front Psychol*, *4*, 863.

**Li, W. & Han, S. (2010). Perspective taking modulates event-related potentials to perceived pain. *Neurosci Lett*, *469*(3), 328–332. DOI: 10.1016/j.neulet.2009.12.021.

Luck, S. J. & Gaspelin, N. (2017). How to get statistically significant effects in any erp experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. DOI: 10.1111/psyp.12639.

**Lyu, Z., Meng, J., & Jackson, T. (2014). Effects of cause of pain on the processing of pain in others: an erp study. *Exp Brain Res*, *232*(9), 2731–2739. DOI:10.1007/s00221-014-3952-7.

**Mella, N., Studer, J., Gilet, A.-L., & Labouvie-Vief, G. (2012). Empathy for pain from adolescence through adulthood: An event-related brain potential study. *Front Psychol*, *3*, 501. DOI: 10.3389/fpsyg.2012.00501.

**Meng, J., Hu, L., Shen, L., Yang, Z., Chen, H., Huang, X., & Jackson, T. (2012). Emotional primes modulate the responses to others' pain: an erp study. *Exp Brain Res*, *220*(3-4), 277–286. DOI: 10.1007/s00221-012-3136-2.

**Meng, J., Jackson, T., Chen, H., Hu, L., Yang, Z., Su, Y., & Huang, X. (2013). Pain perception in the self and observation of others: an erp investigation. *Neuroimage*, *72*, 164–173. DOI: 10.1016/j.neuroimage.2013.01.024.

Munafo, M., Nosek, B., Bishop, D., Button, K., Chambers, C., Percie du Sert, N., Simonsohn, U., Wagenmakers, E., Ware, J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nat Hum Behav*, *1*(1), 0021. DOI: 10.1038/s41562-016-0021.

Olejnik, S. & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychol Methods*, *8*(4), 434–447.

Pernet, C. R., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015). Cluster-based computational methods for mass univariate analyses of event-related brain

potentials/fields: A simulation study. *J Neurosci Methods*, *250*, 85–93. DOI: 10.1016/j.jneumeth.2014.08.003.

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fmri study. *Neuroimage*, *40*(2), 409–414. DOI: 10.1016/j.neuroimage.2007.11.048.

Poldrack, R. A. & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nat Neurosci*, *17*(11), 1510–1517. DOI: 10.1038/nn.3818.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol Bull*, *86*(3), 638–641.

**Sessa, P., Meconi, F., Castelli, L., & Dell'Acqua, R. (2014). Taking one's time in feeling other-race pain: an event-related potential investigation on the time-course of cross-racial empathy. *Soc Cogn Affect Neurosci*, *9*(4), 454–463. DOI: 10.1093/scan/nst003.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*, *22*(11), 1359–1366. DOI: 10.1177/0956797611417632.

Vachon-Presseau, E., Roy, M., Martel, M. O., Albouy, G., Chen, J., Budell, L., Sullivan, M. J., Jackson, P. L., & Rainville, P. (2012). Neural processing of sensory and emotional-communicative information associated with the perception of vicarious pain. *Neuroimage*, *63*(1), 54–62.

**Vaes, J., Meconi, F., Sessa, P., & Olechowski, M. (2016). Minimal humanity cues induce neural empathic reactions towards non-human entities. *Neuropsychologia*, *89*, 132–140. DOI: 10.1016/j.neuropsychologia.2016.06.004.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. DOI: 10.1177/1745691612463078.

**Wang, T., Ge, Y., Zhang, J., Liu, J., & Luo, W. (2014). The capacity for pain empathy among urban internet-addicted left-behind children in china: An event-related potential study. *Comput Human Behav*, *33*, 56–62. DOI:10.1016/j.chb.2013.12.020.

**Wexler, B. E., Ikezawa, S., & Corbera, S. (2014). Increasing stimulus duration can normalize late-positive event related potentials in people with schizophrenia: Possible implications for understanding cognitive deficits. *Schizophr Res*, *158*(1-3), 163–169. DOI: 10.1016/j.schres.2014.07.012.

**Yang, J., Hu, X., Li, X., Zhang, L., Dong, Y., Li, X., Zhu, C., Xie, W., Mu, J., Yuan, S., Chen, J., Chen, F., Yu, F., & Wang, K. (2017). Decreased empathy response to other people's pain in bipolar disorder: evidence from an event-related potential study. *Sci Rep*, *7*, 39903. DOI: 10.1038/srep39903.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods*, *8*(8), 665–670. DOI: 10.1038/nmeth.1635.