Received: 5 January 2021 | Revised: 28 October 2021 | Accepted: 29 October 2021

DOI: 10.1002/dev.22217

RESEARCH ARTICLE

Developmental Psychobiology WILEY

Understanding the causes and consequences of variability in infant ERP editing practices

Claire Monroy[1] | Estefanía Domínguez-Martínez[2,3] | Benjamin Taylor[3,4] | Oscar Portolés Marin[5] | Eugenio Parise[3,6] | Vincent M. Reid[3,7]

[1] School of Psychology, Keele University, Keele, UK

[2] Tobii AB, Danderyd, Sweden

[3] Department of Psychology, Lancaster University, Lancaster, UK

[4] Blackpool Teaching Hospitals NHS Foundation Trust, Blackpool, UK

[5] Department of Artificial Intelligence and Cognitive Modelling, University of Groningen, Groningen, the Netherlands

[6] Department of Psychology and Cognitive Science, CIMeC, Center for Mind/Brain Sciences, University of Trento, Trento, Italy

[7] School of Psychology, University of Waikato, Waikato, New Zealand

Correspondence
Claire Monroy, Keele University, Keele, UK.
Email: c.d.monroy@keele.ac.uk

Claire Monroy and Estefanía Domínguez-Martínez are the joint first authors.

Funding information
FP7-People/Initial Training Network of the European Union, Grant/Award Number: 289404; Economic and Social Research Council Centres and Large Grants Scheme, Grant/Award Number: ES/L008955/1

Abstract

The current study examined the effects of variability on infant event-related potential (ERP) data editing methods. A widespread approach for analyzing infant ERPs is through a trial-by-trial editing process. Researchers identify electroencephalogram (EEG) channels containing artifacts and reject trials that are judged to contain excessive noise. This process can be performed manually by experienced researchers, partially automated by specialized software, or completely automated using an artifact-detection algorithm. Here, we compared the editing process from four different editors—three human experts and an automated algorithm—on the final ERP from an existing infant EEG dataset. Findings reveal that agreement between editors was low, for both the numbers of included trials and of interpolated channels. Critically, variability resulted in differences in the final ERP morphology and in the statistical results of the target ERP that each editor obtained. We also analyzed sources of disagreement by estimating the EEG characteristics that each human editor considered for accepting an ERP trial. In sum, our study reveals significant variability in ERP data editing pipelines, which has important consequences for the final ERP results. These findings represent an important step toward developing best practices for ERP editing methods in infancy research.

KEYWORDS
artifact rejection, data editing, ERP methodology, infant EEG, infant event-related potential

1 | INTRODUCTION

Event-related potentials (ERPs) measure brain responses related to external stimuli without the need for overt behavioral responses, making the ERP method an especially valuable tool for research with infants. Over the last two decades, there has been a dramatic rise in the number of published studies using an ERP approach. These studies have illuminated many aspects of infant cognitive and perceptual development (De Haan, 2007; Thierry, 2005). There are many advantages of using ERP methods with infants: for instance, ERPs provide a neuroimaging tool that is safe, noninvasive, and can be used with both typical and clinical infant populations. Researchers have also successfully identified infant ERP components that correspond to adult ERP components with known neural and cognitive substrates.

However, there are special challenges when using ERP methods with developmental populations, as is the case with many methods that require processing data from infant behavioral and/or physiological responses. One known challenge is that the automatic processing algorithms typically used to detect artifacts in the adult electroencephalogram (EEG) are often not suitable for infant EEG. To overcome this

challenge, a common approach is to manually edit the EEG on a trial-by-trial basis to select artifact-free data for inclusion in the final dataset (Hoehl & Wahl, 2012). Critically, however, it is unknown how the subjective nature of the editing process may alter the characteristics of the final dataset, particularly in terms of number of included trials and replicability of ERP effects and waveform morphologies among individual data editors.

Several factors determine the quality of the recorded data in an ERP experiment, and consequently the final ERP waveform and results. First, the EEG signal is sensitive to body and eye movement. To obtain artifact-free trials, the participant should be as still as possible during the experiment, as muscle activity and movements usually contaminates or masks brain signal (Luck, 2005). Second, in the case of visual stimuli, it is essential that the participant directs attention to the stimulus to obtain a brain response that is specific for the stimulus. Adults can be instructed to move as little as possible and to pay attention to the stimuli. When the research participants are infants, these factors, together with their limited attention span, become a substantial challenge for recording EEG (Hoehl & Wahl, 2012).

The ERP methodology typically requires that the brain response to a stimulus must be recorded over many trials and averaged to obtain a final ERP. The number of artifact-free trials per experimental condition needed for ERP analysis depends on the signal-to-noise ratio of the component under study. For adult participants, this ranges from a minimum of 40 with a typical number of around 75–100 trials per condition (Picton et al., 2000). For infants, the amplitude and latency of the brain response measured from the scalp is higher and longer, respectively, than in adults, due to reduced impedances from a thinner skull, larger postsynaptic activity from a larger number of synapses, and differences in myelination (DeBoer et al., 2005; Thierry, 2005). These differences, together with the difficulty of obtaining artifact-free trials from infants, have led researchers to commonly accept a much lower minimum of 10 trials per condition relative to adult studies (DeBoer et al., 2007; e.g., Bakker et al., 2015; Reid et al., 2007; Reynolds & Guy, 2012). Some studies have suggested that implementing a three to five trial per condition minimum criterion for inclusion into the final data analyses—and therefore increasing the potential sample size—compensates for the reduced signal-to-noise ratio (Kaduk et al., 2016; Stets & Reid, 2011; Yrttiaho et al., 2014). In both cases, infant ERP studies typically have relatively few trials per participant and much smaller sample sizes than is recommended for adult studies. Therefore, even minor changes in the trials that are included in data analyses could have large effects on the final average ERP.

As mentioned above, trial-by-trial editing process of infant data is a widespread approach currently used by developmental researchers. Trained data editors will visually analyze each trial and decide to reject trials that contain artifacts according to prespecified criteria. These criteria typically include the visible presence of muscle artifacts, eye movement artifacts or lack of attention to the stimuli (Hoehl & Wahl, 2012). During the editing process, a trial can be immediately accepted or rejected from the final set of trials, or it can be accepted with the interpolation of noisy channels.[1] This editing process is frequently carried out manually by experienced researchers or can be partially assisted by specialized software. It usually comprises two steps: first, trials are rejected in which the infant is not attending to the stimulus (e.g., looking away) based on offline visual inspection of the infant's behavior from video recordings of the experiment (Hoehl & Wahl, 2012). Second, noisy EEG channels are identified. In adults, this step is usually performed by automatic algorithms (Luck, 2005). In infants, automatic channel rejection has been applied in a few studies (e.g., Kouider et al., 2013; Kulke et al., 2016). However, the automatic algorithms for rejecting artifacts are generally designed for the adult EEG signal and therefore may not be appropriate for the infant EEG (Hoehl & Wahl, 2012). Therefore, detection of noisy channels in infant studies is usually supplemented with or replaced by manual editing (Jeschonek et al., 2010; Leppänen et al., 2007; Reid et al., 2007; Righi et al., 2014). In sum, because of the challenges associated with infant EEG data, a significant portion of ERP analyses in infant studies is performed by trained experimenters using manual editing procedures.

A critical limitation of these procedures is the potential variability due to "the human error factor," in which editors may differ in their subjective judgments regarding the quality of obtained EEG data. There exist several recommendations in the literature for identifying artifacts in the infant EEG, such as amplitude changes (e.g., DeBoer et al., 2007; Hoehl & Wahl, 2012), but there are no clear standards for the EEG characteristics or thresholds that should be considered when editing ERP trials. A few software packages are available for editing infant ERPs that synchronize the EEG data with the video recordings (e.g., NetStation®, Brain Vision Analyzer® or the graphical user interface for infant ERP analysis that uses EEGLAB; Delorme & Makeig, 2004; Kaatiala et al., 2013). These tools facilitate the editing process but do not automate it. It is therefore likely that subjective judgments and expertise of an individual researcher may impact the number and characteristics of the ERP trials accepted for final analysis, and consequently the outcomes and interpretations of the results.

The goal of the current study was to quantify the variability in current infant ERP data preprocessing procedures. We focused on the second step in the ERP editing process: the manual rejection of artifacts in EEG channels. We selected a dataset from a previously published infant ERP study (Monroy et al., 2019) that contained characteristics typical of current infant ERP studies, such as age range, experimental design, type of stimuli, and the EEG recording system. This infant EEG dataset was edited four times: three times by human editors and once by an automatic algorithm. These editors used methods and criteria for artifact rejection that they currently use in their own laboratories, but which are all representative of the current practices for editing infant ERP data.

Our study focused on three aims: first, we assessed the level of agreement among the four editors for trial acceptance and channel interpolation. Based on the complexity of the infant EEG signal and the potential "human error factor," we predicted low agreement between

[1] Channel interpolation allows for the reconstruction of the signal from a "bad" or noisy channel using information such as the average response from the surrounding electrodes.

editing processes on selection of valid trials and channels marked for interpolation. Second, we examined whether these potential differences would affect the results of statistical analyses on the average ERP waveforms; in other words, how important is consistency within the editing process? We examined the effects of variability among editors on the final ERP morphology of two ERP components: the Negative Central (Nc) and the N1. Finally, we examined the sources of variability among editors. This third aim was to identify which EEG characteristics human editors consider when manually selecting clean ERP trials. Specifically, we analyzed consistency among the specific EEG characteristics considered by the human editors and the automatic algorithm. In sum, the current study aims to provide a systematic, quantitative analysis of potential variability in trial and channel rejection, which is an important step in the infant ERP data processing pipeline.

2 | MATERIAL AND METHODS

2.1 | EEG dataset

EEG data from 10-month-old infants were selected from a preexisting infant EEG dataset (Monroy et al., 2019). This dataset was selected because it contained characteristics that are typical of current infant ERP studies, such as the infant age range, experimental design, type of stimuli, and the EEG recording system. The authors of this study used a visual ERP paradigm to examine infants' sensitivity to statistical structure within action sequences. Each ERP trial consisted of a fixation cross followed by a still image, displayed for 1 s each. In total, nine different still images were used as stimuli. EEG data were recorded using the Electrical Geodesic Incorporated (Eugene, OR, USA) 128-channel recording system with a sampling rate of 500 Hz and Cz as an online reference. The raw EEG data were filtered using a band-pass filter from 0.3 to 30 Hz and segmented into trials that comprised a 200 ms baseline and 1000 ms after the onset of the stimulus. A video of each participant was recorded and synchronized with the onset of every stimulus.

The video recordings were visually inspected for infant attention and trials during which infants were not looking at the screen were excluded from the datasets, thereby removing this factor as a source of potential variation between edited datasets. Nineteen EEG recordings were selected for the present study. In total, an average of 42.47 (SD = 4.4) trials per participant were included in the study. The trials were split into two conditions that differed from the ones used in the original study: condition 1 ($M = 17.47$) and condition 2 ($M = 25$)[2] (see Figure S1 for examples of the stimuli used in each condition). This was done to maximize the number of trials per condition. An ERP analysis based on the original data editing indicated that there was a Nc component over the frontal area of the scalp in both conditions, with one of the conditions under study having a more negative Nc than the other condition. The N1, an early visual component, was not analyzed in the original study but was selected here to address whether effects of different editing methods generalize across ERP components.

[2] The paradigm from the original study (Monroy et al. 2019) featured seven conditions that corresponded to seven different picture stimuli. Two of the stimuli featured a light ("light on") while the remaining five did not ("light off"). We collapsed these seven conditions into two conditions by combining the "light on" conditions into "condition 1," and then randomly selecting two of the five "light off" conditions and combining these into one ("condition 2"; see Figure S1). Because of how the original study was designed, the "light on" condition had fewer trials than the "light off" condition.

2.2 | ERP editing methods

Four editors—three human editors from three universities (Bangor University, Lancaster University, and Birkbeck College) and an automatic algorithm—edited the same ERP dataset. The three human editors were developmental researchers with substantial experience in editing infant EEG data, with at least three infant EEG-derived papers published per editor. They applied the methodology and criteria used in their laboratories to edit infant ERP data (see Table 1 for details). Two editors used a similar manual approach to edit the data based on trial-by-trial visual inspection. The third editor used a semi-automatic approach to edit the data: an automatic algorithm was applied to detect artifacts, followed by a trial-by-trial visual inspection. Each editor therefore used parameters for identifying artifacts based on their own laboratory methods and personal expertise.

The automatic algorithm chosen as the fourth editor had been previously utilized to edit infant ERP data (Kouider et al., 2013). The selected algorithm automatically marked channels as contaminated if the absolute voltage during an epoch exceeded $\pm 150\ \mu V$, or there was a local deviation higher than $100\ \mu V$ over a 10 samples window. Channels were interpolated using linear interpolation from the nearest electrodes. Epochs with more than 35% contaminated channels were rejected. This algorithm was implemented in Matlab.

2.3 | ERP data editing procedure

The editors' task was to examine the EEG signal for each of the ERP trials and accept or reject each trial for inclusion in the individual ERP average. Additionally, for the trials accepted, the editors reported any channel that they determined should be interpolated during the ERP data preprocessing. Finally, they decided whether they would include each individual participant in the final ERP sample. All editors were instructed to use whichever criteria for rejecting a trial or participant that they normally used with their own data.

Prior to the start of the ERP data editing, the three human editors were given the same information and guidelines: (1) a written explanation of their task as described above, (2) a general introduction to the ERP experiment, including the type of stimuli and the paradigm, (3) a hypothesis about the Nc effects that were expected over the frontal area of the scalp, and (4) a template document with a list containing the trials to be analyzed for each participant (blind to condition). The editors were also given the nineteen ERP data files already filtered and segmented and the corresponding video recordings of the infants with the stimulus onset information embedded.

Each of the four editors returned the following information: (1) a list with the trials accepted and rejected per participant, (2) for the accepted trials, a list with the channels, if any, marked for interpolation,

TABLE 1 Summary of the methodology used by all editors

Editor	Editing software	Editing method	Criteria to include a trial	Criteria to include a participant
Editor 1	NetStation®	- Visual Inspection trial by trial.	- No more than 10 channels with eye movement or other artifacts detected. - No cluster of three or more nearby channels with artifacts detected.	- At least five trials per condition.
Editor 2	Matlab® (ERPLab)	- Automatic algorithm to detect eye and slow wave artifacts. - Visual inspection trial by trial.	- Not many channels with eye artifacts or slow wave artifacts detected by the algorithm.	- At least 10 trials per condition. - 35% or more of trials accepted.
Editor 3	NetStation®	- Visual inspection trial by trial.	- No eye artifacts detected. - No alpha waves or noise over frontal channels detected. - Less than 20% channels marked for interpolation.	- At least 10 trials per condition. - Not clear drowsiness shown by alpha waves or not extremely fidgety.
Editor 4 (automatic algorithm)	Matlab® (EEGLab)	- Automatic algorithm to detect eye and movement artifacts.	- Less than 35% channels marked for interpolation.	- At least 10 trials per condition.

and (3) the decision, whenever possible,[3] of whether the participant would be included in the final ERP sample.

3 | DATA ANALYSIS

3.1 | Part 1: Agreement between editing methods

We used the Krippendorff's alpha coefficient (α) for nominal data to evaluate the level of agreement between editors. Krippendorff's alpha is a reliability coefficient designed to measure agreement between independent coders and can be applied regardless of the number of observers (Hayes & Krippendorff, 2007). The values of α range from −1 to 1, with 1 representing perfect agreement, 0 representing no reliability between coders, and values below 0 representing disagreement that exceeds what can be expected by chance. There is no minimum acceptable value of α coefficient, but a suggested threshold of $\alpha \geq 0.667$ is considered to indicate that the coded data is reliable for subsequent analyses (Krippendorff, 2004). We used the SPSS macro KALPHA to compute all the agreement values reported (Hayes & Krippendorff, 2007).

The agreement among editors was calculated based on: (1) trial assessment (n = 806), (2) participants included in the final sample (n = 19), and (3) channels marked for interpolation on trials accepted by all editors (n = 218). The number of channels marked for interpolation could vary from zero to the maximum number of channels allowed by each editor on any given accepted trial. This translates into a large range in the possible number of channels marked for interpolation and a low chance of agreement between editors. To reduce complexity in the agreement on channels marked for interpolation, we recoded the trials into a binary format: trials with one or more channels marked for interpolation were assigned a value of (1) and trials with no channels marked for interpolation were assigned a (0). The percentage of agreement on the total number of trials accepted or rejected was calculated as a secondary measure. We calculated the agreement among the four editors as well as in groups of three editors to detect any possible outlier.

3.2 | Part 2: ERP data analysis

We preprocessed the original ERP dataset four times to obtain four grand-averaged ERPs per component, one for each of the editors. The only difference in the ERP analyses was the selection of trials accepted for further processing and the channels that were interpolated, based on the method applied by each editor. For each analysis, the following steps were applied: first, trials marked as rejected were discarded. Next, a trial-by-trial channel interpolation was applied to the channels that had been marked for interpolation. ERP trials were rereferenced to the average of all channels and baseline corrected. Finally, the ERP trials were split into conditions 1 and 2. Participants that did not meet the inclusion criteria given by the editors were excluded (Table 1).

The Nc and the N1 were analyzed to explore changes in a final ERP result due to the data editing procedures. To understand whether different editors produced different results, we conducted four independent statistical analyses on the Nc and N1 components for each of the four final ERP datasets.[4] By doing this, we examined each final ERP as

[3] Editors were blinded to which trials corresponded to which condition. If the editor had a minimum criterion for the numbers of trials per condition needed in order to accept a participant for the final dataset, they could not provide this information and this step was performed for them later in the analysis.

[4] The logical analysis given our experimental design would be a repeated-measures ANOVA with the factors of condition (1, 2) and editors (1, 2, 3, 4) as within-subject variables for a specific channel region. However, this was not possible because each editor accepted different participants (only six infants were accepted by all editors, see Supplementary Materials, Table S1). We confirmed that such an ANOVA did not yield any significant main or interaction effect

though four different studies examined the same dataset. If the differences in the editing process had no or little impact, then the statistical results of the components would be the same for all editing methods. For the Nc, we selected three clusters of electrodes over the frontal area of the scalp: left (four electrodes, including F3), central (six electrodes, including Fz), and right (four electrodes, including F4; see Figure 2). These clusters were chosen based on the Nc component observed in the original study, which are also consistent with prior findings for the infant Nc component (e.g., Kaduk et al., 2016). We selected two Nc analyses based on visual inspection of the four grand averages: a mean amplitude analysis within a time window between 300 and 500 ms and a peak latency analysis within a time window between 300 and 600 ms. For each type of analysis and each final ERP, a repeated-measures ANOVA was applied with location (left, central, right) and condition (1, 2) as within-subject variables. In addition, a paired-sample *t*-test was conducted for each location with condition (1, 2) as a within-subject variable.

We also conducted two supplementary analyses for the Nc component to further examine the effect of editing method on the ERP analyses. First, to examine the relationship between editing methods and the number of accepted trials, we collapsed across all trials for editing method and repeated our analysis of mean amplitude (see Section S1). Second, to further examine variability among editing methods, we reduced the overall amount of variability by repeated our primary analyses while holding one criterion constant across all editors (Sections S2). Therefore, the number of trials required to include a participant was fixed at five across all editors. These additional analyses are reported in the Supplementary Materials.

For the N1 component, we selected three clusters of electrodes based on prior research (Richards, 2000; Richards, 2005; Xie & Richards, 2017): a left occipital region, a central occipital region, and a right occipital region (see Figure 3). Based on visual inspection of the four grand averages, we selected a mean amplitude analysis within a time window of 55–75 ms (Xie & Richards, 2017). As before, a repeated-measures ANOVA was conducted with location (left, central, right) and condition (1, 2) as within-subject variables, and a paired-sample *t*-test was applied for each location with condition (1, 2) as a within-subject variable.

3.3 | Part 3: Estimation of EEG characteristics that affected editing methods' agreement

3.3.1 | EEG characteristics

We identified 22 EEG characteristics that have been used in the literature to characterize an EEG signal and identify noise (Delorme et al., 2007; Hoehl & Wahl, 2012; Inuso et al., 2007; Junghöfer et al., 2000; Luck, 2005; Nolan et al., 2010). The selection contained both time-domain EEG characteristics (e.g., kurtosis) and frequency-domain characteristics (e.g., peak frequency). Each EEG characteristic was calculated for each of the 804 ERP trials and each of the 124 electrodes. The reduction of characteristics was based on Pearson correlation analyses. For each group of two or more variables with a correlation higher than |0.7| only one of the variables was selected. The final set contained 11 EEG characteristics. The initial and final set of EEG characteristics used can be found in the Supplementary Materials, Table S2.

3.3.2 | Statistical model

A statistical challenge with edited infant EEG data is that the electrode-level quality ratings are not observed; instead, we only observe the trial-level decision (i.e., trial accepted or rejected). For that reason, and to reduce the complexity of the statistical model, we first eliminated the electrode-level information. To do so, while losing as little information as possible, we constructed a new one-dimensional EEG characteristic for each trial based on all the electrodes. This new one-dimensional EEG characteristic described how unusual the EEG characteristic was during an ERP trial compared with a measure of that same characteristic for trials accepted by all editors. Our one-dimensional EEG characteristics $v\sim ij$ were computed as:

$$v\sim ij=\log\left\{\left(vij-v'\right)^{\mathrm{T}}\sum{}^{-1}{}_{v}\left(\underline{v}ij-v'\right)\right\} \tag{1}$$

where $\underline{v}ij$ are the vectors of each EEG characteristic for each participant i and trial j, and v' is the average EEG characteristic, calculated as the arithmetic mean across participants of the within-participant empirical mean and covariance of trials accepted by all the editors. In other words, to compute the mean, we first averaged the ERP response across accepted trials within each participant, then took the arithmetic mean of those averages across participants, and then repeated these steps for the covariance. To avoid issues of potentially very large EEG characteristics in certain electrodes and trials—which would affect our estimates of the mean and covariance—we replaced extreme observations, defined as those below the lower 2.5% or above the upper 97.5% quantiles respectively, by the 2.5% and 97.5% quantile for that characteristic across all individuals. We used the log transform because according to both the AIC and BIC (Gelman et al., 2014), the fit of our model (see details below) was better compared with the nonlogged measure.

Finally, we further standardized each $v\sim ij$ to allow us to directly compare effect sizes using the estimated regression coefficients. We fitted the following mixed-effect logistic regression model to the resulting data:

$$\log\left(\mathrm{p_i}/\left(1-\mathrm{p}_i\right)\right)=V\sim ij\beta+\varepsilon_{\mathrm{i}}, \tag{2}$$

where p_i is the editor's decision for each participant i and trial j, $V\sim ij$ is a vector with the fixed effects containing the $v\sim ij$ for each property, β is a vector of fixed effect sizes, ϵ_i is the random effect at the participant level. We used backward selection at each step excluding the least significant variable whose estimated p value was over 0.05 to arrive at the

for the mean amplitude Nc over the frontal-left area. In addition, conducting an ANOVA analysis with only the participants accepted by all editors is unrepresentative of the sample accepted by each editor and may be biased.

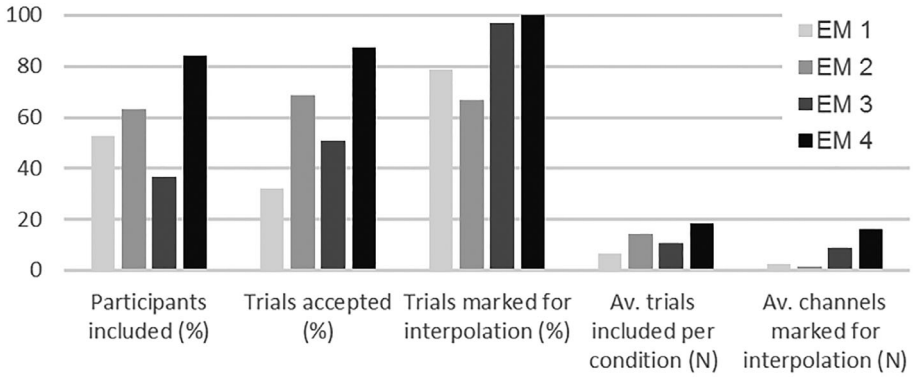

FIGURE 1 Summary of the editing process results for each editor. From left to right: (1) the percentage of participants included in the final ERP sample. (2) The percentage of trials marked as accepted. (3) The percentage of trials marked for interpolation. This percentage was based only on the trials marked as accepted by each editor. (4) The average number of trials per condition that were included in the final ERP sample. This was calculated only from the participants included. (5) The average number of channels marked for interpolation. This average was calculated only from trials that contained at least one channel marked for interpolation

final model. We fitted three models, one for each of the editors, and obtained the significant variables that influenced each editor's decision. We also obtained, for each model, the estimated fixed effects for each significant variable.

4 | RESULTS

4.1 | Part 1: Agreement between editing methods

The number of participants included by each editor in the final ERP sample differed: editor 1 included 10 infants (M = 6.8 trials per condition), editor 2 included 12 participants (M = 14.5 trials per condition), editor 3 included seven participants (M = 10.8 trials per condition) and the automatic algorithm, hereafter labeled editor 4, included 16 participants (M = 18.5 trials per condition; Figure 1). This difference is a direct consequence of each editor's inclusion criteria and the number of trials accepted by each editing method, which ranged from 32% of accepted trials by editor 1 to 87% of accepted trials by editor 4. Regarding interpolation, the percentage of trials with at least one channel marked for interpolation varied from 67% of the accepted trials by editor 2 to 100% of the accepted trials by editor 4. Also, the average number of channels marked for interpolation included high variability, from 1.8 channels on average for editor 2 to 16.05 channels on average for editor 4 (Figure 1).

These differences were confirmed by the Krippendorff's alpha coefficients (Table 2). The overall α agreement among editors was 0.275 for trial acceptance and 0.409 for accepted participants. Very low agreement—close to chance—was obtained for the trials marked for interpolation (0.061). Small differences in α values were obtained when excluding any one editor from the calculations, revealing that no single editor was an outlier that caused the low α values. A slightly higher agreement was found among only the three human editors for all the variables when excluding the algorithm, although for all groupings the alpha coefficient was still close to the suggested reliability threshold of $\alpha \geq 0.667$.

The percentage of agreement also confirms low agreement on the numbers of accepted and rejected trials. 27.08% of the trials were accepted by all the editors, and only 10.06% were rejected by all of them (Table 2). When observing this percentage in groups of three editors, it can be noted that editor 1 substantially affected this result, which reaches 46.21% of trials accepted when not taking editor 1 into account. For the agreement on rejected trials, editor 4 disproportionately affected this result, which reaches 26.46% of rejected trials when not taking editor 4 into account.

4.2 | Part 2: ERP data analysis

4.2.1 | Nc results

Figure 2 shows the frontal left, central, and right clusters of electrodes of the four grand average ERP waveforms calculated for each editor. The final sample that editor 2 used for the statistical analyses was 11 participants. One participant, originally accepted by editor 2, had to be excluded because the averaged ERP contained a large amount of eye artifacts and substantially altered the resulting ERP. An Nc component was observed in the four grand averages over the frontal electrodes, but with variability in their amplitude levels as well as in latency to peak and the amplitude difference between conditions (Figure 2).

4.2.2 | Mean amplitude analysis

The repeated-measures ANOVA indicated no effects in any of the editors' ERP data for location ($ps > .07$) or the location × condition interaction ($ps > .30$). Data from editor 2 revealed a marginally significant effect for condition ($p = .09$). Paired-sample t-tests for each location showed a significant effect of condition in the frontal left cluster only, $t(10) = -2.29$, $p = .045$; $d = 0.69$, revealing that amplitude in condition 1 was significantly more negative ($M = -5.03$, SD = 6.27) than condition 2 ($M = 0.88$, SD = 5.23). Data from editor 1 also revealed a marginally

TABLE 2 Agreement results among all editors and by groups of three

Editors	Krippendorff's alpha			Percentage agreement	
	Trial assessment α (95% CI)	Participants included α (95% CI)	Trials interpolated α (95% CI)	Trials accepted	Trials rejected
All	0.275 (0.085, 0.460)	0.409 (0.224, 0.591)	0.061(−0.209, 0.318)	27.08%	10.06%
Ed1, Ed2, and Ed3	0.381 (0.200, 0.560)	0.517 (0.310, 0.724)	0.146 (−0.091, 0.366)	27.20%	26.46%
Ed1, Ed2, and Ed4	0.139 (−0.069, 0.338)	0.457 (0.186, 0.690)	0.041 (−0.222, 0.283)	29.57%	10.06%
Ed1, Ed3, and Ed4	0.201 (0.002, 0.389)	0.293 (0.046, 0.540)	0.061 (−0.358, 0.469)	29.07%	12.17%
Ed2, Ed3, and Ed4	0.320 (0.113, 0.533)	0.345 (0.091, 0.600)	-0.132 (−0.436, 0.171)	46.21%	10.19%

Note. From left to right: (1) Krippendorff's alpha agreement values and confidence intervals (CI) for trial assessment (accepted, rejected), participants (included, excluded) and number of trials marked for interpolation (one or more channels marked, no channels marked). The agreement on trials interpolated was calculated taking only into account the trials accepted. Alpha confidence intervals at the 95% level were calculated by applying bootstrap analysis of 10,000 samples (Hayes & Krippendorff, 2007). (2) Percentage agreement on trials accepted and rejected where all the editors of each group agreed on the same decision.

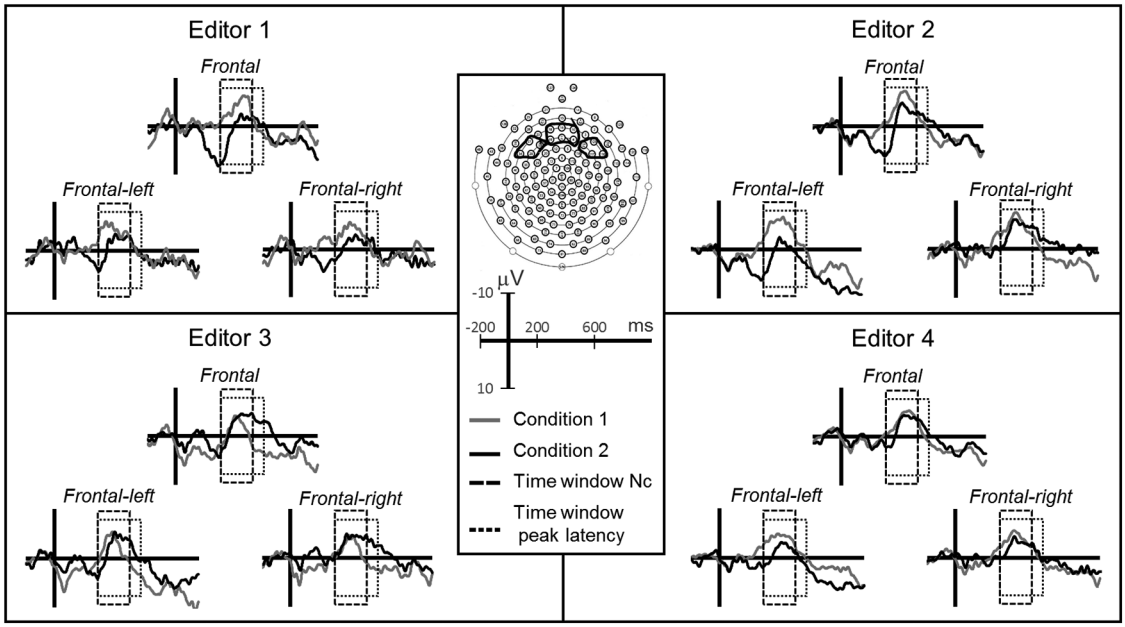

FIGURE 2 The Nc: grand average of each editor across the frontal area split by condition. For each editor, the same three clusters of electrodes are displayed: frontal-left, frontal, and frontal-right. The variability in amplitude and latency of the ERP components is notable when comparing the grand averages

significant effect of condition ($p = .07$), but follow-up paired-sample t-tests did not reveal any significant differences between conditions for any of the locations ($ps > .06$). None of the other editors had any significant effects for condition in any of the locations ($ps > .10$).

4.2.3 | Peak latency analysis

There were two significant effects in the ERP data via editor 1 as indicated by the repeated measures ANOVA analysis. First, there was a significant main effect for condition, $F(1,9) = 5.37$, $p = .046$; $\eta_p^2 = 0.38$. The peak latency of the Nc component in condition 2 was significantly longer ($M = 685.73$ ms, SD = 72.40) compared with condition 1 ($M = 638.93$, SD = 71.25). Second, there was a significant main effect for location, $F(2, 18) = 6.67$, $p = .007$; $\eta_p^2 = 0.43$. The peak latency of Nc in the frontal central cluster was significantly longer ($M = 690.2$ ms, SD = 60.26) compared with the frontal left cluster ($M = 636$ ms, SD = 75.89), $t(9) = -3.737$, $p = .005$; $d = 1.18$. None of the other editors had any significant effects in the peak latency for condition or location.

4.2.4 | N1 results

Figure 3 shows the N1 component observed over the occipital left, central, and right clusters of electrodes of the four grand average ERP waveforms calculated for each editor. The ANOVA indicated no main effects in any of the editors' data for region ($ps > .41$) or condition ($ps > .31$) and no region × condition interaction effects ($ps > .85$).

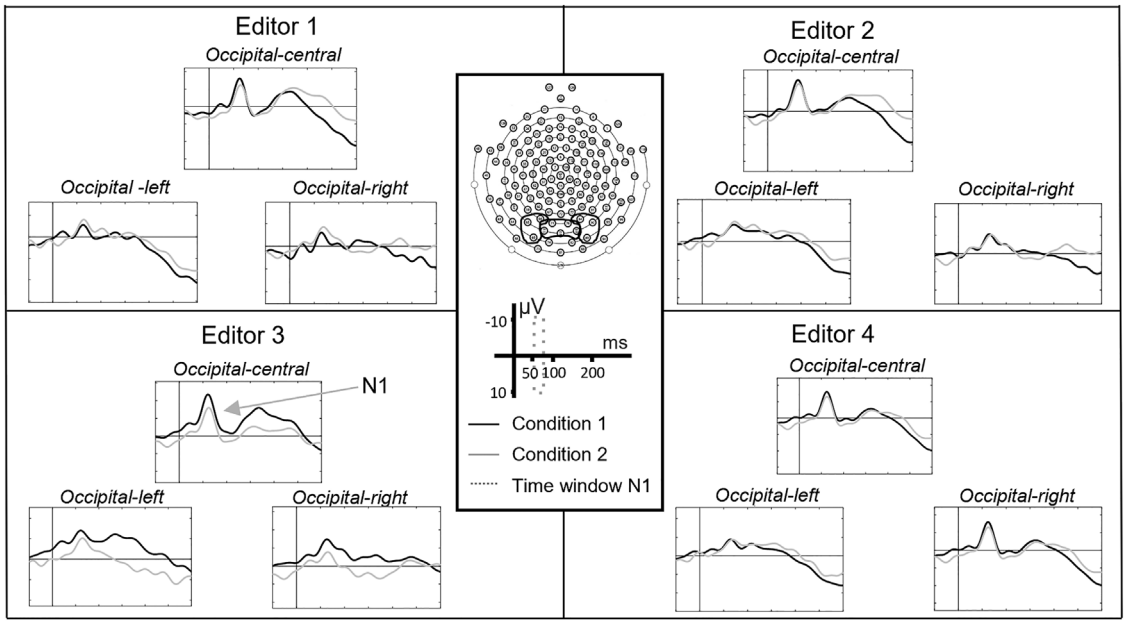

FIGURE 3 The N1: grand average of each editor across the occipital area split by condition. For each editor, the same three clusters of electrodes are displayed: occipital-left, occipital-central, and occipital-right. Like the Nc, there is notable variability in the overall waveform

TABLE 3 List of final included EEG characteristics and EEG characteristics estimated by the three mixed-effect logistic regression models

EEG characteristic	Model 1 (Editor 1)	Model 2 (Editor 2)	Model 3 (Editor 3)
Amplitude range	Ns	Ns	0.547 (0.414; 0.722)
Linear trend	Ns	0.720 (0.539; 0.964)	Ns
Deviation from channel mean	Ns	Ns	Ns
Signal-to-noise ratio	0.369 (0.270; 0.502)	0.553 (0.400; 0.763)	0.505 (0.375; 0.680)
Number of local maxima	Ns	0.711 (0.519; 0.975)	0.621 (0.459; 0.839)
Kurtosis	0.416 (0.300; 0.577)	0.666 (0.520; 0.851)	0.569 (0.442; 0.731)
Power at 0–4 Hz	0.443 (0.337; 0.582)	0.549 (0.399; 0.756)	Ns
Power at 8–13 Hz	Ns	Ns	Ns
Power at 30–60 Hz	Ns	Ns	0.713 (0.523; 0.972)
Spectral edge frequency	Ns	0.757 (0.573; 0.999)	Ns
Peak frequency	1.378 (1.113; 1.70)	Ns	Ns

Note. List of final EEG characteristics included (first column) and EEG characteristics estimated by the three mixed-effect logistic regression models that influenced the decision of each editor applied by human editors. Significant EEG characteristics of an editor include their estimated regression coefficients and, in brackets, their confidence intervals at 95%. The estimated regression coefficients can be interpreted as odds ratios. Thus, a coefficient below 1 can be interpreted as the reduction in odds of accepting a trial given a unit increase in the associated EEG characteristic. A coefficient above 1 can be interpreted as the increase in odds of accepting a trial given a unit increase in the associated variable. Ns: not significant.

Paired-sample *t*-tests also revealed null effects for all regions, among all editors (*ps* > .09).

4.3 | Part 3: Estimation of EEG characteristics that affected editors' agreement

Table 3 shows the results of the three mixed-effect logistic regression models, one for each criterion applied by the human editors. The fixed effects in these models measure the relationship between the EEG characteristics and the probability of accepting the trial. The significant EEG characteristics can be interpreted as the characteristics that influenced each editor's decision.

Our main finding is that the EEG characteristics that were significant vary across editors. Two characteristics—kurtosis and SNR—were significant in the three models (Table 3). Thus, it is estimated that the three editors took them into consideration when editing the ERP data. Two characteristics—number of local maxima and power at

0–4 Hz—were significant in two of three models. Five characteristics—amplitude range, linear trend, power at 30–60 Hz, spectral edge frequency and peak frequency—were significant in only one of the three models. Two characteristics—deviation from the channel mean and power at 8–13 Hz—were not significant in any of the three models. All the regression coefficients except one were estimated to be below 1. This can be interpreted as the reduction in odds of accepting a trial given a unit increase in the specific EEG characteristic. That is, the larger an EEG characteristic value, the higher the chance of the trial being rejected by the editor with that EEG characteristic estimated as significant (recall that a higher EEG characteristic means a more unusual signal). There is only one EEG characteristic for editor 1—peak frequency—with an estimated regression coefficient above 1. The interpretation of this finding is that editor 1 tended to accept trials in which the peak frequency was more unusual. Participant-level random effects from each model can be found in the Supplementary Materials, Table S3.

5 | DISCUSSION

Event-related potentials are a valuable tool in infancy research and offer many advantages to both researchers and clinicians. A current challenge for developmental researchers who use infant ERP methodology is the lack of standardized techniques or guidelines for editing ERP data. The aim of the present study was to evaluate the variability within current infant ERP editing methods and the potential effect of this variability on final ERP results. This study represents a first step toward the long-term goal of establishing standard guidelines for infant ERP data editing methods, by providing data on the consequences of the natural variability among methods for artifact rejection that currently exist "in the wild."

Four editors—three experienced humans and one algorithm—edited the same infant ERP dataset to identify artifact-contaminated ERP trials. One of the human editors used a semi-automatic approach by using manual trial-by-trial rejection after applying the automatic algorithm. We found low agreement among editors in the number of participants included in the final sample, the trials accepted for further analysis and the channels marked for interpolation. Because of the low agreement, the morphology of the grand averages varied in the amplitude and latency of the resulting ERP components. This variability was substantial enough to produce inconsistent statistical results regarding the differences between conditions for the amplitude and latency of the Nc component. On the other hand, consistent null effects were found for all editors for the N1 component despite this variability, though there are visible differences in the overall waveform as with the Nc. These findings demonstrate that the effects of editor variability on the outcome of a statistical analysis depends on the specific component under examination and possibly whether there are true effects in the data.

The method applied by the three human editors relied on a trial-by-trial visual inspection of the EEG data to identify the channels and trials with unacceptable levels of noise in the EEG signal (for editor 2, this was applied in addition to an automatic algorithm). The criteria described by the editors to accept a trial was similar between them (Table 1). They described a focus primarily on detecting physiological artifacts such as eye movements, blinks, body movement, slow waves, or alpha waves. However, the results of the statistical models applied to each editor suggest that one of the causes of the low agreement may be different EEG characteristics considered by each of them to evaluate a trial. This variability highlights the level of complexity of infant ERP data and the lack of standardized methods and definitions to evaluate the noise in infant EEG data.

Certain EEG characteristics included in our model—for example, amplitude range and linear trend—were more likely to be included as significant characteristics considered by the editors, possibly because they are easier to measure or visually inspect. These characteristics were, however, significant only for individual editors. The two EEG characteristics that were significant for all three editors were kurtosis and signal-to-noise ratio. However, these EEG characteristics are not easily assessed by visual inspection. It is unlikely that the editors assessed the level of noise in the EEG data by consciously using the significant EEG characteristics that the model has predicted for each of them. In our view, the main value of the results of the statistical model is not related to the specific significant EEG characteristics obtained for each editor but, rather, the evidence for different criteria and thresholds of what is considered noise within infant EEG by current editors.

The results of this study indicate that one variable to consider when humans edit complex data is the human error factor, which is inherent to any human process. For example, editor 2 included one participant in the final sample whose data contained a high number of eye movement artifacts. As explained in the results section, that participant had to be excluded from the final sample of editor 2 for this reason, which is a common practice in ERP methodology when an ERP grand average is distorted due to artifacts attributable to one participant (Luck, 2005). However, editor 1 and editor 4 included that participant in their final samples and the individual ERP averages were not contaminated with eye movements for those editors. This type of human error contributes to the decrease in agreement between editors and, importantly, contributes to the differences found in the grand averages.

The agreement between editors was reduced when the automatic algorithm (editor 4) was included. From the number of trials accepted and channels interpolated by each editor, the method applied by human editors and the automatic algorithm seem to have two editing strategies: the criteria used by the human editors tended to reject trials when few EEG channels where contaminated with noise, and usually marked only a small number of channels for interpolation. The automatic algorithm tended to accept trials with a higher number of contaminated EEG channels but interpolated them first. Both strategies resulted in grand average with expected ERP morphologies and amplitudes. Since there are no gold standards in infant ERP data editing, none of the editing strategies can be considered better than the others a priori. However, the interpolation of channels creates a new EEG signal based mainly on the nearby channels (Luck, 2005). Therefore, an editing strategy where many channels are interpolated needs to be applied carefully to make sure that the EEG signal is not being altered excessively by creating correlations among channels.

The automatic algorithm used in the present study contained thresholds for accepting trials that were more liberal when compared with the human editors. The higher tolerance of the automatic algorithm suggests that it could have included noisy trials into the averaged ERP. However, the grand average ERP created by the algorithm did not reflect any effect of potential noise. Rather, the higher number of trials and participants accepted seem to have had a positive effect in the signal-to-noise ratio of the grand average ERP. It is unknown whether the higher number of channels interpolated could have influenced the ERP components' morphology and caused false negative results in the algorithm dataset (editor 4).

A limitation of the automatic algorithm is that it was initially applied to adult EEG data and was adapted to infant levels of noise (Kouider et al., 2013). Another limitation of utilizing this type of algorithm for infant ERP data processing is that it is based on general amplitude level rules. However, infant EEG data have a high interindividual variance and greater delta and theta band activity than adult EEG data (Thierry, 2005). Finding a unique amplitude threshold that is valid for all the infant participants may not be possible. More broadly, algorithms based only on one EEG characteristic (such as the amplitude threshold of the algorithm used in the present study) may not capture the complexity of infant EEG data. Although we selected the current algorithm based on the published study by Kouider et al. (2013), there are now newer algorithms that are better suited for infant ERP data (e.g., Bigdely-Shamlo et al., 2015; Gabard-Durnam et al., 2018). In addition, when evaluating the utility of a particular algorithm, it may be important to consider the difference between EEG versus ERP artifact detection (in which the EEG signal is segmented into discrete trials and therefore the number of channels per trial is relevant). Given the continual development and improvement of automated editing methods, future work could compare different algorithms with one another.

As hypothesized, variability among editors had consequences for the morphology of the final Nc component. There was a notable variability in the amplitude and latency of the Nc between the four grand averages. Also, what is commonly more important for ERP studies, the grand averages of each editor showed variability in the amplitude and latency between conditions for the Nc component. The direct explanation of this variability between grand averages is that a different set of accepted trials and participants were included by each editor. Each editor selected varying sets of trials that likely contained different signal-to-noise ratios: (1) because of higher amounts of noise or (2) a larger response to the stimuli. Regarding the former, our results suggest that current methods for infant ERP data editing do not have a common threshold of noise to reject a trial, with the consequence that the grand averages are calculated with trials that contain different amounts of noise. Regarding the latter, the changes in the infant's brain response to the stimulus over time during an ERP experiment also likely contribute to the variability of the ERP morphology (Stets & Reid, 2011).

It should be noted that we are not able to compare the results obtained in this study with any "gold standard" related to data processing. There are no right or wrong results for any of the methods, and therefore we can only speak to the variability that exists between editing methods. Consequently, our aim at this stage was not to provide guidelines for choosing a particular method or how to standardize the data editing approaches. Rather, our aim was to characterize the variability that currently exists in the field as a first step toward the goal of developing such guidelines (e.g., see Picton et al., 2000, for an example of such guidelines in adult ERP research). To make further progress toward this goal, infant ERP researchers should thoroughly report the criteria used to identify artifacts, which is currently often omitted from publications. Another suggestion, in the meantime, would be to always have two independent researchers edit the data and to report inter-rater reliability, as is common and expected for behavioral data that are manually coded.

Regarding sample size, these are quite variable among infant studies, but it is uncommon to find significant effects with fewer than 10 participants due to effect sizes. In the current study, the final sample size of some of the editors would not be considered suitable to extract conclusions. It is quite probable that were this a standard experimental study, the researchers would have continued testing infants until getting a larger sample size could be reported (as indicated in Stets et al., 2012). It should also be pointed out that this study had no strong rationale about the expected Nc or N1 differences between conditions. It is possible that the variability found in the current editors have less impact in the final ERP morphology and differences between conditions of an ERP study where the expected effect size of the component of interest is greater. Despite this, these issues do not change the conclusions inferred from the results about the existing low agreement among current editing processes in infant ERP studies.

Several factors need to be better understood before improved data editing procedures can be adopted by the field. Some examples are whether certain infant participants are easier to edit than others because of their behavior or EEG signal. Another issue is how the level of noise in the infant EEG evolves or changes during an experiment. Until these factors are more robustly understood, it is difficult to determine appropriate pathways for methodological improvement. In addition, a more comprehensive exploration should be made regarding how interpolation of channels affects the final ERP and its components. Using statistical models to address these questions could be a valuable way to learn more about infant EEG data and to continue exploring which EEG characteristics and values are important to consider when assessing noise in infant EEG data. We believe that there would be large benefits for the infant ERP community from examining such methodological questions that may help to establish reliable editing procedures in the future.

6 | CONCLUSIONS

This study highlights the high amount of variability among current infant ERP editing methods within the field. Our results demonstrate high levels of variability in the selection of ERP trials because of noise in the infant EEG signal. This variability introduced by the editing processes can have a substantial effect on the final ERP morphology and in the amplitude and latencies of ERP components. These effects

also depend on the target ERP component. By demonstrating and characterizing the variability in current ERP data editing methods, we provide a starting point from which future work can aim to develop standardized methods for editing infant ERP data, such as those that exist in the field of adult ERP research. This will need to be preceded by a better understanding of the infant EEG signal and noise characteristics in general, most likely on a component-by-component basis.

ACKNOWLEDGMENTS

Special thanks to the three expert EEG researchers who anonymously volunteered their time to edit the data reported in this study. This research was supported by the Marie Curie FP7-PEOPLE / Initial Training Network of the European Union (ACT 289404) and the Economic and Social Research Council (ES/L008955/1).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Claire Monroy https://orcid.org/0000-0002-5044-5185

REFERENCES

Bakker, M., Kaduk, K., Elsner, C., Juvrud, J., & Gredebäck, G. (2015). The neural basis of non-verbal communication-enhanced processing of perceived give-me gestures in 9-month-old girls. *Frontiers in Psychology*, *6*(FEB), 1–6. https://doi.org/10.3389/fpsyg.2015.00059

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*, 16. https://doi.org/10.3389/fninf.2015.00016

De Haan, M. (2007). *Infant EEG and event-related potentials* (1st ed.). Psychology Press.

DeBoer, T., Scott, L., & Nelson, C. A. (2005). Event-Related Potentials in developmental populations. In *Event-related potentials: A methods handbook* (pp. 263–297). MIT press.

DeBoer, T., Scott, L. S., & Nelson, C. A. (2007). Methods for acquiring and analysing infant event-related potentials. *Infant EEG and Event-Related Potentials*, 5–37.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open-source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage*, *34*(4), 1443–1449. https://doi.org/10.1016/j.neuroimage.2006.11.004

Gabard-Durnam, L. J., Mendez Leal, A. S., Wilkinson, C. L., & Levin, A. R. (2018). The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data. *Frontiers in Neuroscience*, *12*, 97. https://doi.org/10.3389/fnins.2018.00097

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Hoehl, S., & Wahl, S. (2012). Recording infant ERP data for cognitive research. *Developmental Neuropsychology*, *37*(3), 187–209. https://doi.org/10.1080/87565641.2011.627958

Inuso, G., La Foresta, F., Mammone, N., & Morabito, F. C. (2007). Brain activity investigation by EEG processing: Wavelet analysis, kurtosis and Renyi's entropy for artifact detection. *Proceedings of the 2007 International Conference on Information Acquisition, ICIA*, 195–200. https://doi.org/10.1109/ICIA.2007.4295725

Jeschonek, S., Marinovic, V., Hoehl, S., Elsner, B., & Pauen, S. (2010). Do animals and furniture items elicit different brain responses in human infants? *Brain & Development*, *32*(10), 863–71. https://doi.org/10.1016/j.braindev.2009.11.010

Junghöfer, M., Elbert, T., Tucker, D. M., & Rockstroh, B. (2000). Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology*, *37*(4), 523–532. https://doi.org/10.1111/1469-8986.3740523

Kaatiala, J., Yrttiaho, S., Forssman, L., Perdue, K., & Leppänen, J. (2013). A graphical user interface for infant ERP analysis. *Behavior Research Methods*, https://doi.org/10.3758/s13428-013-0404-4

Kaduk, K., Bakker, M., Juvrud, J., Gredebäck, G., Westermann, G., Lunn, J., & Reid, V. M. (2016). Semantic processing of actions at 9 months is linked to language proficiency at 9 and 18 months. *Journal of Experimental Child Psychology*, *151*, 96–108, https://doi.org/10.1016/j.jecp.2016.02.003

Kouider, S., Stahlhut, C., Gelskov, S. V., Barbosa, L. S., Dutat, M., de Gardelle, V., Christophe, A., Dehaene, S., & Dehaene-Lambertz, G. (2013). A neural marker of perceptual consciousness in infants. *Science*, *340*(6130), 376-380. https://doi.org/10.1126/science.1232509

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1093/hcr/30.3.411

Kulke, L., Atkinson, J., & Braddick, O. (2016). Neural mechanisms of attention become more specialised during infancy: Insights from combined eye tracking and EEG. *Developmental Psychobiology*, (March), 1–11. https://doi.org/10.1002/dev.21494

Leppänen, J. M., Moulson, M. C., Vogel-farley, V. K., & Nelson, C. A. (2007). An ERP study of emotional face processing in the adult and infant brain. *Child Development*, *78*(1), 232–245. https://doi.org/10.1111/j.1467-8624.2007.00994.x

Luck, S. J. (2005). *An introduction to the event-related potential technique*. MIT press.

Monroy, C. D., Gerson, S. A., Domínguez-Martínez, E., Kaduk, K., Hunnius, S., & Reid, V. (2019). Sensitivity to structure in action sequences: An infant event-related potential study. *Neuropsychologia*, *126*, 92–101. https://doi.org/10.1016/j.neuropsychologia.2017.05.007

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods*, *192*(1), 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. a, Johnson, R., G. A. Miller, W. Ritter, D. S. Ruchkin, M. D. Rugg, & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*(2), 127-152. https://doi.org/10.1111/1469-8986.3720127

Reid, V. M., Csibra, G., Belsky, J., & Johnson, M. H. (2007). Neural correlates of the perception of goal-directed action in infants. *Acta Psychologica*, *124*(1), 129-138. https://doi.org/10.1016/j.actpsy.2006.09.010

Reynolds, G. D., & Guy, M. W. (2012). Brain-behavior relations in infancy: Integrative approaches to examining infant looking behavior and event-related potentials. *Developmental Neuropsychology*, *37*(3), 210-225. https://doi.org/10.1080/87565641.2011.629703

Richards, J. E. (2000). Localizing the development of covert attention in infants with scalp event-related potentials. *Developmental Psychology*, *36*(1), 91. https://doi.org/10.1037/0012-1649.36.1.91

Richards, J. E. (2005). Localizing cortical sources of event-related potentials in infants' covert orienting. *Developmental Science*, *8*(3), 255–278. https://doi.org/10.1111/j.1467-7687.2005.00414.x

Righi, G., Westerlund, A., Congdon, E. L., Troller-Renfree, S., & Nelson, C. a. (2014). Infants' experience-dependent processing of male and female faces: Insights from eye tracking and event-related potentials. *Developmental Cognitive Neuroscience*, *8*, 144–152. https://doi.org/10.1016/j.dcn.2013.09.005

Stets, M., & Reid, V. M. (2011). Infant ERP amplitudes change over the course of an experimental session: Implications for cognitive processes and methodology. *Brain and Development*, *33*(7), 558–568. https://doi.org/10.1016/j.braindev.2010.10.008

Stets, M., Stahl, D., & Reid, V. M. (2012). A meta-analysis investigating factors underlying attrition rates in infant ERP studies. *Developmental Neuropsychology*, *37*(3), 226-252. https://doi.org/10.1080/87565641.2012.654867

Thierry, G. (2005). The use of event related potentials in the study of early cognitive development. *Infant and Child Development*, *14*, 85-94, https://doi.org/10.1002/icd.353

Yrttiaho, S., Forssman, L., Kaatiala, J., & Leppänen, J. M. (2014). Developmental precursors of social brain networks: The emergence of attentional and cortical sensitivity to facial expressions in 5 to 7 months old infants. *PLoS One*, *9*(6), e100811. https://doi.org/10.1371/journal.pone.0100811

Xie, W., & Richards, J. E. (2017). The relation between infant covert orienting, sustained attention, and brain activity. *Brain Topography*, *30*(2), 198–219. https://doi.org/10.1007/s10548-016-0505-3

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Monroy, C., Domínguez-Martínez, E., Taylor, B., Marin, O. P., Parise, E., & Reid, V. M. (2021). Understanding the causes and consequences of variability in infant ERP editing practices. *Developmental Psychobiology*, *63*, e22217. https://doi.org/10.1002/dev.22217