Published in final edited form as:

Psychophysiology. 2020 August ; 57(8): e13566. doi:10.1111/psyp.13566.

Adjusting ADJUST: Optimizing the ADJUST Algorithm for Pediatric Data Using Geodesic Nets

Stephanie C. Leach[1], **Santiago Morales**[1], **Maureen E. Bowers**[2], **George A. Buzzell**[1], **Ranjan Debnath**[1], **Daniel Beall**[3], **Nathan A. Fox**[1]

[1]Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA

[2]Neuroscience and Cognitive Science Program, University of Maryland, College Park, MD, USA

[3]Department of Computer Science, University of Maryland, College Park, MD, USA

Abstract

A major challenge for electroencephalograph (EEG) studies on pediatric populations is that large amounts of data are lost due to artifacts (e.g., movement and blinks). Independent component analysis (ICA) can separate artifactual and neural activity, allowing researchers to remove such artifactual activity and retain a greater percentage of EEG data for analyses. However, manual identification of artifactual components is time consuming and requires subjective judgement. Automated algorithms, like ADJUST and ICLabel, have been validated on adults, but to our knowledge no such algorithms have been optimized for pediatric data. Therefore, in an attempt to automate artifact selection for pediatric data collected with geodesic nets, we modified ADJUST's algorithm. Our "adjusted-ADJUST" algorithm was compared to the "original-ADJUST" algorithm and ICLabel in adults, children, and infants on three different performance measures: respective classification agreement with expert coders, the number of trials retained following artifact removal, and the reliability of the EEG signal after preprocessing with each algorithm. Overall, the adjusted-ADJUST algorithm performed better than the original-ADJUST algorithm and no ICA correction with adult and pediatric data. Moreover, it performed better than ICLabel for pediatric data. These results indicate that optimizing existing algorithms improves artifact classification and retains more trials, potentially facilitating EEG studies with pediatric populations. Adjusted-ADJUST is freely available under the terms of the GNU General Public License at: https://github.com/ChildDevLab/MADE-EEG-preprocessing-pipeline/tree/master/adjusted_adjust_scripts.

Keywords

Electroencephalography; EEG artifacts; automated artifact classification algorithm; developmental research; geodesic sensor net; independent component analysis

Correspondence concerning this article should be addressed to Santiago Morales, Department of Human Development and Quantitative Methodology, 3304 Benjamin Building, University of Maryland, College Park, MD, 20742. Office phone number: (301) 405-8490. moraless@umd.edu.

Conflict of interest statement: The authors have no conflict of interest.

1. Introduction

Electroencephalography (EEG) is a useful, noninvasive method for studying brain functioning (Luck, 2014). EEG is acquired by placing electrodes or sensors on the scalp to record electrical signals generated by the brain. However, in addition to recording neural activity, these sensors also record environmental noise, such as muscle activity from movements and changes of the electrical dipoles of the eyes during blinks or saccades (Luck, 2014). These nonneural sources of activity are often referred to as artifacts, and they can contaminate signals of interest in the EEG and potentially lead to the misinterpretation of results (Luck, 2014). This problem is exacerbated in pediatric research because children often have a hard time sitting still and focusing for long periods of time. As a result, studies involving infants and children are often shorter (incorporating fewer trials) and a greater proportion of trials must ultimately be rejected due to artifact contamination, as compared to adult studies. As such, preprocessing methods that can remove artifacts instead of deleting whole trials are of special importance in developmental research.

One such artifact removal method is independent component analysis (ICA), which decomposes the EEG data into maximally independent components, separating artifacts from neural data (for more detail on how ICA works as an artifact rejection method see Urigüen & Garcia-Zapirain, 2015 or Jung et al., 2000). Once components reflecting artifacts are identified, they can be removed and the data can be reconstructed, thereby subtracting artifacts from the EEG data without having to remove entire epochs of data (Jung et al., 2000). However, the identification of artifactual components takes time and often requires subjective judgement. Moreover, the number of components generated by ICA depends on the number of unique sources of information (i.e., electrodes present in the dataset), such that the number of components is often equal to the number of electrodes used for recording (Luck, 2014). For high-density channel systems, which can have 64, 128, or even 256 channels, manually inspecting each component to determine if it is artifactual can be quite time consuming. Furthermore, while artifacts generally show consistent activation patterns (e.g., high activity over frontal sites and low activity over posterior sites for blinks; Chaumon, Bishop, & Busch, 2015), there are many cases where it is unclear whether a component reflects an artifact or neural activity that follows a similar activation pattern. As such, determining whether or not an ambiguous component reflects neural or artifactual activity is not always straightforward and requires subjective judgement. Relying on subjective judgements, which have the potential for introducing systematic bias into the data, can be especially problematic for large, multi-site studies that rely on multiple people to analyze the data. That is because not all individuals will select the same ICA components for removal and they may unintentionally remove components that reflect neural data, which can affect Event-Related Potentials (ERPs) in a variety of ways (e.g., increasing/decreasing component amplitude). For these reasons, using subjective judgement to identify components reflecting artifactual activity can impair attempts to replicate findings of independent research groups. That is, removing or retaining a different subset of ICA components could potentially be the difference between a significant and a nonsignificant effect.

In response to these issues, numerous automated artifact selection algorithms have been created to expedite the artifact selection process and increase objectivity. One such widely used algorithm is ADJUST (Mognon, Jovicich, Bruzzone, & Buiatti, 2011), which uses a combination of the spatial and temporal features of the components to classify them as artifactual or neural. In our experience, ADJUST often misclassifies components in pediatric data collected with geodesic nets (e.g., components reflecting blinks or horizontal eye movements are not classified as artifactual), which may result from the fact that ADJUST was optimized and validated using adult data.

In addition to ADJUST, a recently released algorithm, called ICLabel (Pion-Tonachini, Kreutz-Delgado, & Makeig, 2019), provides another automated artifact selection method by using a machine learning classifier. While ICLabel exhibited superior performance on adult data, the authors noted that there was anecdotal evidence suggesting underperformance with infant data that was likely due to a lack of infant data in the training dataset (ICLabel's machine-learning approach is optimized based on a labeled set of training data). Moreover, even though there were a few studies with children included in the training dataset for ICLabel, the majority of the studies in the training dataset were with adult populations using the 10–20 system. Ultimately, ICLabel is optimized for adults and the 10–20 system, which means that it may also underperform on pediatric data collected with geodesic nets. In fact, to the best of our knowledge, the performance of these algorithms on pediatric data has not been formally evaluated and no automated component classification algorithms have been optimized for pediatric data.

There are several physiological and behavioral differences between developmental and adult populations that could lead existing algorithms validated in adults to underperform in EEG data from pediatric populations. For example, adults have a higher blink rate than pediatric populations (Lawrenson, Murphy, and Esmaeelpour, 2003). Furthermore, shorter attention spans in pediatric populations result in shorter tasks (less trials) that contain more artifacts/noise than adult data. This means that not only are there less ocular artifacts in pediatric data, but the ocular artifacts that are present often overlap with other types of artifacts (e.g., movement), which can make ICs reflecting these artifacts more ambiguous than they are in adult data. Moreover, the skull is less dense in younger participants and not fully fused in infants. This can cause the spatial layout of the blinks and horizontal eye movements to differ slightly from adult data (e.g., blinks and horizontal eye movements might appear to travel farther into the head or be less symmetric in their spread).

To rectify this, we have made modifications to the ADJUST algorithm in an attempt to optimize component classification for pediatric data that were collected with geodesic nets. More specifically, we modified the ocular artifact selection algorithms to be more robust to the increased noise observed in pediatric data by only using the spatial features and changing the spatial layout. Moreover, to balance this more stringent approach and reduce the overall false alarm rate, we added an algorithm that ensures selected artifacts do not also include evidence of neural activity, such as a local maximum within the alpha band. Importantly, this algorithm takes into consideration the specific features of pediatric data. Because alpha tends to be lower in infants and young children (Marshall, Bar-Haim, & Fox, 2002), the algorithm considers a larger range of frequencies to identify alpha peaks (5–15 Hz).

In order to examine the performance of our updated algorithm, which we refer to as "adjusted-ADJUST", we tested the classification accuracy of our adjusted-ADJUST algorithm, as well as the *original* ADJUST algorithm ("original-ADJUST") and ICLabel algorithm. We further examined EEG data quality following the application of each of these algorithms to pediatric and adult EEG data sets. Specifically, we tested these algorithms by measuring their respective classification agreement with expert coders, the number of trials retained following artifact removal, and the reliability of the EEG signal after preprocessing with each algorithm. We predicted that component classifications using the adjusted-ADJUST algorithm would better match manual classifications by experienced coders as compared to the original-ADJUST algorithm and ICLabel for pediatric data, whereas differences would likely not emerge for adult data. Moreover, we expected that for pediatric but not adult data, the adjusted-ADJUST algorithm (as compared to the original-ADJUST, ICLabel, and no ICA correction at all) would also retain more trials and yield greater internal consistency reliability.

2. Method

2.1 Optimization Data Set

Because the adjusted-ADJUST algorithm uses various thresholds for ocular artifact and alpha peak detection, an independent dataset was utilized to determine thresholds that maximized the hit rate while minimizing the false alarm rate to one of the expert coders. This dataset consisted of 96 children (M = 5.8 years, SD = 1.5 years, range 3 to 9 years) who participated in a larger study examining environmental influences on child health outcomes. Data were collected at two sites in South Dakota. Before participating in the study, all parents provided written informed consent. All participants completed one, six-minute resting state period, an auditory oddball task, and a go/no-go task. Participants aged five or older completed a flanker task in addition to the previous three tasks. These different tasks were combined together into one large data file and preprocessed with the methods described below. For more information on this study, see Dukes et al. (2014).

2.2 Test Data Sets

2.2.1 Data Set 1: The adjusted-ADJUST algorithm was then tested on an existing dataset that contained both child and adult data so that the algorithm's performance on the two populations could be compared. This test data set consisted of 10 children (4 males, M = 7.9 years, SD = 0.8 years, range 7.2 to 9.75 years) and 10 adults (3 males, M = 21.6 years, SD = 0.6 years, range 20.5 to 22.6) who participated in a larger study examining brain activity associated with action observation and action execution. Data were collected at The University of Maryland and its Institutional Review Board approved of all the study protocols. Before participating in the study, all adult participants gave written informed consent while all child participants and their parents gave written informed assent and consent, respectively.

Participants performed an action execution and action observation task. In this task, participants watched three different videos with a jittered onset time. Each trial started with a fixation cross that was present for 500 ms to 1500 ms, followed by one of the three videos

containing two boxes. The three videos corresponded to one of three conditions: action observation, action execution, and scene observation (control). In the action observation condition, the participant watched a hand grasp one of the boxes. For the action execution condition, the participant executed a grasping action while watching a similar video containing boxes. Finally, the scene observation condition contained similar background stimuli as the other two conditions, but it did not depict a grasping action, nor did it require the participant to execute a grasping action. Because the stimuli across conditions were almost identical, we expected no differences across conditions in early visual ERP components and there were no differences in mean P1 ERP amplitude between any of the three video conditions, $F(2,36) = 0.102$, $p = .903$. As such, we did not consider video condition as a factor in any of the later analyses. There were 40 trials in each condition, with a total of 120 trials overall. More details on the task can be found in Morales, Bowman, Velnoskey, Fox, and Redcay (2019).

2.2.2 Data Set 2: The adjusted-ADJUST algorithm was also tested on an existing infant dataset so that the algorithm's performance on infant data could be assessed. This test data set consisted of 10 infants (5 males, $M = 5.2$ months, $SD = 0.6$ months, range 4.2 to 6 months) who participated in a larger study examining temperament and brain development. Data were collected at The University of Maryland and its Institutional Review Board approved of all the study protocols. Before participating in the study, all participant parents provided written informed consent.

Participants completed a resting state task. In order to keep the infants' attention and prevent excessive motor movement, an experimenter showed the infants different colored balls and turned a bingo wheel. The resting state task was divided into six, 30-second blocks. This resulted in a total of 180, one-second segments for analysis. For more details about the experimental procedure see Fox, Henderson, Rubin, Calkins, and Schmidt (2001).

2.3 EEG Preprocessing

Continuous EEG data were collected for the optimization data set using a 64-channel HydroCel Geodesic Sensor Net and EGI Netstation software (version 5; Electrical Geodesic Inc., Eugene, OR) and the test data sets using a 128-channel HydroCel Geodesic Sensor Net and EGI Netstation software (version 4; Electrical Geodesic Inc., Eugene, OR). Consistent with recommended acquisition protocols for high-impedance EEG systems (Electrical Geodesic Inc., Eugene, OR), the target impedance level for the electrodes was below 50kΩ during data collection. The EEG signal was amplified through an EGI NetAmps 300 amplifier with a sampling rate of 500 Hz and referenced online to the vertex electrode (Cz). Data were then preprocessed using the EEGLAB toolbox (Delorme & Makeig, 2004) and custom MATLAB scripts (The MathWorks, Natick, MA). Because the infant data files in the second test data set were shorter (ranging from four to eight minutes in length), the number of electrodes was reduced from 128 to 64 in order to ensure a good ICA decomposition for all infants in the second test data set. Channels were deleted in the infant data set such that the layout matched that of a 64-channel HydroCel Geodesic Sensor net. This reduction in total electrodes for the infant data set arose as a solution to overcome a limitation of ICA with short recordings (Gabard-Durnam, Mendez Leal, Wilkinson, & Levin, 2018). The data

were high-pass filtered at 0.3 Hz and low-pass filtered at 49 Hz. Bad channels were identified and removed globally using the FASTER toolbox (Nolan, Whelan, & Reilly, 2010); On average, 4.4 channels were removed by FASTER, with the number of removed channels ranging from one to eight. Next, the data were copied and put through a 1.0 Hz high-pass filter before being epoched into arbitrary one-second segments. Any epochs where channel voltage exceeded ±1000 μV or power within the 20–40 Hz band (after Fourier analysis) exceeded −100 dB to +30 dB were deleted. Alternatively, if a particular channel caused more than 20% of the epochs to be marked for rejection, that channel was rejected instead of the epochs. After running ICA on this copied dataset, the ICA weights were subsequently applied back to the original dataset (see Debener, Thorne, Schneider, & Viola, 2010, for further details on this approach) and any components that were marked for rejection, using either adjusted-ADJUST (which is the focus of the ensuing methods sections), original-ADJUST, or ICLabel, were removed. Data were then epoched into 1500 millisecond segments that started 500 milliseconds before the video onset. After ICA artifact removal and epoching, a two-step procedure for identifying residual artifacts was employed. First, any epochs where ocular channel (E1, E8, E14, E21, E25, or E32 for dataset 1; E1, E9, E22, or E32 for dataset 2) voltages exceeded +/−125 μV, indicating the presence of residual ocular activity not removed through ICA, were rejected. Second, for any epoch in which nonocular channel voltages exceeded +/−125 μV, these channels were interpolated at the epoch level, unless greater than 10% of the channels exceeded +/−125 μV, in which case the epoch was rejected instead. Finally, all missing channels were interpolated using a spherical spline interpolation and then the data were referenced to the average of all the electrodes. The average number of interpolated channels per epoch for the first (128-channel) data set (including those globally rejected) was 4.8 for adjusted-ADJUST, 4.0 for original-ADJUST, 4.6 for ICLabel, and 3.0 when not using any form of ICA correction. For the second (64-channel) data set, the average number of interpolated channels per epoch (including those globally rejected) was 1.3 for adjusted-ADJUST, 1.1 for original-ADJUST, 1.2 for ICLabel, and 1.0 when not using any form of ICA correction. The number of interpolated channels per epoch for the first (128-channel) data set ranged from one to 19 for both adjusted-ADJUST and ICLabel, one to 18 for original-ADJUST, and one to 20 when not using any form of ICA correction. For the second (64-channel) data set, the number of interpolated channels per epoch ranged from one to 12 for adjusted-ADJUST and ICLabel and one to 11 for original-ADJUST and no form of ICA correction.

2.4 Changes to the ADJUST Algorithm

Similar to Mognon and colleagues (2011), normalized component topographies were used when computing spatial information. The original-ADJUST classified artifacts into four different categories: eye blinks, horizontal eye movements, vertical eye movements, and discontinuities (e.g., pop-offs). Various functions using temporal or spatial information were used to calculate the likelihood that a component belonged to one of the four artifact classes. In order to improve ADJUST performance on pediatric data, changes were made to the eye blink detection and horizontal eye movement detection functions. The vertical eye movement and generic discontinuity detection functions were not altered because anecdotal evidence suggested that these functions already showed high performance on adult and pediatric data collected with geodesic nets. Additionally, new lines of code were added that

check for local maxima within the alpha band for all components that were classified as artifacts. Adjusted-ADJUST is freely available and can be accessed at: https://github.com/ChildDevLab/MADE-EEG-preprocessing-pipeline/tree/master/adjusted_adjust_scripts.

2.2.1 Eye Blink Changes.—Original ADJUST function classified blinks by using two spatial measures (spatial average difference and spatial variance difference) and a temporal measure (kurtosis) of the components. The spatial average difference was calculated by subtracting the average activity at posterior sites from the average activity at anterior sites. Similarly, the spatial variance difference was calculated by subtracting the variance of activity at posterior sites from the variance of activity at anterior sites. Kurtosis was calculated within each epoch and then averaged across epochs to get a single kurtosis value for each component. In order to improve algorithm performance on geodesic nets, the spatial functions were changed to better classify blink artifacts without introducing too many false alarms. The temporal measure was removed because it did not catch any blinks above and beyond the altered spatial functions. Furthermore, because pediatric populations have lower blink rates, the temporal function may miss blink components in participants with especially low blink rates.

The new spatial layout functions used the “zscore” function in MATLAB to standardize component activity at each electrode within a given component. Next, the average z-score was computed for five different electrode clusters (Figure 1): left eye region, right eye region, center eye region, central region, and posterior region. Assignment of electrodes to specific clusters were determined based on their location in polar coordinates (radius, theta). As shown in Figure 1, to be in any of the eye clusters, the radius had to be between 0.45 and 0.60. The left eye cluster required a theta value between −60 and 0 degrees whereas the right eye cluster required a theta value between 0 and 60 degrees. The center eye cluster was restricted to theta values between −20 and 20 degrees. To be in the central cluster, the radius had to be less than 0.45 and the absolute value for theta had to be between 35 and 109 degrees. The posterior cluster required a radius less than 0.55 and an absolute theta value greater than 109 degrees.

In order for a component to be marked as a blink artifact it had to meet three conditions:

1. At least one of the three eye regions (left, right, or center) had to have an average z-score with an absolute value greater than two.
2. The absolute value of the average z-score of the central and posterior regions had to be less than one.
3. The variance of the z-scores in the posterior region had to be less than 0.15 and the variance of the z-scores in the left and right eye regions combined OR the variance of the absolute values of the z-scores in the posterior region had to be less than 0.075 and the variance of the z-scores in the left and right eye regions combined.

Because blinks cause such huge deflections in EEG data at frontal sites (i.e., high activity), the average z-score in components reflecting blinks should well exceed two standard deviations from the mean. Ambiguous components that might reflect neural data, on the

other hand, are far less likely to exceed two standard deviations from the mean. For similar reasons, the activity in the back of the head had to be within one standard deviation of the mean and have low variance because blinks should have low activity at posterior sites. Having greater activity at posterior sites would likely indicate the component contains meaningful neural activity.

The new blink detection function was further modified, such that after selecting potential blink components via the three conditions outlined above, secondary checks were added to determine whether blinks had been incorrectly identified based on component spatial spread (from either of the three eye regions towards the central region). Any electrodes that were part of the central eye region or part of the outermost ring of channels were removed before considering spread. Because artifacts from blinks propagate back from the eyes, the algorithm was set up to consider whether or not component activity decreased along the rostral-to-caudal axis. As long as z-scores were less than 2.5 for electrodes with a radius between .35 and .45 and less than two for electrodes with a radius between .25 and .35, components were not un-marked as blinks. Thresholds for the blink detector were determined on the independent optimization dataset.

2.2.2 Horizontal Eye Movement Changes.—Original-ADJUST also classified horizontal eye movements by using a temporal measure (maximum epoch variance) and a spatial measure (spatial eye difference). The spatial eye difference was calculated by taking the absolute value of the difference in activity between the left and right eyes. Maximum epoch variance was calculated by dividing the greatest epoch variance by the average variance across all epochs. Similar to the blink detection code, the temporal function was removed and the spatial function was altered, the latter for the same reason given in the section entitled “eye blink changes” and the former due to increased noise in pediatric data, which may cause epoch variance from saccades to not reach the required threshold for rejection.

This time, average z-scores were computed for four different electrode clusters: left eye region, right eye region, central region, and posterior region. Electrode assignment to clusters was determined based on the same polar coordinate restrictions that were used for blinks (Figure 1). However, to focus on the lateral electrodes, the left and right eye regions were altered slightly by changing the theta range to be from +/−35 to +/−62 and the radius to be greater than 0.5. In order for a component to be marked as a horizontal eye movement artifact it had to meet three conditions:

1. Either the left or the right eye region had to have an average z-score with an absolute value greater than two
2. The absolute value of the average z-score of the central and posterior regions had to be less than one
3. The variance of the z-scores in the posterior region had to be less than 0.15 and the variance across the left and right eye regions combined had to be greater than 2.5 OR the variance of the absolute values of the z-scores in the posterior region

had to be less than 0.075 and the variance across the left and right eye regions combined had to be greater than 2.5

As with blinks, horizontal eye movements cause deflections in EEG data at frontal sites (i.e., high activity) compared to posterior sites, so the first two conditions are the same as described above in the section entitled “eye blink changes”. In the third condition, because saccade components should exhibit opposing polarity along the right-left axis near the eyes (Mognon et al., 2011), the variance of the left and right eye regions had to be greater than 2.5 standard deviations in order for a component to be marked as a saccade. Similar to the new blink detection function, secondary checks were added to determine whether saccades had been incorrectly identified based on component spatial spread (from the ocular region towards the central region). These secondary checks, based on component spread, were identical to those described above for blinks. Thresholds for the horizontal eye movement detector were determined on the independent optimization dataset.

2.2.3 Alpha Peak Detector.—

Many components reflecting neural activity contain local maxima within the alpha band, which we refer to here, for simplicity, as an “alpha peak”. Even though neural activity exists in other power bands (i.e., delta, theta, beta, and gamma), high activity levels in those power bands do not tend to manifest in the EEG power spectrum as a clear peak. As a result, finding high levels of neural activity in power bands outside of the alpha band is not practical; therefore, we chose to focus exclusively on the alpha band.

The ADJUST algorithm was further modified to check for the presence of alpha peaks within any of the potential ICA artifact components. This “alpha peak detector” loops through all of the components marked as artifacts (to be removed from the data) and de-selects any components containing alpha peaks. Code from the MARA toolbox (Winkler, Haufe, & Tangermann, 2011) was adapted to calculate the power spectrum for components using the “spectopo” function from EEGLAB. Because the EEG power spectrum tends to follow a 1/f curve, the alpha peak detector uses the standardized residual values of the component’s power spectrum after regressing out the estimated 1/f curve. The detector then looks for any local maxima between 5 Hz and 15 Hz. A larger range for the alpha band was selected due to the fact that alpha tends to be lower in children as compared to adults (Marshall et al., 2002) and there are significant individual differences in alpha peaks (Corcoran, Alday, Schlesewsky, & Bornkessel-Schlesewsky, 2018). The detector de-selects components if the local maxima has a prominence greater than 0.3 μV^2/Hz and a width greater than 0.9 Hz; these thresholds were selected to match one of the expert coders on the independent optimization dataset. More specifically, components from an independent dataset were visually inspected in order to determine which thresholds correctly identified the majority of alpha peaks without an appreciable change in the false alarm rate.

2.3 Evaluating adjusted-ADJUST

Performance of the adjusted-ADJUST algorithm was compared to performance of two other algorithms: the original ADJUST algorithm and ICLabel. Performance was measured in three ways: percent agreement with expert coders, number of trials retained after artifact rejection, and reliability (i.e., internal consistency) of the EEG signal. Because the assumption of normality was violated for most variables and the assumption of homogeneity

of variances was violated for most comparisons, rank-based nonparametric tests were conducted using the nparLD package (Noguchi, Gel, Brunner, & Konietschke, 2012) in R (v. 3.5.1; R Core Team, 2015). These tests are a nonparametric alternative to the traditional repeated measures ANOVA and they provide a Wald-Type Statistic (WTS) to examine statistical significance. Significant interactions were further probed with planned contrasts using Wilcoxon signed rank tests. In order to control for the false positive rate, a Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), with a 0.05 false positive discovery rate, was applied to all follow-up comparisons.

2.3.1 Percent Agreement with Expert Coders: As in Mognon and colleagues (2011), three coders with expertise in ICA components derived from EEG classified 3084 components as artifactual or neural. In order for a component to be classified as an artifact, two of the three coders had to classify it as an artifact. Percent agreement scores were then calculated by comparing the artifact selections from each algorithm to the selections of the three expert coders. Additionally, following Mognon and colleagues (2011), the percent agreement scores were weighted based on how much variance was accounted for by each component. Thus, the percent agreement score (PA) was calculated by dividing the variance accounted for by the components that the algorithm and expert coders agreed on (ICs_{Agree}) by the total variance (TV).

$$PA = \Sigma\left(ICs_{Agree}\right) / TV$$

The variance accounted for by each component was calculated using the EEGLAB function “eeg_pvaf”. Because copying the ICA weights from the “copied dataset” back to the original dataset can cause the ICs to become spatially nonorthogonal, it leads to some of the variance accounted for by some ICs to be negative. Thus, the variance accounted for was calculated by using the “eeg_pvaf” function on the “copied dataset.” It is important to note that we observe a similar pattern of results regardless of which dataset the percent agreement is calculated on (copied or original) and this change does not affect their interpretation. Furthermore, the subsequent analyses (trials retained and internal consistency reliability) also show the same pattern of results regardless of which dataset the automated artifact selection algorithms were run on. As an added check, the percent agreement scores were also calculated without considering the variance accounted for by each component and these results also followed a similar pattern. Therefore, the weighted percent agreement scores were used because it is more important for the algorithm to correctly classify components that account for the greatest amount of variance in the data. Next, a 3 (algorithm) × 2 (age group) nonparametric test was conducted for the adult and child data using the weighted percent agreement scores. For the infant data, a three-way repeated measures nonparametric test comparing the different algorithms was conducted using the weighted percent agreement scores. Follow-up pairwise comparisons were conducted using a Wilcoxon signed rank test. All post hoc comparisons were FDR corrected (Benjamini & Hochberg, 1995).

2.3.2 Trials Retained After Removing Artifacts: In order to determine which algorithm not only better matched experts, but also retained more artifact-free data, we

compared the results of preprocessing using the adjusted-ADJUST algorithm to the original-ADJUST algorithm, ICLabel, and a preprocessing stream that employed no ICA correction prior to rejecting epochs with ocular activity or excessive noise. This last comparison evaluates the benefits (or disadvantages) of using ICA correction in preprocessing EEG data given that many studies do not employ ICA for artifact correction. The maximum number of epochs that participants could have was 120 (based on the total number of task trials) for dataset 1 (child and adult data) and 180 (based on the total number of one-second segments) for dataset 2 (infant data). After running each algorithm, or not running an algorithm in the case of the no ICA correction condition, the data went through a final artifact rejection step as detailed in the EEG preprocessing section. The proportion of trials remaining after artifact rejection for each condition for children and adults was then used in a 4 (algorithm) × 2 (age group) repeated measures nonparametric test. For infant data, a three-way repeated measures nonparametric test comparing the different conditions was conducted using the proportion of trials remaining after artifact rejection. Follow-up pairwise comparisons were conducted using a Wilcoxon signed rank test. All post hoc comparisons were FDR corrected (Benjamini & Hochberg, 1995).

2.3.3 Internal Consistency Reliability: Generally, retaining more trials benefits ERP or time frequency analyses by increasing the signal to noise ratio. However, if the trials retained after ICA correction still contain a significant amount of noise, they may decrease the overall data quality rather than increase it. To ensure that including more trials did not reduce the quality of data, the internal consistency reliability was calculated after data correction with adjusted-ADJUST, original-ADJUST, ICLabel, or no ICA correction condition. For adult and child data, we estimated the internal consistency reliability of the mean amplitude of the P1 visual ERP; while, for infant data, we calculated the internal consistency reliability of relative alpha power (6–9 Hz band). The internal consistency reliability estimates were measured via the Spearman-Brown split-half correlation method. Because different methods are expected to differ in the number of artifact-free trials and reliability estimates differ as a function of recording length (i.e., numbers of trials), reliability estimates were obtained for increasing number of trials. This method allowed us to compare the different algorithms while maintaining a constant trial number. Moreover, it allowed us to determine the amount of data necessary to obtain good and excellent reliability. The artifact correction/rejection method that maximizes the signal while reducing noise should be more internally consistent and yield higher reliability estimates even with relatively low amounts of data.

Specifically, internal consistency reliability estimates were obtained using Spearman–Brown-corrected split-half correlation coefficients for an increasing number of trials. From two through nine trials, these correlation coefficients were calculated in steps of one trial; from 10 through 120 trials correlation coefficients were calculated in steps of five trials (i.e., 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, etc.). Following the method proposed by Towers and Allen (2009), for each number of trials, n, 10,000 iterations of split-half correlations were calculated. For each iteration, n trials were randomly selected from all available trials for that participant. The selected trials were then halved by randomly assigning trials/epochs to one of two groups/halves. The mean P1 amplitude or relative alpha power values were then

calculated separately for each half. The Pearson correlation coefficient was obtained using each half across all participants with enough usable trials for any given *n*. Finally, the Spearman-Brown prophecy formula was applied to correct the reliability estimate for test length (i.e. estimate the reliability when the number of trials is doubled given that reliability estimate was calculated with half the original number of trials). This process generated 10,000 reliability estimates for a given *n* for each condition. Notably, when iterating across number of trials (from two to 120), the number of participants included in the split-half reliability estimates decreased as the number of trials increased. Only reliability estimates with a minimum of six participants (the majority of each group) are presented.

For the reliability analyses, whether a method achieved good and excellent reliability for the average reliability coefficients was quantified. Moreover, to provide a measure of the resampling distributions, we present the percentage of iterations that meet or exceed good (.8) and excellent (.9) reliability for each condition. Finally, to quantify differences between conditions, the area under the curve was calculated for each iteration. The average and 95% confidence intervals (CIs) from the resampling distribution were estimated for each condition, providing a measure of overall reliability across increasing numbers of trials – with a greater area under the curve indicating higher internal consistency reliability.

3. Results

3.1 Agreement with Expert Coders

3.1.1 Adult and Child Data: The results of the 3 (algorithm) × 2 (age group) nonparametric test for the percent agreement scores based on all components can be found in Table 1. As expected, the 3 (algorithm) × 2 (age group) nonparametric test showed an interaction between algorithm and age group, $WTS(2) = 7.67$, $p = .022$. Based on *a priori* hypotheses, planned comparisons were conducted and descriptive information can be found in Table 2. These comparisons revealed that, for children, adjusted-ADJUST performed significantly better than both original-ADJUST, $Z = 2.80$, $p = .005$, $q = .015$, and ICLabel, $Z = 2.29$, $p = .022$, $q = .026$. In adults, adjusted-ADJUST performed better than original-ADJUST, $Z = 2.67$, $p = .008$, $q = .016$, but only marginally significantly better than ICLabel, $Z = 1.78$, $p = .074$, $q = .074$. Comparing ICLabel to original-ADJUST revealed significant differences for both children, $Z = 2.45$, $p = .013$, $q = .020$, and adults, $Z = 2.80$, $p = .005$, $q = .015$. As summarized in Figure 2, the adjusted-ADJUST algorithm agreed more with expert coders than both original-ADJUST and ICLabel, for both children and adults. However, as demonstrated by the significant interaction and shown in Figure 2, this difference was larger in children compared to adults, especially for ICLabel.

3.1.2 Infant Data: The nonparametric test with infant data showed a marginally significant main effect of algorithm, $WTS(2) = 4.74$, $p = .093$. Based on *a priori* hypotheses, planned comparisons were conducted and descriptive information can be found in Table 3. These comparisons revealed that adjusted-ADJUST performed significantly better than original-ADJUST, albeit this difference did not survive multiple comparisons correction, $Z = 1.98$, $p = .047$, $q = .141$. Moreover, adjusted-ADJUST did not significantly differ from ICLabel, $Z = 1.38$, $p = .169$, $q = .254$. Furthermore, comparing ICLabel to original-ADJUST

revealed no significant differences, $Z = 0.561$, $p = .575$, $q = .575$. As summarized in Figure 3, adjusted-ADJUST agreed more with expert coders than original-ADJUST, but did not significantly differ from ICLabel or original-ADJUST (after correction for multiple comparisons) for the infant data.

3.2 Trials Retained

3.2.1 Adult and Child Data: The results of the 4 (algorithm) × 2 (age group) nonparametric test for trials retained can be found in Table 4 and descriptive information can be found in Table 5. As expected, there was a significant interaction between algorithm and age group, $WTS(3) = 53.95$, $p < .001$, suggesting that the amount of trials retained after using each correction algorithm differed by age group (Figure 4). As a result, planned comparisons were conducted separately for children and adults between adjusted-ADJUST and the other three algorithm conditions (original-ADJUST, ICLabel, and no ICA correction). These comparisons revealed that adjusted-ADJUST retained significantly more trials than no ICA correction and original-ADJUST for both adults, $Z = 2.80$, $p = .005$, $q = .015$; $Z = 2.52$, $p = .012$, $q = .024$, and children, $Z = 2.81$, $p = .005$, $q = .015$; $Z = 2.24$, $p = .025$, $q = .038$. There were no significant differences between adjusted-ADJUST and ICLabel for children, $Z = 0.95$, $p = .343$, $q = .343$, or adults, $Z = 1.40$, $p = .161$, $q = .193$. These results suggested that, for both children and adults, adjusted-ADJUST retained more trials than both original-ADJUST and no ICA correction, but not ICLabel. However, as shown in Figure 4, the significant interaction suggests that this difference between adjusted-ADJUST and no ICA correction is larger in adults compared to children, while the difference between adjusted- ADJUST and original-ADJUST shows the reverse effect (i.e., the difference is larger in children compared to adults).

3.2.2 Infant Data: The nonparametric test on trials retained showed a significant main effect of algorithm, $WTS(3) = 45.41$, $p < .001$. As such, planned comparisons were conducted and descriptive information can be found in Table 6. The adjusted-ADJUST algorithm retained significantly more trials than no ICA correction, $Z = 2.81$, $p = .005$, $q = .015$, and original-ADJUST, $Z = 2.55$, $p = .011$, $q = .017$, and marginally significantly more trials than ICLabel, $Z = 1.73$, $p = .083$, $q = .083$. These results suggested that, for infants, adjusted-ADJUST retained more trials than no ICA correction, original-ADJUST, and marginally more trials ICLabel (Figure 5).

3.3 Reliability

Overall, reliability estimates increased with the number of trials. However, this was not always the case. In particular, the no ICA correction condition sometimes exhibited a decline in reliability estimates as the number of trials increased. This decline in reliability is likely the result of significantly less trials being retained (see above), and therefore, less subjects being available for computing reliability estimates at higher trial counts.

For adults, as shown in Figure 6A, all ICA-correction procedures reached a mean of excellent reliability. However, as demonstrated in Figure 6B, when examining the resampling distributions, all ICA-correction procedures reached good reliability (> .8) with a 95% CI, but no method reached excellent (> .9) reliability with a 95% CI. In line with these

results, when examining the area under the curve of the mean reliability curves (Figure 6C), the 95% CIs of adjusted-ADJUST overlapped with the 95% CIs of the two other ICA-correction methods. However, the CIs for adjusted-ADJUST, original-ADJUST, and ICLabel did not overlap with the CIs for no ICA correction, suggesting significantly higher internal consistency reliability for the three correction algorithms on adult data as compared to no ICA correction.

For children, as shown in Figure 6D, only adjusted-ADJUST and ICLabel reached an average of excellent reliability, whereas original-ADJUST only reached good reliability. As demonstrated in Figure 6E, when examining the resampling distributions, adjusted-ADJUST and ICLabel reached good reliability with a 95% CI, while original-ADJUST and no ICA correction did not. Moreover, no method reached excellent reliability with a 95% CI. In line with these results, when examining the area under the curve of the mean reliability curves (Figure 6F), the 95% CIs of adjusted-ADJUST overlapped with the 95% CIs of ICLabel. However, the CIs for adjusted-ADJUST and ICLabel did not overlap the CIs for original-ADJUST and no ICA correction, suggesting greater internal consistency reliability for adjusted-ADJUST and ICLabel on child data as compared to original-ADJUST and no ICA correction.

For infants, as shown in Figure 6G, all methods reached a mean of excellent reliability. As demonstrated in Figure 6H, when examining the resampling distributions, all methods reached good reliability with a 95% CI; however, only adjusted-ADJUST reached excellent reliability with a 95% CI. In line with these results, when examining the area under the curve of the mean reliability curves (Figure 6I), the 95% CIs of adjusted-ADJUST did not overlap the 95% CIs of ICLabel, original-ADJUST, and no ICA correction, suggesting greater internal consistency reliability for adjusted-ADJUST on infant data over all other methods tested.

4. Discussion

The goal of the present study was to modify an existing artifact selection algorithm to objectively select artifactual independent components (ICs) derived from pediatric EEG recorded on geodesic nets. Towards this end, the *original* ADJUST algorithm (Mognon et al., 2011) was optimized to detect ocular artifacts using only spatial information and to de-select components that may include neural activity (e.g., alpha peaks). Our results suggest that the adjusted-ADJUST algorithm performs comparably to or better than original-ADJUST and ICLabel with adult data, but outperformed these existing ICs selection algorithms on several measures when applied to pediatric data. These results suggest that the changes to the ADJUST algorithm improve its performance, especially for pediatric data, and using the adjusted-ADJUST algorithm could facilitate EEG studies with infants and children.

Compared to original-ADJUST, the adjusted-ADJUST algorithm had higher percent agreement scores with expert coders for all age groups and retained more trials for children and adults. Importantly, for children and infants, adjusted-ADJUST also yielded a more reliable signal than original-ADJUST. Furthermore, when compared to a novel algorithm,

ICLabel, adjusted-ADJUST better matched expert coders on child and adult data. Although adjusted-ADJUST did not retain significantly more trials than ICLabel on child and adult data, adjusted-ADJUST did retain more trials than ICLabel on infant data – albeit, this difference was marginally significant. Finally, while adjusted-ADJUST and ICLabel generated similar reliability estimates for children and adults, adjusted-ADJUST yielded significantly greater internal reliably estimates than ICLabel for infants. This difference with infant data is in line with anecdotal evidence, which suggests that ICLabel may have subpar performance in populations not included in the ICLabel training dataset, such as infants (Pion-Tonachini et al., 2019). However, ICLabel did yield infant data with good reliability (> .8) and a comparable number of trials.

Comparing the correction algorithms to the no ICA correction condition suggests that it is better to use some form of ICA correction on EEG data prior to artifact rejection. When assessing trials retained and internal consistency reliability, the correction algorithms generally outperformed the no correction condition by retaining significantly more trials or yielding a higher reliability estimate. This implies that, when trial counts are low, EEG studies would benefit from using ICA-correction methods rather than just deleting trials with artifacts. Moreover, this is not to say that all three correction algorithms are equally beneficial. Performance still varied across the three automated algorithms as some algorithms retained more trials or had greater internal consistency reliability than others.

After considering all three performance measures, it appears that both adjusted-ADJUST and ICLabel could be useful automated algorithms for classifying artifactual components derived from EEG data. However, both algorithms have advantages and limitations that should be considered before implementation. For example, while adjusted-ADJUST and ICLabel performed similarly for adults in terms of trials retained and reliability achieved, adjusted-ADJUST outperformed ICLabel in terms of reliability achieved when applied to infant data. Thus, the adjusted-ADJUST algorithm may currently represent a better method for identifying ICA artifacts in pediatric data. Relatedly, ICLabel uses machine learning to classify components as artifactual or neural, but the training data set that ICLabel uses contains mostly adult data (Pion-Tonachini et al., 2019). To optimize ICLabel for pediatric data, a greater proportion of pediatric data may need to be included in its training dataset. It is also worth noting that some users may prefer an approach that does not use machine learning, as this approach can allow users to see how data is being classified as artifactual or neural. Moreover, the specific parameters employed in the adjusted-ADJUST algorithm can be modified by users as needed, to be more or less conservative, depending on the population and the preferences of the researchers. For example, if a researcher finds that adjusted-ADJUST is too conservative on eye blink detection, the z-score value for the frontal eye regions can be modified to match the desired level of blink-identification. Similarly, the thresholds for identifying potential neural activity within a component (i.e., the presence of alpha-peaks) can be adjusted to meet the preference of the researcher. Specifically, thresholds corresponding to the prominence and width of a potential alpha peak, as well as the range of frequencies that are inspected for potential alpha peaks, can all be modified by the user. In summary, if researchers are studying pediatric populations, want to know exactly how their ICA data is being classified as artifactual, or modify their artifact

detection thresholds, the adjusted-ADJUST algorithm may currently provide a better solution for users.

One limitation of the current study is that the expert coders used in this study were from the same lab where the algorithm was developed. Consequently, the expert coders are likely biased in favor of the criteria used for the adjusted-ADJUST algorithm (e.g., the default thresholds used to detect alpha peaks), which could have influenced the measure of agreement of with expert coders. For this reason, the trials retained and reliability analyses were included to provide more objective assessments of algorithm performance. Moreover, only one of the three coders optimized the adjusted-ADJUST algorithm and removing this coder does not change the pattern found in the percent agreement results. In fact, removing this coder makes the results slightly stronger, such that marginally significant comparisons become significant (see Online Supplement). Importantly, threshold optimization for the algorithm was performed in an independent dataset, aiding the generalizability of the method.

Similarly, other sensitivity analyses show that the percent agreement results remain the same regardless of whether the automated algorithms were applied to the “copied dataset” or the original dataset (see Online Supplement). Even though the original dataset contains more information than the copied dataset, the component information looks almost identical with only minor differences in the power between the original and copied dataset. However, the variance accounted for by each component differs between the two datasets. Because the original dataset contains information that was not present during ICA decomposition, the components become nonorthogonal and the variance accounted for by some components (as calculated with the “eeg_pvaf” function) falls below zero. Therefore, the variance accounted for by each component is less meaningful after copying the weights back to the original dataset. As such, we calculated the weighted percent agreement scores using the copied dataset to calculate the variance accounted for by each component. Furthermore, additional analyses were done using raw percent agreement scores (i.e., not considering the variance accounted for by each component), and these also showed a similar pattern to the weighted percent agreement scores (see Online Supplement).

Another limitation is that, although expert coders classified over 3000 components, the sample consisted of only 10 adults, 10 children, and 10 infants, which likely resulted in some of the analyses being underpowered. Future work could test the adjusted-ADJUST algorithm in larger samples. Finally, the adult and child data were generated from an event-related task, while the infant data were generated from a passive viewing task. These two types of tasks (passive versus active) can elicit different types of artifacts due to the different natures of the tasks in addition to the participant age.

In conclusion, the adjusted-ADJUST algorithm performs similarly to the classification of components by expert coders and retains more trials without compromising reliability. Both pediatric and adult data showed higher percent agreement scores with the adjusted-ADJUST algorithm, as compared to original-ADJUST, while children and adults also showed higher percent agreement scores with adjusted-ADJUST as compared to ICLabel. Furthermore, adjusted-ADJUST retained significantly more trials than no ICA correction for both the

pediatric and adult data in addition to retaining more trials than original-ADJUST for both children and adults. The reliability analyses indicated comparable performance across algorithms for adults, better performance for adjusted-ADJUST than original-ADJUST for children and infants, and better performance for adjusted-ADJUST than ICLabel with infants. Thus, the adjusted-ADJUST algorithm is a valuable tool that can assist in the analysis of pediatric EEG data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We thank the many research assistants involved in data collection and the participating families without whom the study would not have been possible.

Author Notes

Grant Information: This work was supported by the National Institutes of Health (P01HD064653 and U01MH093349 to NAF and UH3 OD023279 to Amy J. Elliott).

References

Benjamini Y, & Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289–300. 10.1111/j.2517-6161.1995.tb02031.x

Chaumon M, Bishop DV, & Busch NA (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. Journal of neuroscience methods, 250, 47–63. 10.1016/j.jneumeth.2015.02.025 [PubMed: 25791012]

Corcoran AW, Alday PM, Schlesewsky M, & Bornkessel-Schlesewsky I (2018). Toward a reliable, automated method of individual alpha frequency (IAF) quantification. Psychophysiology, 55(7), 313064 10.1111/psyp.13064

Debener S, Thorne J, Schneider TR, & Viola FC (2010). Using ICA for the analysis of multi-channel EEG data In Ullsperger M & Debener S (Eds.), Simultaneous EEG and fMRI (pp. 121–135). New York, NY: Oxford University Press 10.1093/acprof:oso/9780195372731.003.0008

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. 10.1016/j.jneumeth.2003.10.009 [PubMed: 15102499]

Dukes KA, Burd L, Elliott AJ, Fifer WP, Folkerth RD, Hankins GD, … & Signore C (2014). The Safe Passage Study: Design, Methods, Recruitment, and Follow-Up Approach. Paediatric and perinatal epidemiology, 28(5), 455–465. 10.1111/ppe.12136 [PubMed: 25131605]

Fox NA, Henderson HA, Rubin KH, Calkins SD, & Schmidt LA (2001). Continuity and Discontinuity of Behavioral Inhibition and Exuberance: Psychophysiological and Behavioral Influences across the First Four Years of Life. Child Development, 72(1), 1–21. 10.1111/1467-8624.00262 [PubMed: 11280472]

Gabard-Durnam LJ, Mendez Leal AS, Wilkinson CL, & Levin AR (2018). The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): standardized processing software for developmental and high-artifact data. Frontiers in neuroscience, 12, 97 10.3389/fnins.2018.00097 [PubMed: 29535597]

Jung TP, Makeig S, Westerfield M, Townsend J, Courchesne E, & Sejnowski TJ (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. Clinical Neurophysiology, 111(10), 1745–1758. 10.1016/S1388-2457(00)00386-2 [PubMed: 11018488]

Lawrenson JG, Murphy PJ, & Esmaeelpour M (2003). The neonatal tear film. Contact Lens and Anterior Eye, 26(4), 197–202. 10.1016/j.clae.2003.09.002 [PubMed: 16303518]

Luck SJ (2014). An introduction to the event-related potential technique. Cambridge, MA: MIT press.

Marshall PJ, Bar-Haim Y, & Fox NA (2002). Development of the EEG from 5 months to 4 years of age. Clinical Neurophysiology, 113(8), 1199–1208. 10.1016/S1388-2457(02)00163-3 [PubMed: 12139998]

Morales S, Bowman LC, Velnoskey KR, Fox NA, & Redcay E (2019). An fMRI study of action observation and action execution in childhood. Developmental cognitive neuroscience, 37, 100655 10.1016/j.dcn.2019.100655 [PubMed: 31102960]

Mognon A, Jovicich J, Bruzzone L, & Buiatti M (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. Psychophysiology, 48(2), 229–240. 10.1111/j.1469-8986.2010.01061.x [PubMed: 20636297]

Noguchi K, Gel YR, Brunner E, & Konietschke F (2012). nparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. Journal of Statistical Software, 50(12), 1–23. 10.18637/jss.v050.i12 [PubMed: 25317082]

Nolan H, Whelan R, & Reilly RB, (2010). FASTER: fully automated statistical EEG artifact rejection. Journal of Neuroscience Methods, 192(1), 152–162. 10.1016/j.jneumeth.2010.07.015 [PubMed: 20654646]

Pion-Tonachini L, Kreutz-Delgado K, & Makeig S (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. NeuroImage, 198, 181–197. 10.1016/j.neuroimage.2019.05.026 [PubMed: 31103785]

Team, R. C. (2013). R: A language and environment for statistical computing.

Towers DN, & Allen JJ (2009). A better estimate of the internal consistency reliability of frontal EEG asymmetry scores. Psychophysiology, 46(1), 132–142. 10.1111/j.1469-8986.2008.00759.x [PubMed: 19055503]

Urigüen JA, & Garcia-Zapirain B (2015). EEG artifact removal—state-of-the-art and guidelines. Journal of neural engineering, 12(3), 031001 10.1088/1741-2560/12/3/031001 [PubMed: 25834104]

Winkler I, Haufe S, & Tangermann M (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. Behavioral and Brain Functions, 7(1), 30 10.1186/1744-9081-7-30 [PubMed: 21810266]

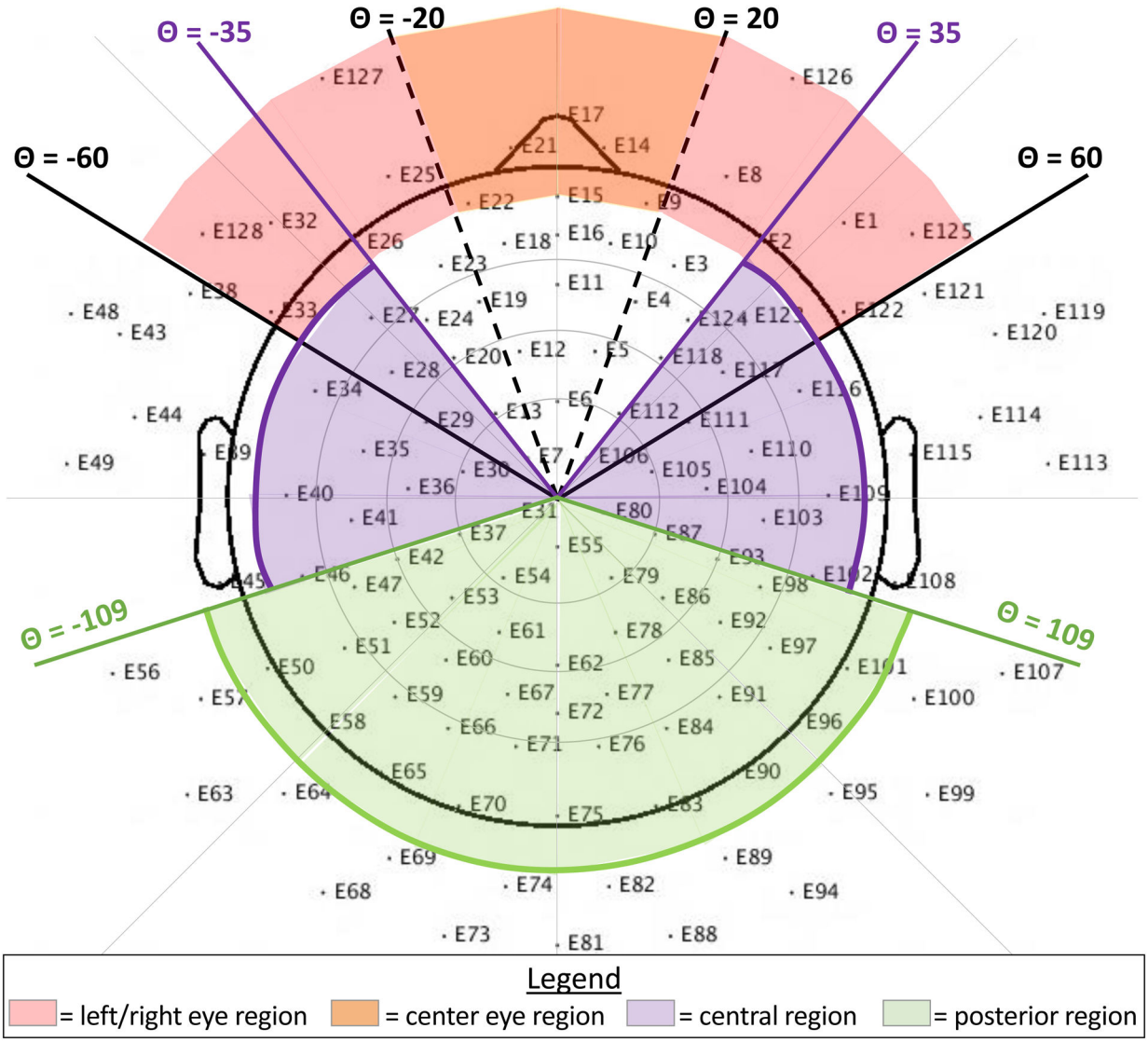

Figure 1.

A 128-channel geodesic net layout showing the five regions of interest (left eye, center eye, right eye, central, and posterior) for the new blink detection function. Note that, for the horizontal eye movement detection function, the theta range (+/−35 to +/−62) and radius (> .5) are slightly different.

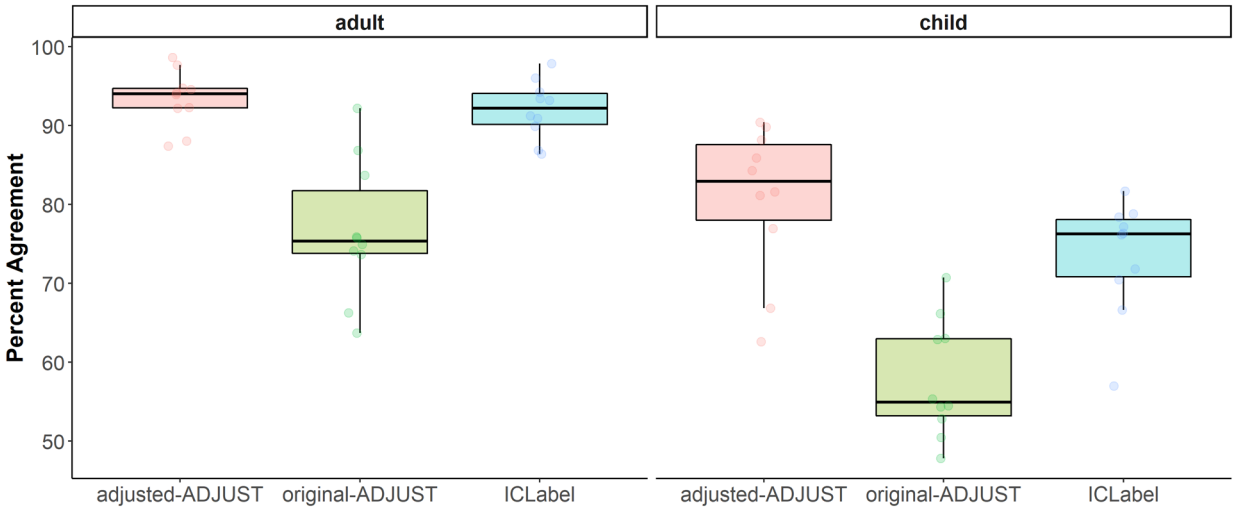

Figure 2.

Box plots showing the distribution of the percent agreement scores for children and adults further split up by each algorithm (adjusted-ADJUST, original-ADJUST, or ICLabel). Each dot represents a score.

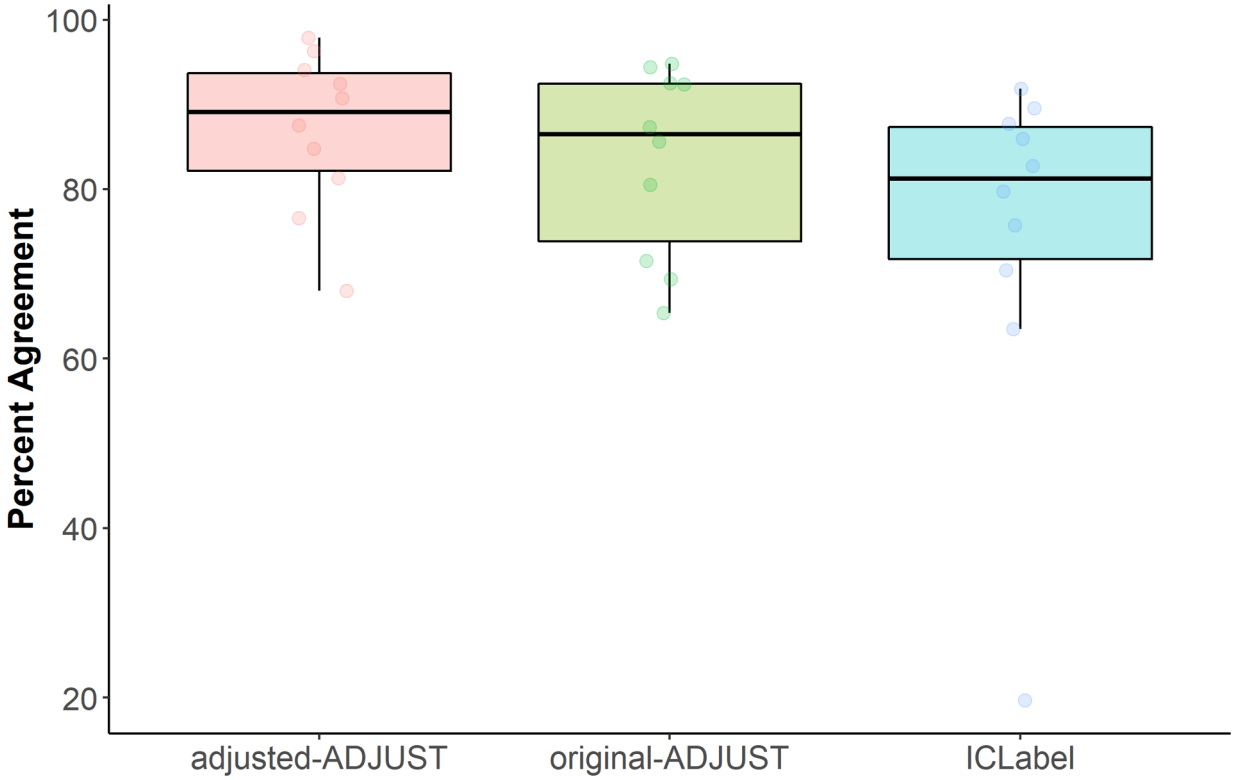

Figure 3.

Box plots showing the distribution of the percent agreement scores for infants further split up by each algorithm (adjusted-ADJUST, original-ADJUST, or ICLabel). Each dot represents a score.

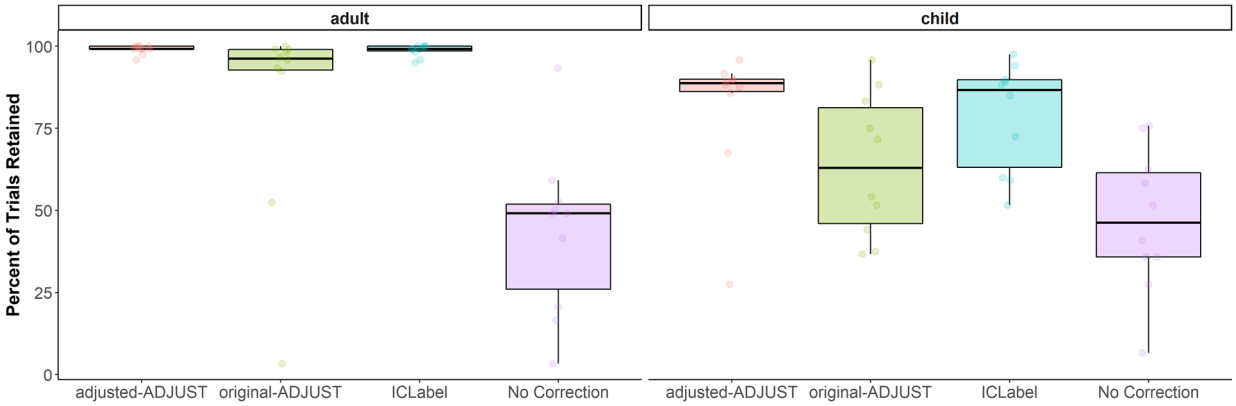

Figure 4.

Box plots showing the percentage of trials retained after artifact rejection for each condition (adjusted-ADJUST, original-ADJUST, ICLabel, or No Correction) further split up by age group (child or adult). Each dot represents a score.

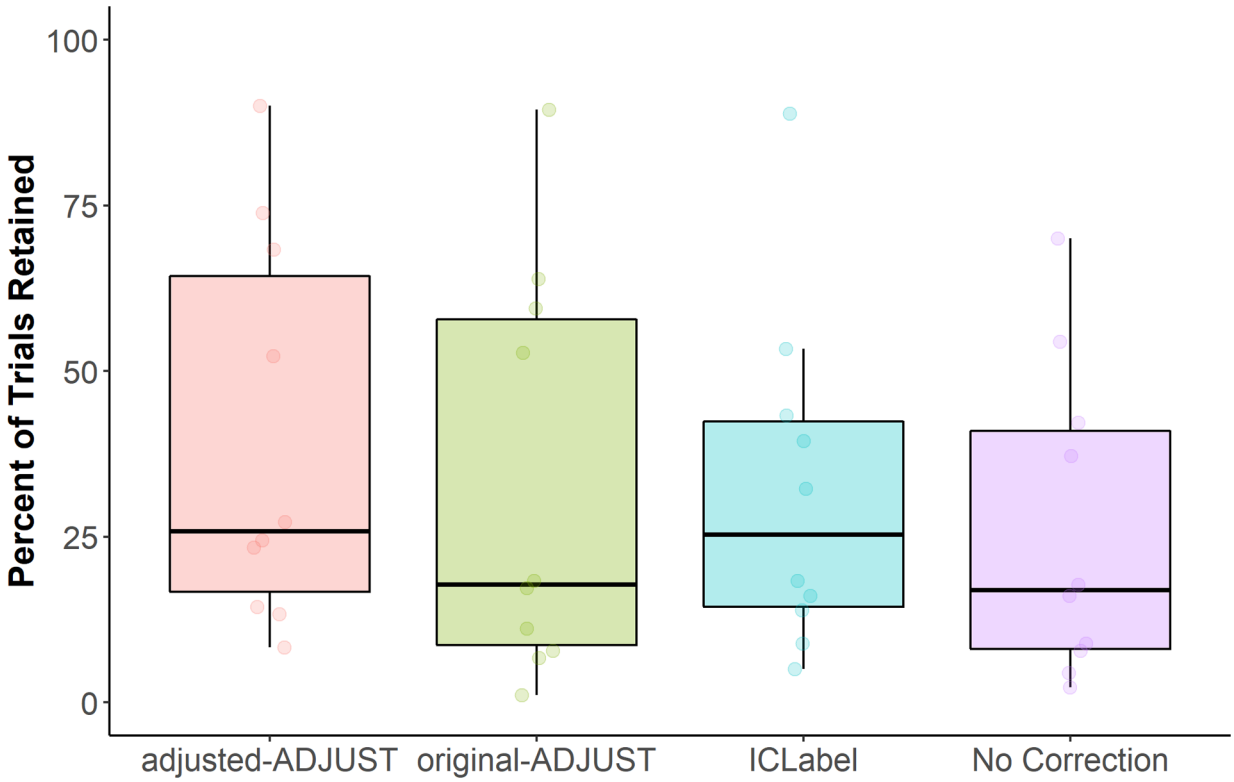

Figure 5.

Box plots showing the percentage of trials retained after artifact rejection for each condition (adjusted-ADJUST, original-ADJUST, ICLabel, or No Correction) for infants. Each dot represents a score.

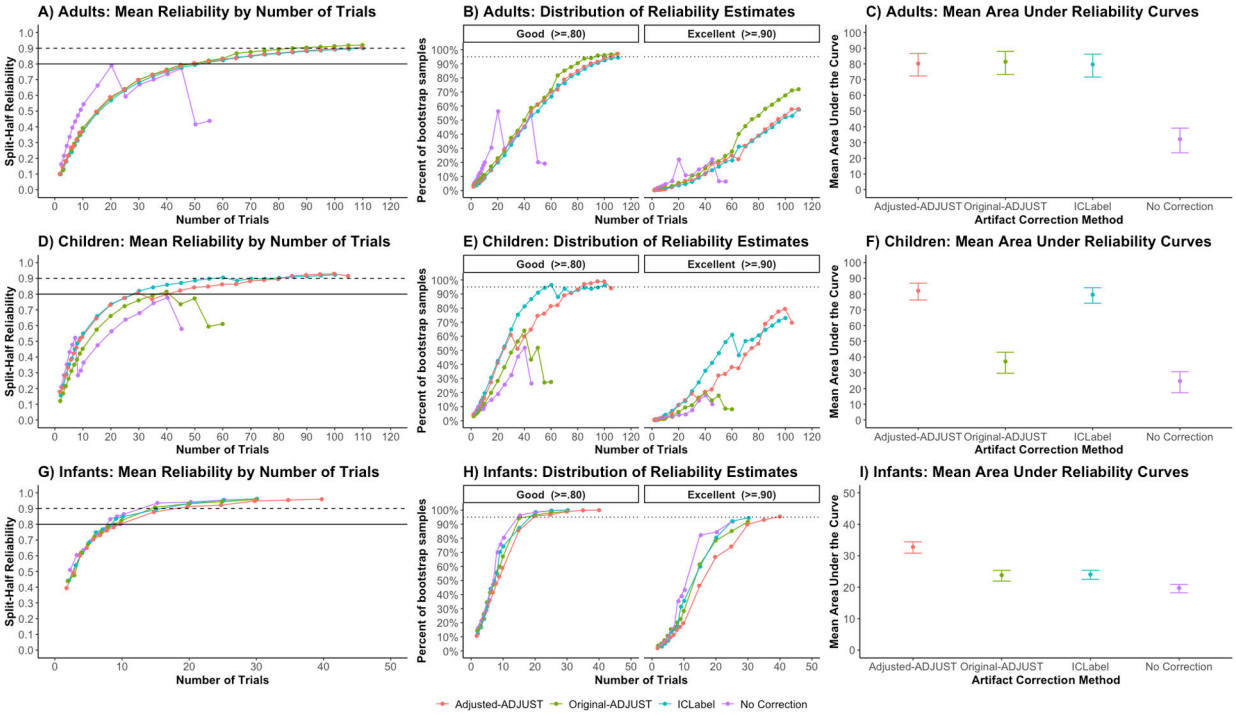

Figure 6.

Internal consistency reliability of the mean P1 ERP amplitude in adults and relative alpha power in infants. Plots A, D, and G show the mean reliability across the 10,000 samples at each number of trials, *n*. The solid line in plots A, D, and G indicates good (> .8) reliability and the dashed line indicates excellent (> .9) reliability. Plots B, E, and H show the percentage of the resampling distribution that achieves good (left) and excellent (right) reliability, which was used to estimate the variability in the reliability estimates across the 10,000 resamples at each number of trials. The dotted line in plots B, E, and H indicates when 95% of the samples reach either good (left) or excellent (right) reliability. Plots C, F, and I show the average area under the curve of the mean reliability curves (plots A, D, and G) with the 95% CIs of the resampling distribution.

Table 1

Repeated Measures nonparametric test of percent agreement scores in adult and child data.

	WTS	df	p-value
Age	52.9579	1	.0001
Algorithm	88.9322	2	.0001
Age*Algorithm	7.6685	2	.0216

Table 2

Descriptive information on percent agreement scores between expert coders and the three correction algorithms for adult and child data.

Age		N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
							25th	50th (Median)	75th
adult	adjusted-ADJUST	10	93.37%	3.61%	87.40%	98.61%	91.16%	94.05%	95.50%
	original-ADJUST	10	76.72%	8.76%	63.72%	92.20%	71.81%	75.37%	84.50%
	ICLabel	10	92.01%	3.70%	86.40%	97.88%	89.14%	92.21%	94.70%
child	adjusted-ADJUST	10	80.78%	9.47%	62.62%	90.42%	74.43%	82.95%	88.59%
	original-ADJUST	10	57.83%	7.43%	47.83%	70.75%	52.27%	54.93%	63.81%
	ICLabel	10	73.46%	7.30%	57.00%	81.70%	69.52%	76.27%	78.50%

Table 3

Descriptive information on percent agreement scores between expert coders and the three correction algorithms for infant data.

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
adjusted-ADJUST	10	86.96%	9.46%	67.98%	97.87%	80.15%	89.11%	94.64%
original-ADJUST	10	83.39%	11.07%	65.41%	94.81%	71.04%	86.46%	92.96%
ICLabel	10	74.69%	21.28%	19.68%	91.84%	68.70%	81.25%	88.21%

Table 4

Repeated Measures nonparametric test on the number of trials retained in adult and child data.

	WTS	df	p-value
Age	14.655	1	.0001
Algorithm	241.979	3	.0000
Age*Algorithm	46.739	3	.0000

Table 5

Descriptive information on the number of trials retained with adult and child data

Age		N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
							25th	50th (Median)	75th
adult	adjusted-ADJUST	10	99.00%	1.35%	95.83%	100.00%	98.75%	99.17%	100.00%
	original-ADJUST	10	83.08%	31.40%	3.33%	100.00%	82.50%	96.25%	99.17%
	ICLabel	10	98.67%	1.81%	95.00%	100.00%	97.71%	99.17%	100.00%
	No Correction	10	43.58%	25.31%	3.33%	93.33%	19.79%	49.17%	54.17%
child	adjusted-ADJUST	10	81.33%	20.35%	27.50%	95.83%	81.25%	88.75%	90.42%
	original-ADJUST	10	63.83%	21.73%	36.67%	95.83%	42.50%	62.92%	84.58%
	ICLabel	10	78.75%	16.54%	51.67%	97.50%	59.79%	86.67%	91.04%
	No Correction	10	47.00%	21.86%	6.67%	75.83%	33.75%	46.25%	65.63%

Table 6

Descriptive information on the number of trials retained with infant data.

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
adjusted-ADJUST	10	39.56%	29.15%	8.33%	90.00%	14.17%	25.83%	69.72%
original-ADJUST	10	32.78%	30.77%	1.11%	89.44%	7.50%	17.78%	60.56%
ICLabel	10	31.94%	25.62%	5.00%	88.89%	12.64%	25.28%	45.83%
No Correction	10	26.11%	23.46%	2.22%	70.00%	6.94%	16.94%	45.28%