UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

Independent Evaluation of the Harvard Automated Processing Pipeline for Electroencephalography using Multi-Site EEG Data from Children with Fragile X Syndrome

A THESIS SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the Degree of

MASTER OF SCIENCE

By

EMMA AUGER

Norman, OK

2020

Independent Evaluation of the Harvard Automated Processing Pipeline using Multi-Site EEG Data from Children with Fragile X Syndrome

A THESIS APPROVED FOR THE

DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Lauren Ethridge

Dr. Hairong Song

Dr. Michael Wenger

Acknowledgements

I would like to acknowledge Dr. Lauren Ethridge for her guidance and encouragement throughout learning programs and pipelines, her support through all the problems encountered, and for the opportunities she has given me. I would like to thank my lab members, Lisa De Stefano, Melody Reese, and Nick Woodruff, for their support and for consistently being the ones who understand and are interesting in my problems and my successes. I would be lost without the knowledge shared by each of these people. Lastly, I would like to thank the technology that made remotely finishing a Master's degree possible.

Table of Contents

Abstract

The Harvard Automatic Processing Pipeline for Electroencephalography (HAPPE) in conjunction with The Batch Electroencephalography Automatic Processing Platform (BEAPP) is a computerized EEG data processing pipeline specifically designed for multiple site analysis of populations with neurodevelopmental disorders. This pipeline has been validated in-house by the developers but external testing using real-world datasets remains to be done. We collected resting and auditory event related data within a clinical trial for 100 children ages 3-6 years with Fragile X Syndrome. The trial encompasses 6 sites using several different EEG systems. This data set represents an ideal test of the new processing pipelines because the data comes from a population with typically high amounts of artifact as well as from different sites and systems. Therefore, we used this rich dataset to evaluate the software's noise reduction techniques, data standardization features, and data integration in comparison to traditional manualized methods of processing. A MANCOVA was used to examine several measures of data post-processing and was found to be significant. Univariate results indicated that the HAPPE/BEAPP pipeline resulted in greater trials retained ($F(4,24) = 5.80$, $p = 0.02$), variance retained through ICA ($F(4,24) = 39.74$, $p < 0.01$), and smaller kurtosis ($F(4,24) = 4.29$, $p = 0.049$) than a manual pipeline for task-related data. No significant differences were found in signal-to-noise ratio (SNR) ($F(1,24) = 0.18$, $p = 0.68$). We did observe an overall loss of signal in the HAPPE/BEAPP pipeline, which is supported by the decrease in kurtosis. In order to further explore the reduction in signal, we processed simulated data in both pipelines. The simulated data was composed of simulated brain, pink noise, and real artifact. We measured correlations between the post-processed data from each pipeline and the pure simulated brain signal. Using a paired samples t-test we determined that the correlation between the pure signal and processed

data was significantly higher for the manually processed data (M = 0.96, SD = 0.03) compared to the HAPPE processed data (M = 0.29, SD = 0.03); t(55) = 105.87, $p < 0.01$. In conclusion, data processed using HAPPE has many benefits including less active processing time and artifact reduction without removing segments. One major drawback is an overall reduction of signal. It eliminates noise and artifact at the cost of reducing signal. Importantly the SNR in the real data was not significantly different between the manually processed data and the HAPPE processed data, so the signal reduction may not negatively affect outcome measures. Therefore recommended implementation of the HAPPE pipeline for neurodevelopmental populations depends on the goals and priorities of the research

Background

Electroencephalography (EEG) can be an effective, relatively inexpensive means for assessing brain activity in a number of contexts. However, use of EEG includes consideration of practical issues related to artifacts in the data. Artifacts can be caused by participants (i.e., as eye blinks, eye movements, muscle tension, or movement) or by environment (i.e., electrical noise, or equipment misuse and malfunction) (Keil et al., 2014). In order to provide accurate assessment of brain activity, artifacts must first be separated from brain signal.

To accomplish this separation, researchers use several methods of artifact removal or correction during preprocessing. There are no established and accepted standards for removal of artifacts, which has resulted in artifact removal techniques varying considerably between researchers. One typical approach to artifact removal is manual selection of artifact free data. This method is time intensive, and due to the reliance on subjective judgment of artifact levels, can vary considerably between processors. It can also reduce the number of trials available for data analysis. However, manual removal allows for specific selection of artifacts and can work especially well for artifacts that are irregular or extreme (Dickter & Kieffaber, 2014).

Another commonly utilized method for artifact correction is independent component analysis (ICA). ICA works by taking the EEG signal, which is a mixture of brain activity and artifact, and blindly separating it from rows of mixed data, separated by channel, into a matrix of temporally independent data sources, separated by a specified number of sources. ICA assumes that the data is linear with minimal delays in measurement, that the time courses of the sources are independent, and that the number of sources is no greater than the number of sensors. EEG data meets each of these assumptions given that the researcher chooses a number of components that is less than or equal to the number of channels used to collect the data. The resulting data

after ICA is divided into components by source rather than by channel. Researchers visually inspect the components in order to determine which are caused by brain activity and which are caused by eye or muscle movements, channel noise, electrical activity, or heart rate. Then the data is reconstructed into its original state, separated by electrode channel, now without the components that contain artifacts and are marked for removal (Jung, 2000; Makeig et al., 1996). ICA is particularly adept at extracting the smaller, more regular artifacts such as blinks, eye-movement, or heart rate because the regularity creates a pattern within the data that ICA can distinguish as originating from one source. However, ICA does not do as well at extracting larger or more irregular artifacts, such as movement or muscle tension, into a singular source. In addition, because these artifacts can create more extreme variance in the data, movement artifacts can disguise the variance of the smaller more regular artifacts (Dickter & Kieffaber, 2014; Jung, 2000).

In order to make ICA component separation more accurate, researchers often employ ICA component removal after removing segments of artifact manually through visual examination and hand selection of the large irregular artifacts (Dickter & Kieffaber, 2014). Manually removing artifacts before ICA prevents larger irregular artifacts from masking smaller regular artifacts like heart rate and eye blinks, which allows for a cleaner separation of components in ICA. However, these methods introduce two points of subjective decision making, in which subjectivity is introduced into the data cleaning process. Researchers are trained to remove specific types of artifacts or to identify components for removal, but the process involves some subjectivity and compromise. Each researcher may choose to remove different sections of data or ICA components.

In attempts to increase standardization, particularly at previously subjective decision points, and decrease labor, some preprocessing software has been developed to automate various parts of the processing. These pipelines have been developed and tested in healthy adults, who generally produce low levels of artifact. Most automated pipelines focus solely on automating the ICA component selection without any additional built in artifact removal (Joyce et al., 2004; Mognon et al., 2011). These automated artifact removal programs can work well in adults or simulated data. However, most are untested in children or neurodevelopmental populations. The automated programs are expected to have more difficulties in removing eye-related artifacts and heartrate in children and other high artifact populations because there are more overall sources of artifacts. Further, these populations tend to have larger amplitude, irregular artifacts related to increased movement and touching of the EEG equipment. To date, systematic evaluations of the discussed automated pipelines have not been done in children or neurodevelopmental populations (Webb et al., 2015).

Another less utilized method for partially automating artifact removal is usage of wavelet transform and thresholding. EEG signals can be decomposed into different time-frequency domains. A wavelet function can be fit to the original EEG signal using the underlying frequency patterns of the data and correlating frequency wavelets to the EEG signal. During this fitting, wavelets can be omitted from the wavelet reconstruction if they fail to meet a certain threshold, which can be and has been defined using several different methods. Some of which emphasize elimination of white noise while other methods aim to get rid of eye-movement or cardiovascular artifacts(Castellanos & Makarov, 2006; Mamun et al., 2013). In order to increase the accuracy of wavelet thresholding, some researchers have attempted systems of combined wavelet thresholding and ICA. One study decomposed the data into wavelets then further decomposed it

using ICA denoising before filtering the wavelets and reconstructing the signal (Walters-Williams & Li, 2011). Another paper used a form of ICA, Second Order Blind Identification (SOBI), to decompose the data into independent components (ICs) and then ran a soft wavelet threshold on the ICs before reconstructing the data (Kaur & Singh, 2015). Each of these methods show some promise. However, most of these techniques have been tested on simulated or partially simulated data that often does not contain realistic artifacts above and beyond white noise, and none of them were tested in children or high artifact contaminated data.

EEG is increasingly being used in studying developmental populations, populations with neurodevelopmental disorders, and more generally in disorders with increased prevalence of artifact within EEG data. Specifically, EEG is being used to develop biomarkers for use in diagnosis and clinical testing for many disorders including fragile X syndrome, autism spectrum disorder (ASD), and epilepsy (Bosl, 2017; Sahin et al., 2018). Although the use of EEG in these populations can inform treatments and diagnoses and is a promising avenue for research, the discussed currently available automated processing pipelines for removing artifact are not designed for high artifact populations and therefore cannot be successfully deployed. The large amplitude artifacts found in these populations can mask the variance in the smaller more regular artifacts resulting in poor separation of artifact from brain activity during ICA. Instead, manual artifact removal and manual ICA component selection are typically used. These manual processes allow for more complicated and precise decision making in the artifact removal process, but they also introduce both increased processing time and subjectivity into the preprocessing.

The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE) attempts to alleviate these issues and create a fully automated preprocessing pipeline that can

handle data with high levels of artifact (Gabard-Durnam et al., 2018). It was originally designed and tested using a dataset from a developmental population with diagnosed autism spectrum disorder. It can process event-related and resting state EEG data. The pipeline puts data through a series of filtering, channel selection, electrical noise removal, channel rejection, wavelet-enhanced ICA, ICA with component rejection, channel interpolation, and re-referencing. It also provides some standardized outputs for assessing data post-processing (Gabard-Durnam et al., 2018). The Batch Electroencephalography Automated Processing Platform (BEAPP) is a platform that integrates various EEG processing tools including HAPPE for use on multiple files, also addressing some issues with multiple data collection systems and sites (Levin et al., 2018).

Purpose and Research Objectives

In the current study, we aim to independently evaluate HAPPE through BEAPP using a dataset from a large multisite clinical trial. The dataset is uniquely fitting for the evaluation of HAPPE for several reasons. It contains data from children with Fragile X Syndrome, which is a single gene neuro-developmental disorder with symptoms of cognitive impairments, extreme social anxiety, sensory hypersensitivity, and some slight physical characteristics such as facial dysmorphologies (Hagerman & Hagerman, 2002). EEG data collected from children with Fragile X syndrome often has a high level of artifact (Hagerman & Hagerman, 2002). In this study, the data was collected using several different EEG systems, in different sites across the US, and has a series of repeated visits per subject. The dataset has both resting-state data, and event-related data. Therefore, this dataset represents an ideal test of HAPPE in the contexts for which it was designed. We evaluated HAPPE in comparison to the typical methods of manual artifact rejection in conjunction with manual artifact component removal after ICA.

Although using real data from children with FXS allows for testing HAPPE in the exact conditions for which it was designed and will be implemented in, there are some limitations in assessing the outcome of the pipelines. We have efficient methods for estimating signal to noise ratio in real EEG data, but true signal and true noise cannot be known in real data. Therefore, a supporting analysis was done using simulated data. HAPPE was developed and tested using real data (Gabard-Durnam et al., 2018). The Multiple Artifact Rejection Algorithm (MARA), a component of the HAPPE pipeline, in particular was developed using machine learning on distinct features of real EEG data (Winkler et al., 2011). Therefore, it may not perform optimally on simulated brain data and artifacts if the simulations are not realistic. Current methods of simulating data often focus largely on simulating only clean brainwaves, or in a single channel, or only a short period of time (Haufe & Ewald, 2019; Pontifex et al., 2017). Simulations typically use white noise as an introduction of noise to their signal, but they often fail to introduce artifacts such as eye-movements, blinks, heart-rate, muscle tension, or general movement and high amplitude artifacts. In order to use simulated data in these pipelines and represent how they would perform in the real world, this data needs to be as realistic as possible by inclusion of realistic artifacts. Our simulated data is composed of simulated signal, pink noise, and artifact components taken from the FXS dataset. Our method of using simulated signal, simulated pink noise, and real artifacts, creates a balance in which the signal can be known and measured before and after processing. Further the artifact is realistic enough to be processed by the pipelines similarly to real data. The data was processed using both pipelines, and signal to noise ratio was determined by assessing the correlation between the processed data and the simulated signal that went into the pipelines. This outcome aids in the interpretation of the results of the main analysis.

For the main analysis, we hypothesized that data preprocessed using manual artifact rejection techniques would have a smaller signal-to-noise ratio and fewer trials retained in comparison to the automated processing pipeline using BEAPP to integrate data and run HAPPE.

We used variance retained and number of components removed to explore the results of any signal to noise ratio differences found between the pipelines. We used variance retained and number of epochs retained to evaluate data retention in resting-state data.

Additionally, for each pipeline, we examined the test-retest reliability of the P1 ERP peak amplitude of an auditory oddball task by evaluating the difference between data collected at two different timepoints for subjects in the FXS dataset. As this is a within-subjects assessment across both time and pipeline, differences in intraclass correlation between pipelines highlight differences in standardization of data cleaning and signal retention.

Lastly, we evaluated and compared qualitative aspects of the pipeline, such as, amount of time needed to process the data, number of salvageable datasets, and problems faced in using the pipeline.

The richness of the current dataset in combination with the simulated data work together to highlight strengths and limitations to the HAPPE/BEAPP pipeline, including specific circumstances or artifactual features which are not reliably classified by automated processing.

Methods

FXS Dataset

For this evaluation, we used data from a double-blind randomized clinical trial which includes 100 children ages 32 months to 6 years that have a diagnosis of fragile X syndrome (FXS), a full FMR1 mutation. All patients participated in a 4 month placebo lead-in period. Task-related EEG as well as resting-state EEG data were collected at baseline, and then again before starting treatment or placebo. The task is a passive listening auditory oddball task lasting approximately 8 minutes. Participants completed a passive auditory oddball task presented using Presentation software (Neurobehavioral Systems, Albany, CA). Stimuli consisted of 432 "standard" tones (1000 Hz; 90% of stimuli) and 48 "oddball" tones (2000 Hz; 10% of stimuli) presented at 70 dB SPL via bilateral speakers. Tones were 70 ms in duration including a 10ms rise/fall with 1000 ms inter-stimulus interval. Order of stimuli was pseudorandomized with the caveat that at least 6 standard stimuli must be presented sequentially before an oddball stimulus would occur once at either the 7^{th}, 8^{th}, 9^{th}, or 10^{th} position in a 10 stimulus train (Schneider et al., 2013). Participants watched a silent video of their choice during stimulus presentation to improve comfort and reduce movement. For the purposes of this study, only the standard stimuli were analyzed.

Resting-state collection lasted approximately 3 minutes. Participants were instructed to sit quietly with eyes open for 1 minute, 30 seconds with eyes closed, then repeat this sequence once more. If a participant was unable to comply with instructions to maintain eye closure, 2 minutes of eyes open resting EEG data was collected. Due to compliance issues in this age range and level of intellectual ability, the majority of participants provided eyes open EEG only, therefore only eyes open resting data was analyzed for the purposes of this study. We use the

EEG data from the initial screening visit, and a baseline follow-up. Both visits occur prior to randomization and treatment administration. The data was collected from 6 different sites and 4 EEG systems, a Biosemi 32 channel system (1 site), an EGI 32 channel system (1 site), an EGI 64 channel system (1 site), an EGI 128 channel system (3 sites). Of the 100 children in the overall study, we were able to collect usable EEG data from 25 children for our analysis. As a result of the population and diverse data collection circumstances, we are able to test the HAPPE/BEAPP pipeline in the exact conditions for which it was designed. For the comparison, all data is preprocessed using both a manual pipeline and the HAPPE/BEAPP pipeline separately in order to evaluate the effectiveness of the automatic pipeline in removing artifact against the known effectiveness of the manual pipeline (Ethridge et al., 2019).

Simulated Dataset

A power analysis was done using G*Power (Erdfelder et al., 1996). Specifically, we calculated how many participants would be needed to find the difference in SNR from the SNR univariate test in the FXS dataset analysis at a power of .8 and alpha of .05. From this, we determined 56 simulated datasets would be sufficient for the simulated data analysis. The simulated data was created to have the same 33 channel layout as the FXS dataset, the same 1000 HZ sampling rate, and be 244.25 seconds long or approximately 4 minutes. It was constructed from 32 components falling into 3 categories. The first category was simulated brain signal. Eight of the 32 components were simulated brain signal. The 8 components were 2 delta components, 2 theta components, 2 alpha components, and 2 beta components. Each had topographies typical of their frequency band but none had identical topographies. The simulated signal was created using EEG Simulation Scripts from the SimEEG program (Bridwell et al., 2018). The second category was pink noise. Pink noise made up 16 of the components in the

simulated data. The pink noise was generated with the power spectral density slope and ranged from -3 dB/oct. to -10 dB/dec with a zero mean (Zhivomirov, 2020).The last category was artifact. Artifacts were selected by examining the ICA components from the FXS data in order to make the simulated data as realistic as possible so that the quality of processing assessed for the pipelines could be more accurately generalized to real data. The artifact components selected contained clear artifacts with as little brain activity as possible. 8 artifacts were selected: 2 eye movement artifact components, 2 blink artifact components, 2 heartrate artifact components, 1 ear muscle artifact component, and 1 neck muscle artifact component. Each artifact component was shifted a varying amount of time for every simulated dataset so that artifacts would not occur at the same time in each dataset. In addition, each artifact component was multiplied by a random number from 0.2 to 2 in order to vary the amplitudes and relative variance accounted for of each artifact.

Manual pipeline – Real EEG Data

Before processing, all FXS data was re-montaged to a 33-channel EEGLAB standard channel layout following the 10-20 system. Then raw data was digitally filtered at .5-120 HZ (12 and 24 db/octave rolloff, respectively; zero-phase; 60 Hz notch) and re-referenced to average reference. The data was visually inspected. During this inspection, sections of high amplitude artifacts as well as sections of artifacts that are present across the majority of channels, such as some muscle tension artifacts, were removed by hand selection. The periods of data selected for rejection range from one short burst of muscle tension or high amplitude to a larger range of several seconds of muscle tension or high amplitude artifact. The amount removed differed between subjects depending on the presence of the larger artifacts. The shortest task-related data file after segment rejection was 266 seconds, and the shortest resting-state file used after segment

rejection was 90 seconds. The lower length limit required to perform ICA on this data was 22 seconds (Groppe et al., 2009). Bad channels were interpolated (no more than one interpolated channel per file) using spherical spline interpolation implemented in BESA 6.0 (MEGIS Software, Grafelfing, Germany). These bad channels were identified visually by pervasive high amplitude irregular artifact throughout the majority of the recording. However, a limit of one interpolated channel was used to retain data integrity given that the data only has 33 channels and interpolating more than 5% of channels would introduce bias into the data. Heart-rate, eye-movement, and muscle movement artifacts were removed using independent component analysis (ICA) implemented in EEGLAB (Delorme & Makeig, 2004) through MATLAB (The Mathworks, Natick, MA) to separate the variance within the data into source components. Then visual inspection was done to remove the components containing artifact.

Decisions regarding component removal or retention were made by examining several features of the data in each component. Brain related components that were marked for retention were recognized by scalp topographies with dipoles, as well as a slowly decreasing power spectrum with peaks between 5 and 30 HZ, and/or visible ERP in the epoched data. As for components marked for removal, eye-related components were identified by topographies in which the variance is localized towards the front of the scalp, as well as through characteristic patterns of moderately high amplitude spikes or steps in the component's waveform, and a power spectrum with low frequencies. Heart-rate related components were identified by diagonally oriented near linear gradients across the scalp in the topography, either no peaks or a very small theta peak in the power spectrum, and a clear regular QRS complex in the waveform. Muscle-tension related components have power concentrated in higher frequencies, and have highly specialized localized topographies over single or very few electrodes, such as outside the head,

the back of the neck, or over an ear. Line noise components were identified by a strong peak in the power spectrum at 50 or 60 HZ. Lastly, channel noise components were identified by topographies localized to one electrode. Movement artifacts should have largely been removed from the data prior to ICA, but were recognized by large amplitude in the waveform, concentrated variance in the epoched data, or variance contained in a small number of trials.

Some components may contain both artifact and brain data or are more difficult to match to a specific artifact (see Figure 1 for examples of each of these artifacts). Researchers had to make judgements regarding level of artifact and amount of brain data to be lost in order to determine if the components should be retained or removed. This can be particularly difficult and increasingly subjective if the source separation between brain and artifact is not clear. The flexibility of a researcher can be a benefit when perfect source separation is not possible, but it can also introduce differences in processing between researchers and datasets.

After the components containing artifact were identified and marked for removal, the data was reconstructed into its original form of separation by channel without the components that were marked as artifact. Task-related data was then segmented into 1500 ms trials (-500 to 1000 ms). Resting state data was blocked into 2 second epochs. Trials and epochs were considered residual artifact if waveform amplitudes exceeded 120μV and were removed. Task data was then averaged across trials for standard stimuli. Event marker timing was adjusted prior to segmentation if necessary according to event timing tests performed at each site prior to data collection. Overall the manual processing pipeline uses subjective reasoning in 3 parts of its process. We had to decide which channel, if any, needed to be interpolated, the segments of data that needed to be removed, and which of the components from the ICA needed to be removed.

Automatic pipeline – Real EEG Data

In order to combine the different EEG systems, data was first re-montaged from their original system layouts to an EEGLAB standard 33 channel montage, and for each file a maximum of one bad channel was selected and interpolated if necessary. BEAPP was used in order to choose preprocessing settings, integrate EEG systems, and run HAPPE on the data (Levin et al., 2018). In BEAPP, settings for preprocessing were decided based on recommended settings for HAPPE as well as with the knowledge that ERP data needs a slightly lower high-pass filter. Those included filtering at 0.5-250 HZ and notch filtering at 60 HZ in order to remove electrical noise. The automated pipeline filtered the data, labeled and removed bad channels, and ran the data through wavelet-enhanced ICA (W-ICA), ICA, Multiple Artifact Rejection Algorithm (MARA), referencing, and then segmentation for task-related data (Gabard-Durnam et al., 2018). The main aspects of this pipeline that are unique compared to the manual pipeline are the automatic selection of bad channels, the W-ICA, and the use of MARA for component classification. Our data was not suited for automatic selection and interpolation of bad channels through HAPPE because it does not allow for a maximum number of channels to be set for interpolation. Since the data is limited to 33 channels, interpolation of more than one channel could bias the data. Therefore, we preselected a maximum of one channel for interpolation and allowed HAPPE to select other channels for removal without reintroduction.

The W-ICA works by decomposing the data through ICA into components, and then the components are subjected to a wavelet transform to further separate the neural data from the high amplitude artifacts and remove the statistically separated out artifact from each component without removing the component as a whole. In this way, it eliminates larger artifacts from the data while retaining neural signal. Then the data is reconstituted into its original form organized

by electrode channel (Castellanos & Makarov, 2006). After W-ICA, the data is put through ICA again and broken down into components. Then the components were examined for retention or rejection by MARA. MARA uses 6 features of EEG data to give a rating of artifact for each component. Those features include current density norm, range within pattern, mean local skewness, λ and fit error, and the average log band power of the alpha band (8-13 HZ). The six features are a result of the reduction of 38 features of ICA to the 6 features that best discriminate brain activity from artifact. Current density norm is an estimation of the complexity of the underlying source location and spread with more complex sources more likely to be artifact. Range within pattern is the logarithm of the differences between the minimum and maximum activations in a component's pattern. Mean local skewness gives the mean absolute local skewness for 15 second time intervals over the component waveforms. λ indicates wavelength deviation of each component's power spectrum from a 1/frequency curve. Fit error also examines the difference between a components power spectrum and a 1/frequency wave by using the logarithm of the mean squared error. The last feature MARA uses is the average log band power of the alpha band (8-13 HZ). After rating each of these features, MARA outputs an overall rating of the probability of artifact in a component to choose which components to remove (Winkler et al., 2011, 2014). From the MARA output, HAPPE generally uses a cut off of 0.5 for eliminating components, but suggests an increase to 0.7 for high artifact contaminated data. After examining processed data with both the 0.5 threshold and the 0.7 threshold, we determined that the 0.7 threshold was a better fit for this dataset. It did a better job retaining relevant signal in some files and was not noticeably detrimental compared to the 0.5 threshold in any files. Once data was through ICA and MARA, data was average referenced, and then task-related data was segmented into 1500 ms trials (-500 to 1000 ms). HAPPE outputs the processed

data as well as some descriptive results for each file. Resting state data was blocked into 2 second epochs. Trials and epochs were considered residual artifact if waveform amplitudes exceeded 120μV and were removed. For the task related data, we then averaged the pre-processed data across trials for standard stimuli. Overall, algorithms within the HAPPE/BEAPP pipeline make discriminations about artifact and its removal in two parts of the pipeline, in W-ICA and in determining which ICA components to reject or retain. In this particular dataset an additional subjective element was introduced because HAPPE had no method for placing an upper limit on the number of channels to be interpolated, so we preemptively interpolated a channel if needed before running the data through the pipeline.

Processing Simulated Data

For the simulated data, both processing pipelines were run with all of the same settings as the real data with two notable exceptions. First, the simulated data was created using artifact components from ICA run after data had been through manual interpolation of channels and manual selection of data for removal. This means that although the data contains real artifact, it is retains less large-scale artifact than raw data from children with FXS. Due to this caveat, for the simulated data in the manual pipeline, no additional channels were removed and no additional segments of artifact were removed. The primary method for artifact removal in the simulated data was ICA and manual ICA component removal. In addition, one problem encountered with processing simulated data is that artifacts were too easily identifiable for both researchers using the manual pipeline and for the automated pipeline. ICA was able to perfectly separate the simulated data into its 32 components with ease due to their reduced statistics dependence relative to real EEG data. Therefore, we limited ICA decomposition in both the manual pipeline and the ICA part of the automated pipeline to 16 components. This forced the

ICA algorithm to find fewer sources of variance, combine components and introduced some minor mixing between different types of components as well as some ambiguity in the artifact removal process, and more accurately simulated real-world decision making with mixed sources.

Comparisons between manual and automated pipelines in FXS data

We evaluated the differences between the pre-processed data from the manual and automated pipelines in several areas: signal to noise ratio (SNR), number of trials/epochs retained, variance retained, and system and site differences. In order to estimate the level of noise for the SNR, we multiplied the values of every other epoch for the standard waveform by -1 to flip the waveform so that positive peaks would be negative and negative peaks would be positive. Then we combined the altered standards with the normal standard trials. This combination retains power and variance throughout the dataset, but cancels out any event-related effects (van Drongelen, 2007). SNR was calculated by the amplitude at peak amplitude of the P1 ERP (40-150 ms) for all standard trials divided by the amplitude of the flipped and combined standard trials during the same time window (Thigpen et al., 2018). Percent variance retained from ICA component rejection is automatically output by HAPPE and was calculated for the manually preprocessed data using the same EEGLAB function used in HAPPE. For the task-related ERP data, a MANCOVA was conducted in SPSS with the dependent variables: SNR, number of trials retained, variance retained, and kurtosis. The covariates used were system and site. The dependent variable was the pipeline (e.g, manual or automated), and was treated as a repeated measure. Signal-to-noise ratio allows us to compare how well the pipeline reduces artifact or noise by evaluating the known and expected signal of a reliably activated sensory processing ERP peak (the P1 response) to the standard tone in our oddball task paradigm. Number of trials retained allows us to assess data retention, which is specifically impacted by

manual segment rejection in the manual pipeline as well as the final artifact threshold of 120μV for both pipelines. Variance retained only assessed the proportion of variance in the data after ICA compared to before ICA. It is relevant to both pipelines as both pipelines use ICA and component rejection, but it is influenced by the different steps preceding it in the different pipelines. Lastly, kurtosis allows us to examine the variance and outliers left in the data after processing. For resting-state data, a MANCOVA was conducted with the following independent variables: variance retained, epochs retained, and kurtosis. Again, the dependent variable was the pipeline, which was treated as a repeated measure. The covariates were system and site. Test-retest reliability was calculated via intraclass correlation for the ERP peak amplitude of P1 for the standard trials in the auditory oddball task from the screening visit to the visit at baseline for each pipeline separately. Lastly, we compared the intraclass correlations by transforming the r-scores to z-scores and running a difference test.

Comparison for OB $Y_i = \beta_0 + \beta_1 Pipeline_i + \beta_2 System_i + \beta_3 Site_i + e_i$, where SNR, Variance retained, Kurtosis, and Trials Retained are the DVs, and Pipeline is the IV.

Comparison for Resting $Y_i = \beta_0 + \beta_1 Pipeline_i + \beta_2 System_i + \beta_3 Site_i + \beta_4 Days_i + e_i$, where Seconds retained, kurtosis, and variance retained are the DVs, and Pipeline is the IV.

Intraclass correlations and comparisons

For test-retest: intraclass correlation of the peak amplitude of the P1 for standard trials for each dataset in each pipeline. In order to compare the pipelines we transformed the resulting Rs into z-scores and found the difference between them using the following formula:

$$Z_{obs} = \frac{Z_{manual} - Z_{HAPPE}}{\sqrt{\left(\frac{1}{N_{manual}-3}\right)+\left(\frac{1}{N_{HAPPE}-3}\right)}}$$

Comparisons between manual and automated pipelines in simulated data

The correlation between the manually processed cleaned data and the pure simulated signal that went into the simulated data were calculated for all 56 simulated datasets. The correlation between the HAPPE, automatically processed cleaned data and the pure simulated signal was also calculated for every dataset. These correlations provide an overall measure of how well the pipelines retain signal and eliminate noise and artifact. A dependent t-test was used to compare correlations for the manually processed data and correlations for the automatically processed data. In addition to this analysis I calculated the power of the manually processed data, the HAPPE processed data, and the pure signal within delta, theta, alpha, and the beta frequency bands. I then calculated correlations between the processed power data and the pure signal power in order to further examine if any particular frequency band was differentially reduced in the cleaning process. Lastly, I ran time-series multiple regression analyses for every channel and every subject in both pipelines with the cleaned data as the dependent variable, and the original frequency band simulated signals as the regressors.

Results

Task-related MANCOVA

In order to examine the difference in processed task-related data from a manual processing pipeline as compared to HAPPE, a one-way repeated measures MANCOVA was used to allow an overarching look at multiple aspects of the resulting processed data for 25 files processed through each pipeline. The MANCOVA had one independent variable, the pipeline used to process the file, and four dependent variables, variance retained after ICA, trials retained, kurtosis of the waveform, and signal-to-noise ratio. Site and System were included as covariates. The model was highly significant ($F(4,21) = 14.85$, $p < 0.01$) demonstrating that there were differences in the processed data resulting from the two pipelines. There was a significant interaction between pipeline and site ($F(4,21) = 4.35$, $p = 0.01$), and no significant interaction between pipeline and system ($F(4,21) = 2.16$, $p = 0.11$)

The univariate ANOVAs for the model allow investigation of the significance of the overall MANCOVA and of each outcome. The univariate test for SNR was positively skewed, so we log-transformed the variable, and it was not significant ($F(1,24) = 2.33$, $p = 0.14$). The univariate test for kurtosis was significant ($F(4,24) = 4.29$, $p = 0.049$), with a smaller kurtosis for HAPPE processed data. Univariate tests for trials retained and variance retained were each significant (respectively, $F(4,24) = 5.80$, $p = 0.02$, $F(4,24) = 39.74$, $p < 0.01$), with more variance retained and trials retained in the HAPPE processed data (See means in table 1 and univariate results in table 2). The significant difference in kurtosis between the pipelines without a significant difference in SNR may reflect a change in the scaling of the HAPPE processed data.

Beyond the within-subject effects, between-subjects effects were used to examine the effects of site and system. No significant differences were found in SNR ($F(1,24) = 0.18$, $p =$

0.68), trials retained ($F(1,24) = 1.81$, $p = 0.19$), nor kurtosis ($F(1,24) = 0.04$, $p = 0.84$) between systems. Also no significant differences were found in SNR ($F(1,24) = 1.86$, $p = 0.19$), trials retained ($F(1,24) = 2.39$, $p = 0.14$), nor kurtosis ($F(1,24) = 1.06$, $p = 0.31$) between sites. However, there was a significant difference in variance retained for both systems ($F(1,24) = 6.97$, $p = 0.01$) and sites ($F(1,24) = 4.77$, $p = 0.04$). This reflects differences in the amount of artifact removed in the ICA step. Therefore, although the different sites and EEG systems may differ in the amount of artifact removed, there seems to be no significant difference in the resulting cleaned data.

Resting-state MANCOVA

Although signal retention is more difficult to assess in resting-state data given the lack of a defined and expected signal to examine, it is still important to assess how the pipelines handled resting-state data given its abundant use and generally shorter file length. Therefore, we ran a MANCOVA assessing data retention of resting-state data between the two pipelines with 4 DVs, kurtosis, variance retained in ICA, components retained in ICA, and seconds retained, with pipeline as the IV. Site and system were treated as covariates. The model was not significant ($F(3,20)=1.91$, $p = 0.15$). Further none of the univariate tests were significant: kurtosis ($F(1,20) = 3.50$, $p = 0.08$), variance retained ($F(1,20) = 2.04$, $p = 0.17$), seconds retained ($F(1,20) = 0.13$, $p = 0.725$). Therefore, data retention of resting-state data was not different between processing pipelines.

Intraclass Correlation

Of the 25 files processed through both pipelines, 10 had both a screening and a follow-up visit. Therefore, an intraclass correlation was run to examine the reliability of the P1 peak amplitude for standard trials in the task related data from the screening to the follow-up. For the

data processed with the HAPPE pipeline, the intraclass correlation coefficient was 0.35. This indicates a weak to moderate relationship between P1 peak amplitude between visits. For the data processed with the manual pipeline, the intraclass correlation coefficient was 0.45 showing a moderate relationship between visits. After transforming these correlation scores to z-scores, the resulting difference test found no significant difference between the reliability of the P1 amplitude processed with the HAPPE pipeline and the reliability of the P1 amplitude processed with the manual pipeline ($z = 0.22$, $p = 0.83$).

Simulated Data Correlation and Multiple Regression

Pearson correlations were calculated between each processed set of data within both pipelines and the original simulated signal. Using a paired samples t-test we determined that the correlation between the pure signal and processed data was significantly higher for the manually processed data ($M = 0.96$, $SD = 0.03$) compared to the HAPPE processed data ($M = 0.29$, $SD = 0.03$); $t(55) = 105.87$, $p < 0.01$. Therefore, the manual processing pipeline recovered the original simulated signal to a greater extent than the HAPPE pipeline. Based on the main analysis we predicted that there would be a decrease in the scale of the signal when data is processed through HAPPE, but this comparison suggests that the difference occurs in both scale and in cleaning capabilities for simulated data. In order to further examine where those differences occur and if certain frequency bands are differentially affected, we also calculated power for delta, alpha, theta, and beta bands in the manually processed data, HAPPE processed data, and pure signal. Then we ran correlations between the processed datasets and the pure signal (see Table 3). The manually processed data had strong correlations with the pure signal for theta and alpha, a moderate correlation for delta, and a weak non-significant correlation between beta measurements. The HAPPE processed data and the pure signal had a weak significant correlation

for delta, and weak non-significant correlations for alpha, theta and beta. This shows greater reduction of signal in the beta band for both pipelines, and greater reduction of alpha and theta for the HAPPE pipeline. Regression coefficients for each signal and each channel were calculated and averaged across simulated subjects. Then the averaged beta weights were plotted on topographies for delta, theta, alpha, and beta predictors as well as for the residual variance (shown in Figure 6). The manual pipeline regression coefficients appear to be greater in all four frequency bands compared to the HAPPE processed regression coefficients, and the residual variance is smaller for the manual regression across the scalp.

Discussion

The implementation of a fully automated processing pipeline for removal of artifact from high artifact EEG data could be revolutionary for EEG research and its expansion and standardization. If the success of HAPPE and its use through BEAPP can be independently replicated and supported, then it would fulfill this need. The main MANCOVA model for the task-related data found significant differences between the processed data from the manual pipeline compared to processed data from the HAPPE/BEAPP pipeline. The main differences were driven by increased data retention shown by greater number of trials retained in task related data. In addition, variance retained in the ICA step of each pipeline was compared, and it was found that the HAPPE/BEAPP pipeline retained more variance. However, this difference is most likely driven by greater reduction in variance in the steps leading up to the ICA in the HAPPE/BEAPP pipeline, e.g. channel rejection and wavelet-threshold ICA, compared to the steps leading up to ICA in the manual pipeline, e.g. visual segment rejection. This idea is supported by the significant difference in kurtosis, which was significantly different between pipelines and showed decreased kurtosis for the data processed in the HAPPE pipeline. This reduction in kurtosis indicates reduction of variance and specifically artifact reduction. Perhaps the most important indicator of signal retention, artifact rejection, or both is SNR. The model found no significant difference between the SNR of data cleaned by the two processing pipelines. These results fit with the conclusion that there is an overall reduction in variance when files go through HAPPE, and this leads to a decrease in the scale of the HAPPE processed data in comparison to the manually processed data. The change in scale is shown in the range of the standard deviations of the processed data: the automatically processed data (min = 1.02, max = 4.87), the manually processed data (min = 7.09, max = 18.57).

The between subject effects for site and system showed that the outcome measures quantifying cleanliness and length of the processed files did not differ between sites or systems. The only outcome measure that did differ between the sites and EEG systems was variance retained. Variance retained focuses solely on the amount of variance that was removed in ICA and can be a measure of data retention, but is difficult to interpret without other factors such as quantifications of signal and noise within the variance removed or retained. However, one likely explanation for the significance in variance retained between sites and systems is that there are differing levels of artifact in the pre-ICA data. If that is the driving force behind this difference then these results show that even if sites and systems had different levels of artifact in their raw data, they did not have different measures of kurtosis, SNR, or trials retained post analysis, which in combination supports the use of either pipeline for use with integrating different sites and systems.

The comparison of the HAPPE/BEAPP pipeline and manual pipeline processing for the resting-state data showed no significant results, leading to the conclusion that the pipelines did not differ in data retention for the resting-state data. This result is particularly interesting given that the processing of the data in both pipelines should be no different for resting-state data compared to task related data. Since we included some of the same measures in the resting-state MANCOVA we expected similar results for variance retained, seconds retained (in place of trials retained), and kurtosis as were obtained for the task-related data. The most obvious explanation for this discrepancy is the length of the files. The resting-state files in this dataset were much shorter than the task related data, ranging from 138 seconds to 364 seconds with an average of 191.22 seconds. They are adequate length for the measures needed for the clinical study such as measures of power and are an adequate length for cleaning and processing procedures. However,

the length of the file may influence the performance of the pipelines in the wICA step of the automated pipeline or the ICA steps (Delorme et al., 2007; Makeig et al., 1996). This influence may or may not impact the final cleaning steps of the data. Since this MANCOVA only examined some aspects of overall data retention and not signal retention or noise reduction, we cannot make conclusions about how well either pipeline performed in cleaning the resting-state data. However, we can note that although the manual pipeline included segment rejection, the two pipelines did not differ significantly in seconds retained. The ICA step in the pipelines did not perform significantly differently in either components rejected and variance retained. Lastly, kurtosis was not significantly different in the cleaned data from the two pipelines possibly indicating similar levels of artifact reduction, or more generally, similar levels of reduction of variance and outliers.

The results of the intraclass correlation were also not significant. Both pipelines showed weak to moderate intraclass correlation of the peak amplitude of the P1 for standard trials from the screening visit to the baseline visit. The low intraclass correlation may be driven by the young age of the participants (2.5 – 6 years) in combination with the time between visits (~4 months). There was no significant difference in the intraclass correlations between the two pipelines. Given that the two pipelines were processing the same files and the HAPPE/BEAPP pipeline should increase standardization, we expected that would result in increased reliability of ERP measurements, but the HAPPE/BEAPP pipeline did not have a greater intraclass correlation. A main reason for using automated pipelines is increased standardization and the increased reliability that presumably can come from the increased standardization. This could potentially indicate that the increased standardization does not result in better reliability and therefore negates a major advantage for using this automated pipeline.

There are many aspects of signal retention, artifact rejection, and overall processing that are difficult to quantify and evaluate given the complex nature of EEG and the difficulty in parsing out true signal and true artifact. If one could determine true brain and true artifact easily then cleaning EEG data would be simple and perfectly automating the cleaning process would be a standard procedure in lab settings. Therefore, in addition to quantifying the differences between data processed by each pipeline in the ways already discussed, the cleaned data was examined qualitatively to see if there were trends in how each processing pipeline was successful or unsuccessful.

Qualitative Differences

A side-by-side comparison immediately reveals differences in the appearance of the cleaned waveforms after being processing by the two pipelines. While every HAPPE/BEAPP processed file appears very similar and clean, the level of cleanliness of the manually processed files varies somewhat from file to file. The only aspect of the data that was distinctly visually different between the automatically processed files was the degree to which high frequency artifact contaminated the data (see Figure 4). Beyond this, the most apparent difference between the two sets of processed files is the scale. The data processed with the HAPPE/BEAPP pipeline is reduced in absolute scale (representative examples shown in Figure 4 and Figure 5). Again, this reduction in scale can be seen in the minimum and maximum standard deviation values for the automatically processed data (min = 1.02, max = 4.87) compared to the manually processed data (min = 7.09, max = 18.57).One of the main strengths of the automated pipeline is that it consistently removed eye movements and large amplitude artifacts, and if the reduction in scale does not influence signal retention then it may be suitable for certain analyses.

In order to further investigate where the differences between the cleaned data arise we investigated the performance of critical steps in the HAPPE/BEAPP pipeline. The HAPPE/BEAPP pipeline allows for data output at almost every step of the process with the exception of between the wICA step and the ICA and component removal step. However, because wICA is one of the main computational steps differentiating HAPPE from this manual pipeline and other automated pipelines, and in order to examine the extent of the cleaning occurring in wICA compared to ICA, the data were output after wICA and before ICA. From visually inspecting the data between the wICA and ICA steps, it appears that the greater overall reduction in artifact and variance observed in the HAPPE processed waveforms is driven largely by the wICA step. After this step, the majority of the artifacts in the waveform are removed, particularly, eye-movements, blinks, and low frequency movement artifacts. It is unsurprising after this observation that the ICA step of the HAPPE pipeline removed much less variance in the ICA step compared to the manual pipeline. In the HAPPE pipeline, ICA seems to act as a back-up for left over artifact rather than a main method of removing artifact. Whereas, in the manual pipeline, the ICA step is a key part of the artifact removal process and the only method used for certain artifacts including eye-related artifacts and cardiac artifact.

Lastly, the resulting ERPs for the data processed with the two pipelines showed similar visually observable differences as the waveforms. When examining the ERPs, the noise and signal of the manually processed data were much larger than the noise and signal of the automatically processed data (see Figure 5). The model showed the SNRs did not significantly differ between pipelines. Therefore, it seems to be the case that the automated processing pipeline is reducing both signal and noise more than the manual processing, but not providing increase in SNR compared to the manual pipeline.

Both processing pipelines took about the same amount of time to process a single file, about 40 minutes. However, the time spent processing in the manual pipeline was time in which a researcher had to be actively processing. The HAPPE pipeline processes files without intervention from the researcher, and the BEAPP processor allowed for files to be run in batches, so many can be run at once without interference from a researcher. This is a major strength in using the automated pipeline. If the researcher decides to make a change such as in level of filtering or resampling, it is fairly easy to reprocess all of the files. However, we did encounter several technical issues in using both BEAPP and HAPPE that could perhaps be improved upon in future releases.

First although BEAPP is designed to analyze data from multiple EEG systems and integrate it, it does not have a way to montage the data to a new channel layout. Since our data was coming from different layouts, we needed to montage it to a standard layout prior to cleaning in order to retain standardization of processing between sites. Therefore, we had to montage the data in EEGLAB prior to putting it through the automated pipeline. Since full automation is the goal, this is one area in which the pipeline fails to meet that goal. One option it does have for integrating the systems could be to select only the channels which overlap most in location coordinates between systems to use and discard the rest of the channels. However, the channels may not be located in exactly the same place and this can introduce systematic differences between sites or systems. Another problem encountered is that the HAPPE/BEAPP pipeline removes any channels from the data that are three standard deviations outside of the mean amplitude. It also interpolates data back into those channels after the data cleaning process is over. This could be a good and standard way to remove bad channels from the data. However, no upper limit can be placed on the number of channels removed and interpolated. Our data

contains 33 channels so interpolation of more than one electrode would be over 5% of the data and interpolation would bias the data. Therefore, we had to inspect the data visually and interpolate an electrode, if needed, before running the data through the automated pipeline. This introduces subjectivity into the automated pipeline and undermines one of the main reasons to use it which is the removal of subjectivity from the data cleaning process. However, BEAPP did give an option for removing bad channels without interpolating them back in after processing. This allowed for HAPPE to work within the conditions it was designed, eliminating further bad channels, without biasing the data through interpolation. In contrast, in the manual pipeline, we did not remove channels above and beyond the maximum of one channel interpolated.

One major strength and limitation of this analysis is our use of real data. Since our dataset was made up of various EEG systems and contained data from children with FXS, it allowed us to test HAPPE and BEAPP together in the exact conditions for which they were developed: from neurodevelopmental populations with high artifact contamination and multiple sites and systems integrated together. However, it was difficult to accurately capture and assess performance of the pipelines given the inability to know and parse out what is truly brain activity and truly artifact. Therefore, our next step was to include simulated data in addition to the FXS dataset in order measure the exact levels of signal and of noise in the data pre and post processing. This allows us to capture both signal and noise reduction resulting from the two pipelines.

Simulated Data

The manually processed simulated data was significantly more correlated with the original simulated signal than the HAPPE processed simulated data was. This correlation indicates a relationship between the processed data and the original pure signal that went into the data. The strong correlation between the two implies that the manual pipeline removed most of

the artifact and noise and retained most of the signal. However, the data going into the manual pipeline was already cleaner than data in the FXS dataset because it was made up of artifacts from the FXS dataset that had been through the interpolation and segment rejection steps of the pipeline. Therefore, the interpolation and segmentation steps are not included in this assessment. Another limitation was a lack of blindedness in the manual processing. The researcher removing artifact was aware that the data was simulated, which could have influenced decision making. However, this same adjustment in processing is what makes automating EEG processing such a difficult task. The weak correlation between the HAPPE processed data and the pure simulated signal suggests either a reduction of signal or retention of artifact or a combination of both. There was a reduction of signal in the HAPPE processed files of the main analysis, so this could be indicative of that change, but reduction in scale alone could not affect the relationship between the processed data and the pure signal to that extent, so the processed data likely retains some of the artifact or noise. In order to further investigate how the signal is affected, we examined the power of each frequency band in the processed data in comparison to the pure signal. That analysis showed that all four frequency bands had signal loss and a decrease in power. In particular the higher frequencies, beta and theta did not correlate significantly for the HAPPE processed data. Beta power was also not significantly correlated for the manually processed data and the pure signal. This may indicate that the higher frequency beta signal mixed more with one of the artifacts or with the pink noise resulting in increased loss with the removal of artifact. Overall, the manual pipeline did a better job of recovering the original signal and removing the artifacts and noise for the simulated data. One major limitation to this conclusion is that this is simulated data and its generalizability to real data is uncertain. The HAPPE artifact removal algorithms were trained on real EEG data and may not operate as effectively on

simulated datasets, regardless of the effort made here to make the simulated data as realistic as possible using artifact waveforms derived from real data. However, similar patterns arise in the simulated data as in the FXS data. An overall decrease in power for the HAPPE processed data is seen in both analyses.

Although HAPPE is a fairly new processing pipeline, it has been used successfully in a number of studies. So far, it has been applied in studying resting-state power bands in infants and young children with autism (Gabard-Durnam et al., 2019; Wilkinson, Gabard-Durnam, et al., 2019; Wilkinson, Levin, et al., 2019). These studies are the perfect candidates for HAPPE in that they are working with populations that may have shorter recordings with more artifact, and they are focusing on power related outcomes. HAPPE makes it easier to process the data, and although there may be signal reduction, they are looking at differences in power. Therefore, as long as the signal reduction is consistent, it should still work as a measure of change within a population. HAPPE has also been used in a study of perceived maternal stress and its relationship with infant beta and theta power (Pierce et al., 2019). While this population is ideal for HAPPE use, there is concern over how power is affected by the processing when comparing it to other outcomes. It is possible that the relationship found between perceived maternal stress and EEG power could be impacted by the signal loss. Additionally, our results showed that HAPPE and BEAPP could be used to standardize processing between sites and EEG systems. Therefore, it could be used as a tool for future meta-analyses where raw data can be obtained.

If implemented successfully, HAPPE in combination with BEAPP could potentially decrease the labor cost of EEG processing and could increase standardization of EEG processing. However, there are issues with signal retention that could impact certain outcome measurements. It eliminates noise and artifact at the cost of reducing signal. The SNR in the real data was not

significantly different between the manually processed data and the HAPPE processed data, so the signal reduction may not negatively affect outcome measures. However, there was a trend toward a decrease in SNR, and the signal in the simulated data was disproportionally impacted by HAPPE. Therefore, recommended implementation of the HAPPE pipeline for neurodevelopmental populations depends on the goals and priorities of the research. HAPPE is useful for integration across sites and systems and for some resting-state power comparisons. It would not be recommended for ERP analyses where the signal is less robust, where outcomes need to be compared to outcomes from other studies that did not use HAPPE, or for comparison between power and other variables.

Tables

Table 1: Descriptive statistics

	Variance Retained		Trials Retained		Signal-to-Noise Ratio		Kurtosis	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Automated Pipeline	88.00	8.07	99.97	0.13	0.34	0.48	3.13	0.17
Manual Pipeline	32.51	14.42	79.51	16.04	0.66	0.69	9.37	5.95

Table 2: Univariate MANCOVA results

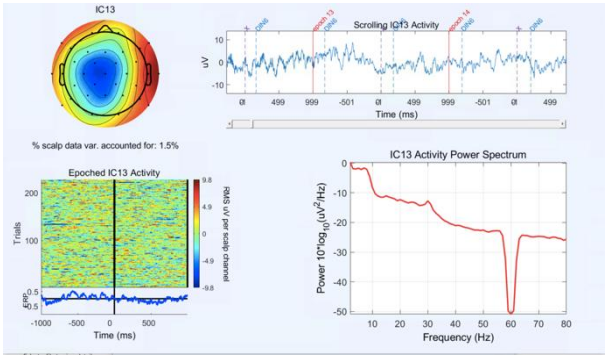
	df	MS	F	*p-value*
SNR	1	.625	2.33	0.14
Variance Retained	1	4461.1	39.74	<0.01
Trials Retained	1	826.29	5.80	0.02
Kurtosis	1	73.74	4.29	0.049

Table 3: Power Correlations

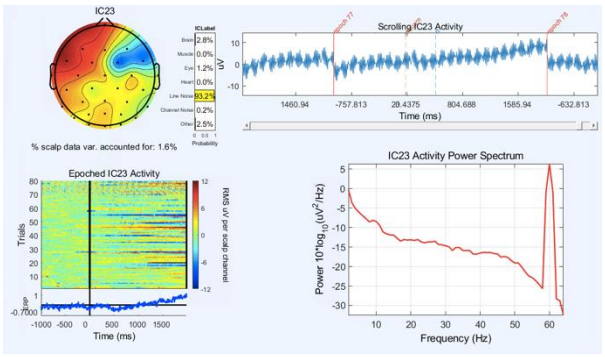
		Mean (uV^2)	SD	Delta	Theta	Alpha	Beta
Manually Processed	Delta	4.07	0.40	0.35*			
	Theta	1.77	0.08		0.71*		
	Alpha	0.68	0.03			0.60*	
	Beta	0.53	0.07				0.18
HAPPE Processed	Delta	0.17	0.04	0.27*			
	Theta	0.07	0.01		0.34*		
	Alpha	0.04	0.01			0.23	
	Beta	0.07	0.01				0.16

Note: '*' marks significance at alpha 0.05. The power for the processed data from the manual pipeline and HAPPE, and the correlation with the power for the corresponding pure signal.

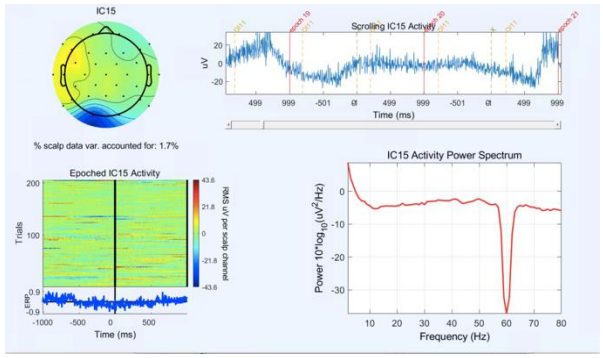
Figures



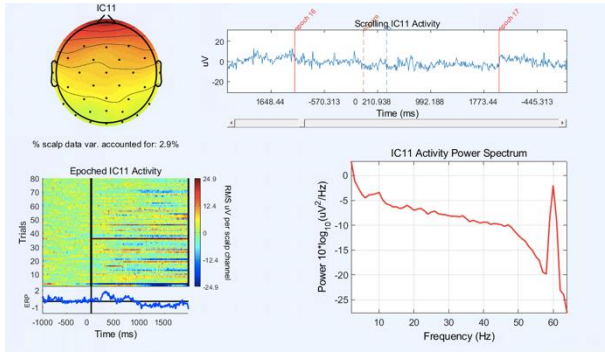
1. Brain Component



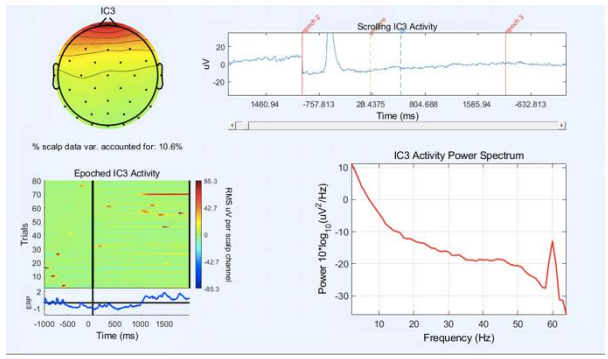
2. Line noise component



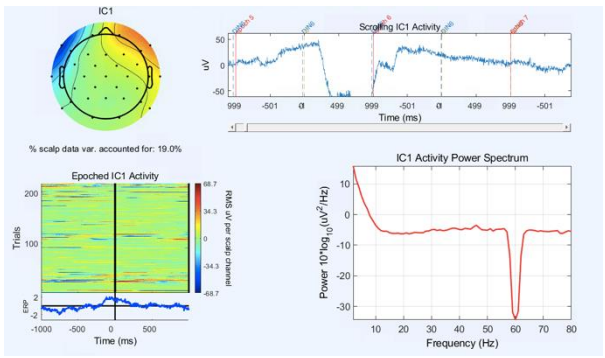
3. Muscle component



4. Heart rate component



5. Blink component



6. Lateral eye movement component

Figure 1. Examples of each type of component and the output used to determine its status as artifact or brain. Each component output contains: a topography, the waveform, epoched activity, and the activity power spectrum. The spikes at 60 HZ are where line noise has or has not been removed and can be ignored.

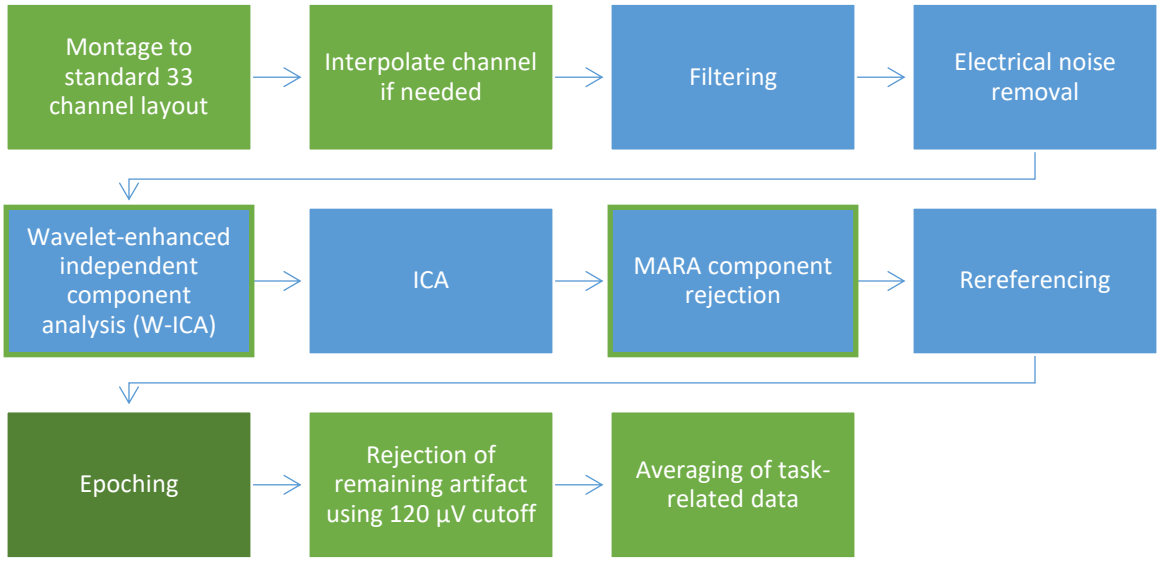


Figure 2. Automated pipeline (HAPPE/BEAPP) processing diagram. Blue boxes are done by HAPPE though BEAPP. Green boxes are done manually in EEGLAB. Epoching is done manually for resting-state data but by HAPPE for task-related data. Steps where decisions are made by HAPPE are marked by a green outline.

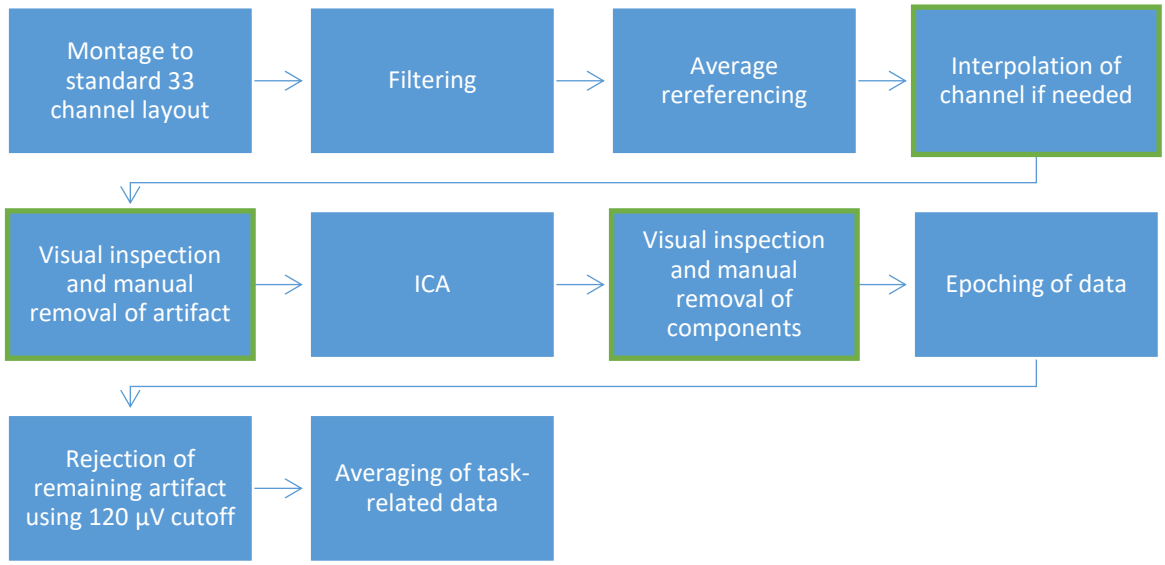


Figure 3. Manual pipeline processing diagram. All steps are done in either EEGLAB or Matlab. Steps where decisions are made by a researcher are marked by a green outline.

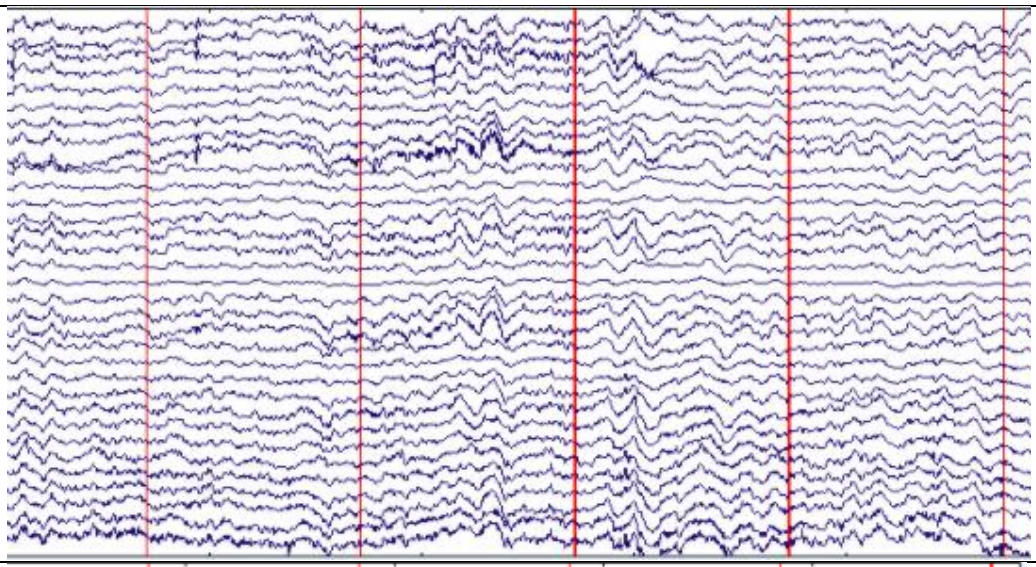
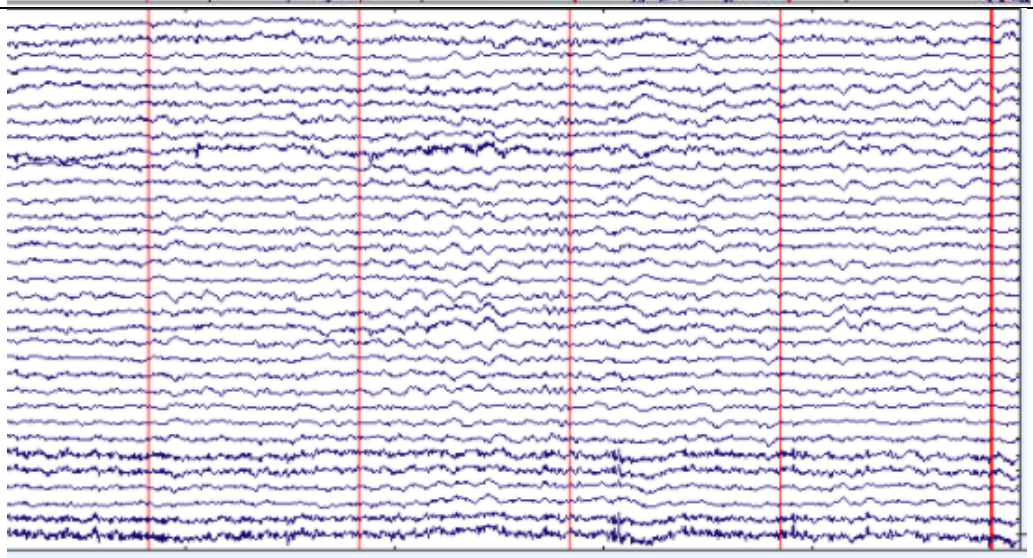
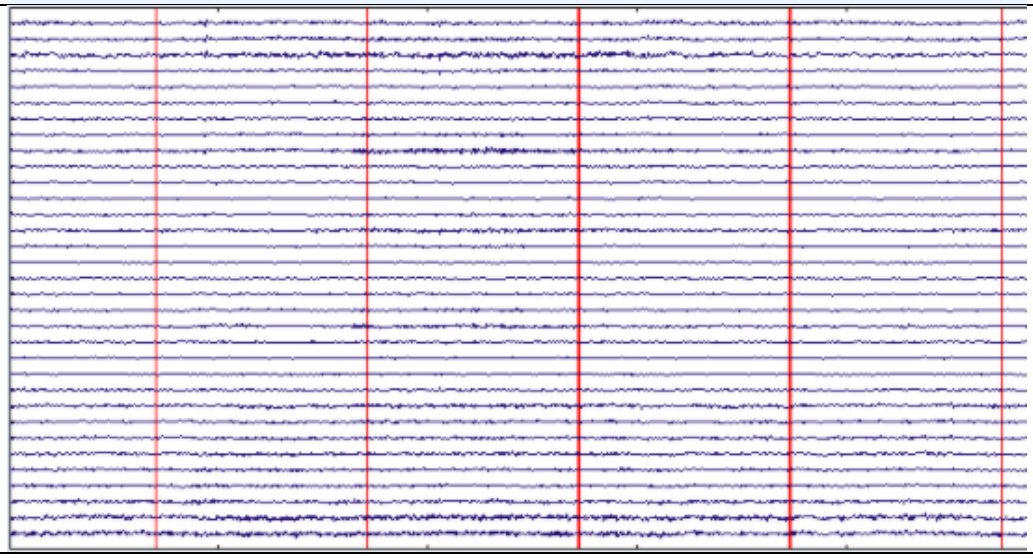
Filtered Data before Cleaning	
Manually Cleaned Data	
HAPPE Cleaned Data	

Figure 4. A representative example of 5 seconds of data before undergoing cleaning, after going through the manual processing pipeline, and after going through HAPPE. All segments are scaled at 50 microvolts.

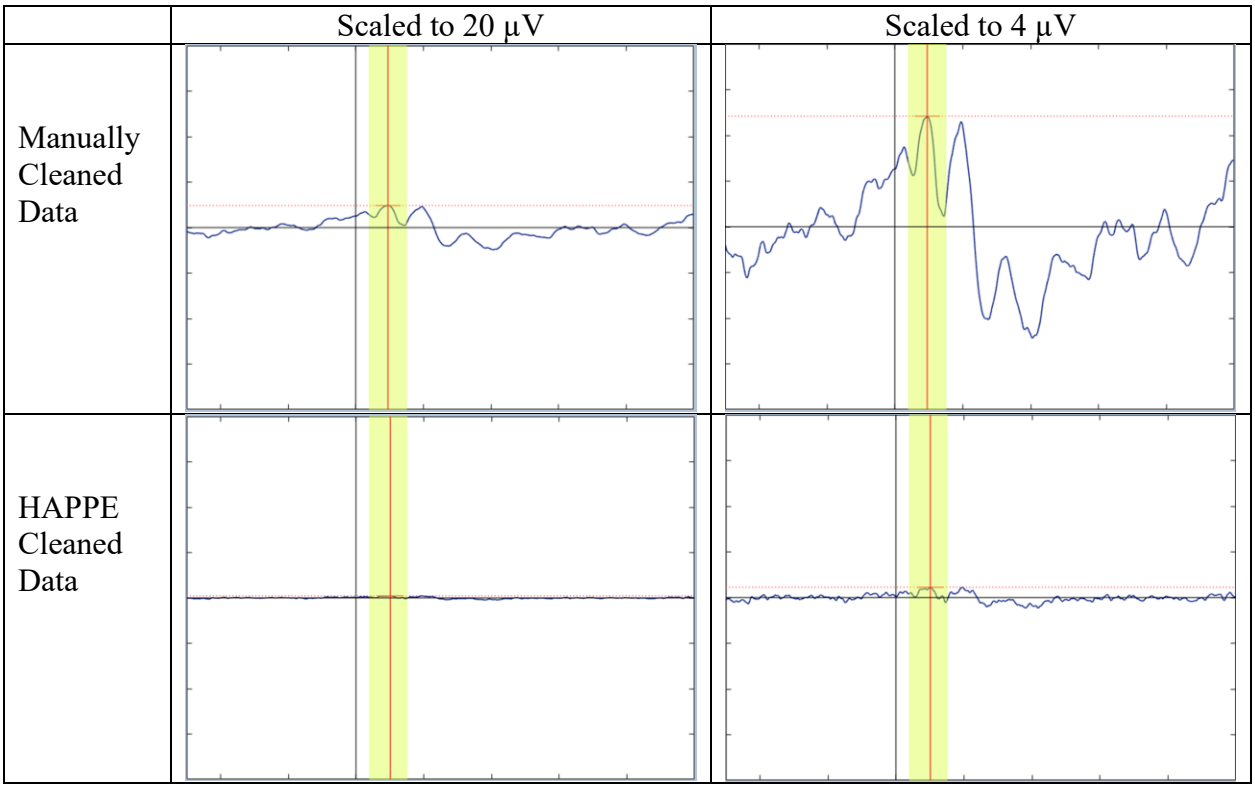


Figure 5. Example ERP from both pipelines scaled to 20 µV and 4 µV. The highlighted section shows the window used for determining the P1 peak (40-150 ms). The red line shows the highest positive peak within the window.

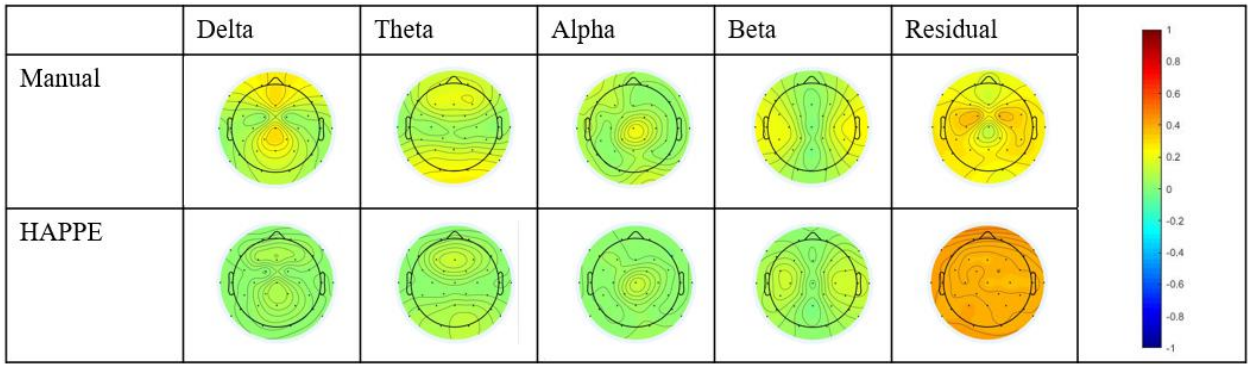


Figure 6. Heat maps showing standardized Beta weights from single channel multiple regressions averaged across simulated subjects. Regressions investigated relationship between pipeline cleaned data and the pure signals for each channel. Scaled from -1 to 1.

References

Bosl, W. J. (2017). *Informatics for EEG biomarker discovery in clinical neuroscience* [ProQuest Information & Learning (US)]. http://search.proquest.com/docview/1872259962/2239EE393B51435FPQ/6

Bridwell, D. A., Rachakonda, S., Rogers, F. S., Pearlson, G. D., & Calhoun, V. D. (2018). Spatiospectral decomposition of multi-subject EEG: Evaluating blind source separation algorithms on real and realistic simulated data. *Brain Topography*, *31*(1), 47–61. https://doi.org/10.1007/s10548-016-0479-1

Castellanos, N. P., & Makarov, V. A. (2006). Recovering EEG brain signals: Artifact suppression with wavelet enhanced independent component analysis. *Journal of Neuroscience Methods*, *158*(2), 300–312. https://doi.org/10.1016/j.jneumeth.2006.05.033

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

Dickter, C., & Kieffaber, P. (2014). *EEG Methods for the Psychological Sciences*. Sage Publications. https://ou-primo.hosted.exlibrisgroup.com/primo-explore/fulldisplay?docid=TN_wj10.1111%2Fpsyp.12827&context=U&vid=OUNEW&lang=en_US

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Ethridge, L. E., De Stefano, L. A., Schmitt, L. M., Woodruff, N. E., Brown, K. L., Tran, M., Wang, J., Pedapati, E. V., Erickson, C. A., & Sweeney, J. A. (2019). Auditory EEG Biomarkers in Fragile X Syndrome: Clinical Relevance. *Frontiers in Integrative Neuroscience*, *13*. https://doi.org/10.3389/fnint.2019.00060

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

Gabard-Durnam, L. J., Wilkinson, C., Kapur, K., Tager-Flusberg, H., Levin, A. R., & Nelson, C. A. (2019). Longitudinal EEG power in the first postnatal year differentiates autism outcomes. *Nature Communications*, *10*(1), 4188. https://doi.org/10.1038/s41467-019-12202-9

Groppe, D. M., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage*, *45*(4), 1199–1211. https://doi.org/10.1016/j.neuroimage.2008.12.038

Hagerman, R. J., & Hagerman, P. J. (2002). *Fragile X Syndrome: Diagnosis, Treatment, and Research*. Taylor & Francis US.

Haufe, S., & Ewald, A. (2019). A Simulation Framework for Benchmarking EEG-Based Brain Connectivity Estimation Methodologies. *Brain Topography*, *32*(4), 625–642. https://doi.org/10.1007/s10548-016-0498-y

Joyce, C. A., Gorodnitsky, I. F., & Kutas, M. (2004). Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology*, *41*(2), 313–325. https://doi.org/10.1111/j.1469-8986.2003.00141.x

Jung. (2000). *Removing electroencephalographic artifacts by blind source separation*. https://onlinelibrary.wiley.com/doi/abs/10.1111/1469-8986.3720163?sid=nlm%3Apubmed

Kaur, C., & Singh, P. (2015). EEG artifact suppression based on SOBI based ICA using wavelet thresholding. *2015 2nd International Conference on Recent Advances in Engineering Computational Sciences (RAECS)*, 1–4. https://doi.org/10.1109/RAECS.2015.7453319

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

Levin, A. R., Méndez Leal, A. S., Gabard-Durnam, L. J., & O'Leary, H. M. (2018). BEAPP: The Batch Electroencephalography Automated Processing Platform. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00513

Makeig, S., Bell, A. J., Jung, T.-P., & Sejnowski, T. J. (1996). Independent Component Analysis of Electroencephalographic Data. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8* (pp. 145–151). MIT Press.

http://papers.nips.cc/paper/1091-independent-component-analysis-of-electroencephalographic-data.pdf

Mamun, Md., Al-Kadi, M., & Marufuzzaman, Mohd. (2013). Effectiveness of Wavelet Denoising on Electroencephalogram Signals. *Journal of Applied Research and Technology*, *11*(1), 156–160. https://doi.org/10.1016/S1665-6423(13)71524-4

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.

Pierce, L. J., Thompson, B. L., Gharib, A., Schlueter, L., Reilly, E., Valdes, V., Roberts, S., Conroy, K., Levitt, P., & Nelson, C. A. (2019). Association of Perceived Maternal Stress During the Perinatal Period With Electroencephalography Patterns in 2-Month-Old Infants. *JAMA Pediatrics*, *173*(6), 561. https://doi.org/10.1001/jamapediatrics.2019.0492

Pontifex, M. B., Miskovic, V., & Laszlo, S. (2017). Evaluating the efficacy of fully automated approaches for the selection of eyeblink ICA components. *Psychophysiology*, *54*(5), 780–791. https://doi.org/10.1111/psyp.12827

Sahin, M., Jones, S. R., Sweeney, J. A., Berry-Kravis, E., Connors, B. W., Ewen, J. B., Hartman, A. L., Levin, A. R., Potter, W. Z., & Mamounas, L. A. (2018). Discovering translational biomarkers in neurodevelopmental disorders. *Nature Reviews Drug Discovery*, *18*(4), 235. https://doi.org/10.1038/d41573-018-00010-7

Thigpen, N., Kappenman, E., & Keil, A. (2018). *Assessing the internal consistency of the event-related potential: An example analysis*. 36.

van Drongelen, W. (2007). *Signal Processing for Neuroscientists—Introduction to the Analysis of Physiological Signals*. Elsevier. https://app-knovel-com.ezproxy.lib.ou.edu/web/toc.v/cid:kpSPNIAPS1/viewerType:toc//root_slug:signal-processing-for?kpromoter=marc

Walters-Williams, J., & Li, Y. (2011). Performance Comparison of Known ICA Algorithms to a Wavelet-ICA Merger. *Signal Processing: An International Journal*, *5*(3), 80–92.

Webb, S. J., Bernier, R., Henderson, H. A., Johnson, M. H., Jones, E. J. H., Lerner, M. D., McPartland, J. C., Nelson, C. A., Rojas, D. C., Townsend, J., & Westerfield, M. (2015). Guidelines and best practices for electrophysiological data collection, analysis and reporting in autism. *Journal of Autism and Developmental Disorders*, *45*(2), 425–443. https://doi.org/10.1007/s10803-013-1916-6

Wilkinson, C. L., Gabard-Durnam, L. J., Kapur, K., Tager-Flusberg, H., Levin, A. R., & Nelson, C. A. (2019). Use of Longitudinal EEG Measures in Estimating Language Development in Infants With and Without Familial Risk for Autism Spectrum Disorder. *Neurobiology of Language*, *1*(1), 33–53. https://doi.org/10.1162/nol_a_00002

Wilkinson, C. L., Levin, A. R., Gabard-Durnam, L. J., Tager-Flusberg, H., & Nelson, C. A. (2019). Reduced frontal gamma power at 24 months is associated with better expressive language in toddlers at risk for autism. *Autism Research*, *12*(8), 1211–1224. https://doi.org/10.1002/aur.2131

Winkler, I., Brandl, S., Horn, F., Waldburger, E., Allefeld, C., & Tangermann, M. (2014). Robust artifactual independent component classification for BCI practitioners. *Journal of Neural Engineering*, *11*(3), 035013. https://doi.org/10.1088/1741-2560/11/3/035013

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. https://doi.org/10.1186/1744-9081-7-30

Zhivomirov, H. (2020). *Pink, Red, Blue and Violet Noise Generation with Matlab*. https://www.mathworks.com/matlabcentral/fileexchange/42919-pink-red-blue-and-violet-noise-generation-with-matlab