Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG

Brenton W. McMenamin a,⁎,1, Alexander J. Shackman b,⁎,1, Jeffrey S. Maxwell c, David R.W. Bachhuber b, Adam M. Koppenhaver b, Lawrence L. Greischar b, Richard J. Davidson b,⁎

a Department of Psychology, Center for Cognitive Science, University of Minnesota, Twin Cities, Minneapolis, MN, USA
b Laboratory for Affective Neuroscience, Weisman Laboratory for Brain Imaging and Behavior, University of Wisconsin, Madison, WI, USA
c U.S. Army Research Laboratory, Aberdeen, MD, Laboratory for Affective Neuroscience, University of Wisconsin, Madison, WI, USA

⁎ Corresponding authors. Laboratory for Affective Neuroscience, University of Wisconsin, Madison, 1202 West Johnson Street, Madison, WI 53706, USA.
E-mail addresses: mcmen020@umn.edu (B.W. McMenamin), shackman@wisc.edu (A.J. Shackman), rjdavids@wisc.edu (R.J. Davidson).
1 Brenton McMenamin and Alexander Shackman are co-first authors on this manuscript.

ARTICLE INFO

Article history:
Received 8 July 2009
Revised 5 October 2009
Accepted 6 October 2009
Available online 13 October 2009

ABSTRACT

Muscle electrical activity, or “electromyogenic” (EMG) artifact, poses a serious threat to the validity of electroencephalography (EEG) investigations in the frequency domain. EMG is sensitive to a variety of psychological processes and can mask genuine effects or masquerade as legitimate neurogenic effects across the scalp in frequencies as low as the alpha band (8–13 Hz). Although several techniques for correcting myogenic activity have been described, most are subjected to only limited validation attempts. Attempts to gauge the impact of EMG correction on intracerebral source models (source “localization” analyses) are rarer still. Accordingly, we assessed the sensitivity and specificity of one prominent correction tool, independent component analysis (ICA), on the scalp and in the source-space using high-resolution EEG. Data were collected from 17 participants while neurogenic and myogenic activity was independently varied. Several protocols for classifying and discarding components classified as myogenic and non-myogenic artifact (e.g., ocular) were systematically assessed, leading to the exclusion of one-third to as much as three-quarters of the variance in the EEG. Some, but not all, of these protocols showed adequate performance on the scalp. Indeed, performance was superior to previously validated regression-based techniques. Nevertheless, ICA-based EMG correction exhibited low validity in the intracerebral source-space, likely owing to incomplete separation of neurogenic from myogenic sources. Taken with prior work, this indicates that EMG artifact can substantially distort estimates of intracerebral spectral activity. Neither regression- nor ICA-based EMG correction techniques provide complete safeguards against such distortions. In light of these results, several practical suggestions and recommendations are made for intelligently using ICA to minimize EMG and other common artifacts.

© 2009 Elsevier Inc. All rights reserved.

Peri-cranial muscle or myogenic activity is distinguished by its relatively high amplitude, broad spectral and often anatomical distributions, and exquisite sensitivity to a variety of psychologically interesting processes. Consequently, it poses a serious inferential hazard for any electroencephalography (EEG) investigation in the frequency domain. This artifact can compromise sensitivity by masking effects of interest or diminish specificity by masquerading as a neurogenic effect. Although several techniques have been developed to correct myogenic activity (Shackman et al., 2009), many have been subjected to only limited attempts at validation, rendering their utility questionable. In particular, the sensitivity and specificity of one prominent electromyography (EMG) correction tool, independent component analysis (ICA), remains unclear.

Several properties of cranial EMG are collectively responsible for its pernicious effects. First, EMG is sufficiently sizable to perturb all classic EEG bands. Goncharova, McFarland, Vaughan, and Wolpaw (2003) report myogenic artifact reliably as low as 2 Hz, making even the widely used alpha band (8–13 Hz) vulnerable to muscle artifacts (Lee and Buchsbaum, 1987; Willis et al., 1993; Van Boxtel, 2001). Second, EMG can often be detected across the entire scalp (Goncharova et al., 2003) due to volume conduction of myogenic activity independently generated by muscles across the head, face and neck. Anterior electrodes are sensitive to facial muscles, such as the corrugator supercilii and frontalis; lateral electrodes are sensitive to the muscles of mastication, masseter and temporalis; and posterior electrodes are sensitive to muscles at the intersection of the cranium, spine, and torso, such as occipitalis (Supplementary Fig. 1). Third, EMG is temporally confounded with a variety of experimental manipulations. Facial EMG, in particular, is sensitive to numerous cognitive and affective processes, including cognitive load (Cohen et al., 1992; Waterlink and van Boxtel, 1994), facial mimicry (Dimberg et al., 2000), vocalization (Brooker and Donald, 1980), and induced emotional states (Borden et al., 1991; Coan and Allen, 2003; Bradley et al., 2001).
EMG also exhibits less stereotypy than other biological artifacts. Ocular and cardiac artifacts, for example, arise from fixed sources and do not qualitatively differ across individuals. EMG, however, arises from the activity of spatially distributed, functionally independent muscle groups, with distinct topographic and spectral signatures (Goncharova et al., 2003). For instance, *frontalis* activity peaks around 25 Hz, whereas *temporalis* generates a low peak around 20 Hz and broad plateau centered around 40–80 Hz (Goncharova et al., 2003).

The spectral composition of myogenic activity also varies as a function of contraction intensity (Goncharova et al., 2003) and fatigue (Chung et al., 2002). This is compounded by the fact that the relative contributions of each muscle group to the cranial EMG vary substantially across elicitors and individuals (Tassinary et al., 2007) and may differ somewhat between spontaneous and voluntary contractions (Davidson et al., 2004; Morecraft and Tanji, 2009).

Given the inferential hazards posed by EMG, there is substantial interest in developing tools to remedy myogenic artifact. Generally, EEG artifacts can be addressed in one of two ways, rejecting contaminated epochs of data or filtering artifact from neurogenic activity. Rejection-based techniques are most appropriate for transient artifacts, such as blinks, that influence a small portion of the data record. The protracted time-course of EMG makes such a solution impractical—the high data rejection rate would markedly erode the signal to noise ratio (Jung et al., 2000b; Talsma, 2008). Moreover, because EMG covaries with cognitive and affective processes of interest, rejecting data laden with EMG artifact would likely entail discarding some of the most interesting, discriminative periods of neural activity (Davidson et al., 1990). For these reasons, EMG mandates the use of filtering techniques capable of separating myogenic from neurogenic activity. Given marked individual differences in the spectral and anatomical profile of myogenic activity, spatial or spectral filters or templates that are fixed across subjects cannot be fruitfully applied to the correction of EMG artifact (cf. Frank and Frishkoff, 2006; Ille et al., 2002; Koskinen and Vartiainen, 2009). Instead, a more flexible approach is required.

One class of techniques for correcting EMG artifact employs variants of the general linear model (GLM), such as multiple regression and ANCOVA, to identify and discard variance in a neurogenic band of interest (e.g., alpha) that is predicted by activity in an *a priori* EMG band (e.g., 70–80 Hz). The advantage of this technique is that it does not require dedicated EMG channels or manual intervention, and, by performing separate corrections at each site, can accommodate individual differences in artifact topography. These GLM-based techniques have proven quite popular (Allen et al., 2004; Davidson et al., 2000), and McMenamin et al. (2009) have shown that at least one variant of this technique displays adequate sensitivity and specificity on the scalp.

Despite these strengths, GLM-based EMG correction techniques suffer from two key limitations. First, they do not permit reconstruction of the EEG time-series. Thus, while useful for investigations of tonic ("resting") and induced changes in the EEG spectra (e.g., Lutz et al., 2004), GLM-based tools cannot be applied to studies relying on event-related spectral perturbation (ERSP) measures (Onton et al., 2006). Second, McMenamin et al. (2009) reported that applying GLM-based techniques to source-estimated EEG in a voxelwise manner is not appropriate if the data has been corrupted by EMG prior to localization. Source-estimation ("localization") is a technique that estimates neurogenic signals from scalp EEG recordings. This is achieved by developing a forward-model that uses the biophysics of the EEG (e.g., the spatial filtering imposed on neurogenic signals by the cerebrospinal fluid, skull and scalp) to predict signals on the scalp given a particular neural generator. Source-estimation occurs when this model is inverted and used to estimate a probable neural generator given scalp-recorded signals (Pizzagalli, 2007). McMenamin et al. (2009) speculated that the EMG-contaminated data cannot be properly localized because a solution space that only allows intra-cranial dipoles cannot account for a scalp-recording that contain both intra-cranial (neurogenic) and extra-cranial (myogenic) sources. The resulting attempt at localization will be corrupted and the true neurogenic solution rendered unrecoverable. Removing extra-cranial source activity from the data prior to source-estimation may circumvent this problem. Unfortunately, this is not possible using GLM-based techniques because they cannot reconstruct the artifact-free EEG time-series.

A second class of EMG correction methods employs ICA to decompose the EEG time-series into a set of temporally independent components (Delorme, 2007a; James and Hesse, 2005; Onton et al., 2006; Onton and Makeig, 2006; Makeig et al., 2004). Components are inspected visually for the presence of artifact and those classified as predominantly artificial (e.g. EMG or blinks) are discarded. Like GLM-based correction techniques, ICA does not require dedicated EMG channels and can accommodate variation across the scalp. More importantly, unlike GLM-based techniques, ICA allows reconstruction of the artifact-filtered time-series, which can then be used for analyses employing averaging, spectral decomposition, or source modeling.

Although ICA shows great promise as a tool for correcting EMG and other kinds of biological artifact (e.g. Jung et al., 2000a), attempts to assess its validity have been limited. Many validation studies have relied on small samples of *ad hoc* data (Delorme et al., 2007b; Jung et al., 2000a,b; Wallstrom et al., 2004; Flexer et al., 2005; Ting et al., 2006; Frank and Frishkoff, 2006). While others have used simulations (e.g., Crespo-Garcia et al., 2008; De Clercq et al., 2005; Delorme et al., 2007a,b; Fitzgbibbon et al., 2007; Frank and Frishkoff, 2006; Romero et al., 2008). In simulations, real or artificial EMG activity is mathematically "injected" into otherwise artifact-free EEG. The potential problem with this strategy is that the assumptions underlying injection (e.g., the degree of temporal and spatial correlation with neurogenic signals) may not characterize real EMG contamination, potentially limiting external validity and biasing the results in favor of correction techniques founded on similar assumptions (Grouiller et al., 2007; Hoffmann and Falkenstein, 2009).

Accordingly, the major aim of the present study was to quantitatively assess the quality of EMG artifact correction afforded by ICA. Ideally, validation would quantitatively establish that a technique possesses a high degree of sensitivity (i.e., attenuates myogenic artifact) and specificity (i.e., preserves neurogenic signals) in a reasonably large and varied dataset. This requires data in which the presence or absence of EMG ("ground truth") is definitive or can be reasonably assumed. To this end, the dataset previously employed by McMenamin et al. (2009) for testing the validity of GLM-based correction techniques was reanalyzed using ICA. This had the advantage of facilitating direct comparisons across correction techniques. In this dataset, 128-channel EEG was acquired while neurogenic and myogenic activity were independently varied. Alpha band neurogenic activity was selectively increased or decreased by instructing participants to close or open their eyes, a procedure sometimes termed the “Berger maneuver” (cf. Berger, 1929/1969). Myogenic activity was manipulated by instructing participants to alternately tense and relax their cranial muscles. The sensitivity and specificity of ICA-based EMG correction were then quantitatively assessed in the alpha band using methods similar to those described in our prior report (McMenamin et al., 2009). Several considerations led us to focus on the alpha band. First, it is relatively easy to manipulate neurogenic activity in this frequency. To our knowledge, comparably robust manipulations do not exist for the other classical EEG bands (Niedermeyer, 2005). Second, alpha activity has been among the most widely used spectral indices of neural activity, from the earliest EEG research (Berger, 1929/1969), to contemporary

---

2 Source modeling in the frequency-domain requires phase information in the form of the cross-spectra. Extant GLM-correction techniques operate on estimates of spectral power (squared amplitude) and discard information about the phase of EEG oscillations required to compute the cross-spectra.
investigations of memory (Freunberger et al., 2009; Gevins and Smith, 2000; Hamidi et al., 2009), perception and attention (Romei et al., 2008; Thut and Miniussi, 2009), emotion (Coan and Allen, 2003; Davidson et al., 1990), temperament and individual differences (Carver and Harmon-Jones, 2009; Shackman et al., in press), and psychopathology (Thibodeau et al., 2006; DeRubeis et al., 2008).

The other major aim of this study was to test whether ICA-based techniques constitute a valid EMG correction technique for distributed intracerebral source modeling. Source modeling is an increasingly popular technique for maximizing the anatomical information yielded by scalp-recorded EEG (Pizzagalli, 2007) and the dissemination of commercial and freely available software for performing distributed source localisation, such as Cartool (http://brainmapping.unige.ch/Cartool.htm), EMSE (http://www nguồnignal.com), LORETA-KEY (http://www.unizh.ch/keyinst/) and SPM5 (http://www.fil.ion.ucl.ac.uk/spm/), is likely to accelerate this trend. Furthermore, prior work indicates that EMG correction techniques deemed valid on the scalp do not necessarily confer validity in the intracerebral source-space (McMenamin et al., 2009). Accordingly, ICA-based procedures that proved valid on the scalp were also tested with source solutions estimated using the low-resolution electromagnetic tomography (LORETA) algorithm.

A minor aim of this study was to evaluate the degree to which variation in the protocol for filtering non-myogenic artifacts, such as eye movements, impacts the quality of EMG correction. To date, existing methodological and empirical reports employing ICA provide little direct guidance on the question of which components ought to be discarded (Shackman et al., 2009). Furthermore, despite the fact that ICA requires trained raters to inspect hundreds or even thousands of components for a single high-resolution EEG study (number of components ≈ channels × participants; see Supplementary Method), the reliability of component classification has only rarely been reported (Viola et al., 2009). Without such evidence, poor validity might simply reflect inadequate training or an ambiguous classification protocol. Accordingly, the inter-rater reliability was computed.

Methods

Participants

The dataset consisted of seventeen individuals recruited from the University of Wisconsin–Madison campus (16 female; M = 24.1 years, SD = 7.1) and described in an earlier report assessing the validity of GLM-based EMG correction techniques (McMenamin et al., 2009). Each received US$20 for their participation. Participants provided informed consent in accord with guidelines prescribed by the local Institutional Review Board.

Design

In order to independently manipulate neurogenic and myogenic activity in the alpha band (8–13 Hz), the experiment took the form of a 2 (Eyes Open/Closed) × 2 (Muscles Tense/Relaxed) repeated-measures design. We anticipated that participants would generate greater broad-spectrum power, including increases in alpha power, indicative of reduced neural activity (Allen et al., 2004; Oakes et al., 2004), during the eyes-closed condition. We further expected participants to generate greater alpha power, indicative of increased muscle activity, during the muscles-tense condition. Hereafter, these four conditions are referred to using the following acronyms: Open-Relaxed (OR), Open-Tense (OT), Closed-Relaxed (CR), and Closed-Tense (CT).

Procedure

Procedures were identical to those detailed by McMenamin et al. (2009; see also Bonnett and Arand, 2001; Freeman et al., 2003). In brief, participants were instructed how to properly tense facial muscles at the outset of the session. Frontalis and corrugator muscles were contracted by lifting and squeezing the eyebrows together; masseter and temporalis were contracted by lightly clenching the jaw. EEG was acquired during sixteen 32-second blocks (order counterbalanced; 4 blocks/condition). Participants were continuously monitored via a closed-circuit audio-video circuit and real-time EEG.

EEG acquisition and preliminary reduction

EEG were collected using a 128-channel Geodesic Sensor Net (GSN128; Electrical Geodesics Inc., Eugene, OR) referenced to vertex (Cz) and sampled at 500-Hz (analog anti-aliasing: 0.1–250 Hz). Data reduction used a combination of EEGLAB (Delorme and Makeig, 2004; http://www.sccn.ucsd.edu/eeeglab) and in-house code written for MATLAB (http://www.mathworks.com). A zero-phase 60-Hz notch filter removed line noise from calibrated (μV) data, and bad channels (±100 μV for –20 s) or gross artifacts (±100 μV for >4 channels) were manually identified and rejected. A 0.5–Hz high-pass filter was used to attenuate channel drift and better satisfy ICA’s stationarity assumption (Onton et al., 2006). Such artifacts were rejected to better approximate the subtle contamination of signal that can occur when EMG covaries with an experimental treatment. Removal of non-stereotyped artifact also maximizes the quality of the ICA (Onton et al., 2006).

ICA

Overview

Consistent with other high-resolution EEG studies (Delorme, 2007b), a spatial Principal Components Analysis (PCA) was used to reduce the dimensionality of the EEG from 128 channels to 64 principal components (PCs) prior to performing ICA. This was done in a single step as part of the ICA using the EEGLAB runcia command, implementing the extended Infomax algorithm (Bell and Sejnowski, 1995; Lee et al., 1999).

The primary aim of this study was to assess the validity of ICA for EMG artifact correction. Accordingly, three protocols for the correction of EMG artifact, described below, were investigated. A secondary aim of this study was to investigate the degree to which the quality of EMG correction was dependent on the protocol for removing non-myogenic sources of variance (e.g., ocular artifact, noise components). Consequently, three ICA-based protocols for the correction of non-neurogenic/non-myogenic (NNNM) components, described below, 5 there were two reasons for doing so, aside from computational and classification efficiency. First, preliminary inspection of the 128 components extracted from the native electrode array indicated over-fitting, evidenced by fragmentation of artifacts across components (Li et al., 2007; Lawrence and Hancock, 1995). By contrast, exploratory analyses (not reported) showed that reduction to 48 or fewer PCs prior to ICA led to under-fitting, evidenced by cross-contamination of EEG, physiological artifacts, and noise (Fava and Velicer, 1996). Second, it has been suggested (Onton et al., 2006; Romero et al., 2008) that Infomax ICA requires a minimum of 20 × 2 × 2 samples, where c is the number of channels or, equivalently, PCs. For the native electrode array, this would require 20 × 128² = 327680 samples, whereas we had at most 16 blocks × 32 × 500 Hz = 256000 samples. Reducing the model order by half allowed us to satisfy this criterion (20 × 64² = 81920 samples). Quantitative estimates of “model order,” the number of components required to adequately but parsimoniously describe the data, suggested that this was sufficient (see Supplementary Method and Results). A viable alternative to PCA-based dimension reduction is to simply prune the number channels submitted to ICA (Milne et al., 2009), at the potential expense of spatial resolution (Srinivasan et al., 1998; Michel et al., 2004).
were also examined. The quality of EMG artifact correction was evaluated for all nine factorial combinations of the EMG and NNNM protocols. Following removal of the relevant components, the filtered 128-channel time-series were reconstructed. Subsequent analyses used only the 107 cephalic electrodes. Exploratory analyses (not reported) using the complete 128-channel array indicated worsened performance when the peri-cephalic electrodes on the face and along the posterior edge of the array were retained (Supplementary Fig. 1).

Component classification

Using in-house code, the variance accounted for by each of the ICs was assessed. By default, components that individually accounted for <0.2% of the variance were categorized as Low-Variance. In cases where the determination was ambiguous, exceptions were made. As described in the Supplementary Method and Results (Supplementary Figs. 2–13), the remaining components were classified as neurogenic (Neuro), myogenic (Myo), a combination of the two sources (Neuro-Dominant or Myo-Dominant), or artifact (residual Gross or Ocular). Components classified as Gross included reference, ground, electrocardiographic, and alternating current artifacts. Components that met the minimum variance criterion, but proved impossible to unambiguously categorize were classified as Noise. Classifications were made by two raters based on inspection of the component’s time-series, power spectrum, and topography. When disagreements occurred, final classification was by consensus. Interrater reliability, assessed prior to consensus using Krippendorff’s alpha (Hayes and Krippendorff, 2007), was excellent, α = 0.98 (for details, see Supplementary Method). We urge investigators with a practical interest in using ICA for artifact reduction to examine our detailed classification protocol (see Supplementary Method and Results).

Correction of EMG artifact

Three different ICA-based protocols for removing myogenic artifact were assessed. The Minimal-EMG protocol discarded only those components that contained clear EMG activity in the absence of any identifiable neurogenic activity (i.e., rejected Myo components). The Intermediate-EMG protocol expanded this definition to include mixed components in which myogenic activity was more prominent than neurogenic activity (i.e. rejected both Myo and Myo-Dominant components). The Maximal-EMG protocol rejected any component containing myogenic signal, even if myogenic activity was less prominent than neurogenic activity (i.e., rejected Myo, Myo-Dominant, and Neuro-Dominant components). Thus, the Maximal-EMG protocol performs the strictest filtering of the data, at the potential expense of discarding neurogenic signals of interest.

Filtering of non-neurogenic/non-myogenic (NNNM) signals

To provide a specific test of ICA’s utility for removing EMG artifact, it is necessary to first filter signals that are not clearly neurogenic or myogenic (cf. McMenamin et al., 2009). However, the choice of which components to remove is subjective and has a marked impact on the number of components and percentage of variance retained (see Results). Accordingly, three different ICA-based protocols for filtering non-neurogenic/non-myogenic signals were used. The Minimal-NNNM protocol made the fewest assumptions, filtering only those components that were explicitly classified as Gross or Ocular artifact, similar to the method used in McMenamin et al. (2009). The Intermediate-NNNM protocol made the additional assumption that components categorized as Noise do not contain meaningful neurogenic signal and filtered them as well. The Maximal-NNNM protocol further assumed that Low-Variance components do not contain significant neurogenic signal and filters them as well.

Scalp spectral power density estimation

Following reconstruction of the filtered time-series, epochs with residual artifact (i.e., deviations exceeding ±200 μV for more than half an epoch or variance exceeding 1000 μV²) or flat channels (epoch variance less than 0.25 μV²) were automatically rejected (Delorme et al., 2007a,b). After residual artifact-rejection, the rejected channels were interpolated with a spherical spline when at least one neighboring electrode was usable (Greischar et al., 2004). Data were re-referenced to an average montage (Davidson et al., 2000; Dien, 1998) and spectral power density (μV²/Hz) estimated for the alpha (8–13 Hz) band using Welch’s (1967) method on sliding Hanning-windowed epochs (50% overlap). Estimates were log₁₀ transformed to normalize the distribution (Allen et al., 2004; Gasser et al., 1982).

LORETA distributed source current density modeling

The modeling of distributed sources from scalp-recorded electrical activity was performed using previously published procedures (McMenamin et al., 2009; Shackman et al., in press) via in-house MATLAB code implementing the LORETA algorithm (Pascual-Marqui et al., 1994) to estimate intracerebral current density. LORETA has undergone extensive cross-modal validation (reviewed in Shackman et al., in press; Pizzagalli, 2007).

An inverse operator distributed with the LORETA-Key software suite (Pascual-Marqui, 1999; http://www.unizh.ch/keyinst/; λ = 10−5) was used to generate three-dimensional intracerebral current density estimates (A/m²) from cross-spectra calculated using the artifact-free Hanning-windowed epochs from the scalp analyses. The forward-model is a 3-shell spherical head model using 107 cephalic EEG electrodes (Shackman et al., in press). The source-space is normalized to the Montreal Neurological Institute’s probabilistic MRI anatomical template (i.e., MNI305; Evans et al., 1993; Collins et al., 1994), restricted to the cerebral gray matter, hippocampi, and amygdalae on a 7-mm³ isotropic lattice. LORETA source-estimates were log₁₀-transformed prior to analysis (Thatcher et al., 2005). Results are displayed on the rendered canonical brain distributed with LORETA-Key.

Analytic strategy

Overview

A valid correction technique should render EMG-contaminated data statistically equivalent to data collected under the same conditions in the absence of myogenic artifact (Frank and Frishkoff, 2006; Debener et al., 2007; Freyer et al., 2009). Accordingly, each combination of the EMG correction and NNNM filtering protocols was evaluated in terms of its (i) sensitivity, the attenuation of myogenic artifact (i.e., Tense vs. Relaxed) in the alpha band, (ii) specificity, the preservation of neurogenic effects (i.e., alpha-blocking: Eyes-Closed vs. Eyes-Open) in the alpha band, and (iii) the degree to which each protocol introduced correction artifacts, artificial effects generated by the correction. Sensitivity and specificity were assessed using regions of interest (ROIs) defined by the areas of peak myogenic and neurogenic activation, respectively. An ROI approach was used to constrain the number of comparisons in both scalp and LORETA source-space analyses. Only those filtering protocols that proved sufficiently valid on the scalp were assessed with LORETA. To permit a direct comparison of ICA- and GLM-based EMG correction techniques, key analyses reported in McMenamin et al. (2009) were recomputed using the identical validation techniques used here. These analyses are detailed in the Supplementary Method and Results.
Sensitivity

On the scalp, a myogenic ROI was created for each of the three NNNM filters using electrodes exhibiting a significant ($p<0.05$) myogenic effect (OR-OT). Channels that were situated at the edge of the 107-channel electrode-array, were spatially discontiguous (i.e., lacked at least one nearest neighbor meeting the significance criterion), or also met the inclusion criteria for the neurogenic (i.e., specificity) ROI were excluded (range: 8–15 electrodes; many located at the posterior base of the array). The effect of the spatial contiguity criterion was minor, resulting in a single electrode being dropped. In the LORETA source-space, myogenic ROIs were created by identifying voxels in the OR-OT contrast with $p<0.001$, using a cluster-extent threshold to correct for multiple comparisons (Nichols and Holmes, 2002; Shackman et al., in press).

Using the resulting ROIs, the degree to which each EMG correction protocol attenuated the myogenic contrast (i.e., EMG-corrected OR-OT vs. 0) was tested. Additional contrasts tested the degree to which each EMG-correction protocol removed myogenic effects using double differences that compared three EMG-corrected contrasts of interest and their uncorrected, artifact-free analogs. ICA's ability to correct EMG artifact that negatively covaried with neurogenic signals was tested using the (EMG-corrected OT-CT) contrast. The amount of EMG artifact surviving each correction was indexed using median and peak ROI $t$-values, viewed as indices of typical and "worst-case" correction, respectively. Significant $t$-tests for these contrasts indicate that the EMG-corrected EEG signals deviate from their artifact-free analogs, evidence of poor sensitivity.

Conversely, failure to reject the null hypothesis does not indicate the absence of residual myogenic activity. In order to rigorously test whether the EMG-corrected contrasts were significantly equivalent to artifact-free data, the Westlake–Schuirmann test (Seaman and Serlin, 1998) was employed as a follow-up test to non-significant contrasts. Sometimes termed the two one-sided tests (TOST) method, a number of fields (e.g., the US Food and Drug Agency; Department of Health and Human Services, 2001) consider TOST the gold standard for testing statistical equivalence. The null hypothesis for TOST is that the mean difference lies outside of the range $[-\epsilon, \epsilon]$, where $\epsilon$ is an a priori error tolerance. To reject the null (i.e., demonstrate significant equivalence) for $\alpha=0.05$, one must demonstrate that the 90th-percentile confidence interval of the mean difference between the artifact-free and EMG-corrected data lies completely within the interval $[-\epsilon, \epsilon]$. Following our prior report (McMenamin et al., 2009), $\epsilon$ was set to 0.5 standard deviations of the artifact-free contrast (i.e., OR for the OR-OT contrast, OR-CR for positively/negatively covarying contrasts).

Specificity

Neurogenic ROIs were generated by thresholding the neurogenic contrast (OR-CR). Owing to the large size of this effect ($p<0.001$ at all electrodes), it proved useful to threshold the contrast using a percentile approach. This had the advantage of creating ROIs that were similar in size to the myogenic ROIs used to interrogate sensitivity. On the scalp, this entailed selecting the upper tercile of channels. As before, channels that were situated on the edge of the array, were not spatially contiguous, or met the inclusion criteria for the myogenic (i.e., sensitivity) ROI were excluded. The latter two criteria led us to drop one electrode. In the source-space, voxels with absolute $t$-values in the upper tercile for the OR-CR contrast were selected.

As with the sensitivity analysis, $t$-tests and follow-up TOSTs were used to test whether neurogenic effects were preserved following EMG correction. First, to test the impact of correction per se on neurogenic activation, (EMG-corrected OR-CR) was compared against (uncorrected OR-CR). Second, to test correction's impact on negatively covarying neurogenic and myogenic signals, (EMG-corrected OT-CR) was compared to (uncorrected OR-CR). Third, to test correction's impact on positively covarying signals, (EMG-corrected OR-CR) was compared to (uncorrected OR-CR). The $\epsilon$ error tolerance for TOST follow-ups was defined using the uncorrected OR-CR contrast.

Correction artifacts

To investigate the degree to which protocols generated artificial results, two kinds of tests were conducted. The first test determined whether correction of the EMG-contaminated myogenic contrast produced artificial effects in the neurogenic ROI (i.e., EMG-corrected OR-OT). The second test assessed whether correction of the EMG-free neurogenic contrast yielded artificial effects in the myogenic ROI (i.e., [EMG-corrected OR-CR] − [uncorrected OR-CR]). The presence of artifactual effects was assessed using the same logic as the sensitivity and specificity tests.

Fig. 1. Alpha-band contrasts prior to correction. Topographic maps depict spline-interpolated $t$-maps for each condition-contrast (columns) and non-neurogenic/non-myogenic (NNNM) artifact filter (rows). There were four conditions, reflecting the factorial manipulation of myogenic (Muscles: Relaxed, Tensed) and neurogenic activity (Eyes: Open, Closed). Contrasts were computed to isolate myogenic (OR-OT), neurogenic (OR-CR), positively-covarying (OT-CR), and negatively-covarying (OR-CT) activity. Note the least extreme values for the myogenic contrast (OR-OT; first column).
Performance ratings
For each contrast and correction protocol, sensitivity was rated as: 
- poor ($p_{\text{median} \cdot t\text{-test}} < 0.05$ or median $p_{\text{TOST}>0.05}$), 
- questionable (0.10 > $p_{\text{median} \cdot t\text{-test}} > 0.05$ or median $p_{\text{TOST}<0.05}$), 
- adequate ($p_{\text{median} \cdot t\text{-test}} > 0.10$ and median $p_{\text{TOST}<0.05}$) or 
- excellent ($p_{\text{peak} \cdot t\text{-test}} > 0.05$ and all $p_{\text{TOST}<0.05}$). 
Thus, the significance of the peak t-test was only considered in cases where a particular EMG-correction protocol showed evidence of adequate or excellent sensitivity using the median-based tests.

Results

Scalp

Effects prior to EMG correction
Visual inspection indicated that the topography of the four alpha-band contrasts was similar across the three NNNM protocols (Fig. 1).

Myogenic ROI
Consistent with expectation, scripted muscle tensing increased alpha power near facial muscles at midline-, left- and right-frontal electrodes. The myogenic contrast (OR-OT) was significant at 34–39 anterior electrodes (Fig. 2) with qualitatively similar peak locations across the three NNNM filters. The median t-scores ($t = −2.30$ to $−2.44$, $p < 0.04$; $t_{0.05} = 0.24$ to 0.27) and extreme t-scores ($t = −2.82$ to $−3.21$, $p < 0.01$; $t_{0.05} = 0.33$ to 0.39) were similar across the three protocols, indicating that the degree of EMG contamination was also similar. This contrast was used to form the Myogenic ROIs (Fig. 3), resulting in clusters of 23–25 contiguous anterior electrodes, extending to mid-frontal and fronto-central leads (e.g., AF7, F2, F3, F5, F7, FC2, FC3, FT7, T7).

Neurogenic ROI
Consistent with expectation, the Berger maneuver (OR-CR) altered power at all electrodes ($t = −4.37$, $p < 0.001$). The peak difference occurred at midline parietal sites (Fig. 1), and defined the neurogenic ROIs (Fig. 3). The three neurogenic ROIs contained 21–26 contiguous electrodes (e.g., Pz, P1, P2, P3, P5, P7, P9, POz, PO3, PO4, PO8, Oz, CPz) with comparable median ($t = −7.35$ to $−7.49$, $p < 0.001$; $t_{0.05} = 0.77$ to 0.78) and extreme t-scores ($t = −8.85$ to $−9.12$, $p < 0.001$; $t_{0.05} = 0.83$ to 0.84), indicating that the strength of neurogenic effect was minimally affected by the choice of NNNM protocol.

Covarying effects. In the absence of EMG correction, myogenic activity distorted the magnitude of neurogenic effects in the alpha band. For instance, when changes in EMG and EEG negatively covaried (OT-CR), the effect remained significant at all electrodes; but, the magnitude of the neurogenic effect was significantly attenuated at 34–39 electrodes relative to the uncontaminated effect (OR-CR). This resulted in a slightly shifted topography that deemphasizes activation at anterior sites. Notably, significant attenuation was present at posterior electrodes far removed from the area of peak myogenic artifact (Fig. 4). A parallel, albeit non-significant, pattern of amplification occurred at anterior sites (not shown) when changes in EMG and EEG positively covaried (OR-CT). In this case, significant attenuation was only observed at a small number of parietal sites (Fig. 5).

Descriptive statistics for ICA

Classification. Fig. 6 depicts the relative frequency and percentage of scalp variance predicted by each class of components. A similar pattern was found using means. Visual inspection indicates that the most frequent classification, comprising about one-fifth of the total, was Noise (Fig. 6A). The frequency of the other classifications was somewhat smaller, but similar to one another (11–15%), Neuro-Dominant and Gross Artifact components were infrequent (2–3%). The frequency of the Myogenic and Myogenic-Dominant components

**Fig. 2.** Myogenic contrast (OR-OT) after EMG correction. Topographic maps depict thresholded p-values at each electrode after applying each method of non-neurogenic/non-myogenic (NNNM) artifact filtering and ICA-based EMG correction. Negative values are depicted in blue (dark-blue: $p < 0.05$; light-blue: $p < 0.10$; green: $p > 0.10$). Note that the row labeled “None” depicts the thresholded OR-OT contrast from Fig. 1.
showed marked variability and was positively skewed. That is, a few individuals displayed many more of these two EMG-related components than the remainder of the group.

Collectively, the Neurogenic (25%) and Ocular components (21%) accounted for nearly half of the variance in scalp electrical activity (Fig. 6B). Myogenic, Myogenic-Dominant, and Noise components each accounted for another 8–10%, while the Neurogenic-Dominant, Low Variance, and Gross Artifact components each accounted for less than 2% of the variance. There was substantial variability and positive skew in the percentage of variance accounted for by several artifact components, particularly Ocular and Myogenic-Dominant. It is worth emphasizing that a sizable proportion of the variance—equal to that accounted for by the purely myogenic component—was predicted by the more heterogeneous Myogenic-Dominant component.

Validity of ICA-based EMG correction

As summarized in Table 1, only four protocols showed questionable or better performance across all tests of sensitivity (Figs. 2, 4, and 5), specificity (Figs. 4, 5, and 8), and correction-induced artifact (Figs. 2 and 8): the Minimal-EMG protocol paired with Minimal- or Intermediate-NNNM filtering and the Maximal-EMG protocol paired with
Minimal- or Maximal-NNNM filtering. For detailed results, see Supplementary Tables 1–3.

Moreover, inspection of Table 1 indicates that among these four, the Minimal-EMG/Intermediate-NNNM protocol invariably equaled or exceeded the performance of the Minimal-EMG/Minimal-NNNM protocol; likewise, the Maximal-EMG/Maximal–NNNM protocol always outperformed the Maximal-EMG/Minimal-NNNM protocol. Accordingly, these two combinations of EMG correction and NNNM filtering were subjected to additional testing in the intracerebral source-space using LORETA.

Intracerebral source modeling

Effects prior to EMG correction

Myogenic ROI. Muscle tensing increased current density in a large number of frontopolar and ventral prefrontal voxels near the facial muscles (Fig. 9). Across correction protocols, the myogenic contrast (OR-OT) was significant at 717–789 voxels (~30% of the source-space). Myogenic ROIs (715–749 voxels) were created from this contrast after applying a cluster-extent threshold (Supplementary Fig. 14). The median t-scores ($t_s = -3.75$ to $-3.83$, $ps < 0.01$, $f_{g} = 0.47$ to 0.48) and extreme t-scores ($t_s = -6.19$ to $-6.33$, $ps < 0.001$, $f_{g} = 0.71$) were similar across the two filters.

Neurogenic ROI. The Berger maneuver was associated with widespread attenuation of current density across the posterior cortex (Fig. 9). Using the neurogenic (OR-CR) contrast, ROIs were generated for the Intermediate- and Maximal-NNNM filters (765–766 voxels) across posterior voxels (Supplementary Fig. 15). Median (t$s = -3.92$ to $-3.96$, $ps < 0.01$, $f_{g} = 0.49$) and extreme t-scores (t$s = -7.26$ to $-7.68$, $ps < 0.001$, $f_{g} = 0.77$ to 0.79) were similar across the two filters.

Covarying effects. The presence of uncorrected EMG artifact altered the magnitude of the neurogenic effects produced by the Berger maneuver (Fig. 9). In particular, when changes in neurogenic and myogenic activity negatively covaried (OT-CR) there was a substantial attenuation of effects in the myogenic ROI for both median (t$s = -3.75$, $ps < 0.01$, $f_{g} = 0.47$) and extreme t-scores (t$s = -6.19$, $ps < 0.001$, $f_{g} = 0.71$). The attenuation was reduced, but still observable, in the posterior neurogenic ROI (median $ps < 0.08$, $f_{g} = 0.18$; extreme t$s > 4.53$, $ps < 0.01$, $f_{g} = 0.56$). Conversely, when changes in neurogenic and myogenic activity positively covaried (OR-CR), effects in the myogenic ROI were amplified as indexed by both the median (t$s < -2.30$, $ps < 0.04$, $f_{g} = 0.25$) and extreme t-scores (t$s < -4.32$, $ps < 0.001$, $f_{g} = 0.54$). These deleterious effects were weaker in the neurogenic ROI, reaching significance for the extreme (t$s < -3.81$, $ps < 0.01$, $f_{g} = 0.48$) but not the median t-scores ($ps > 0.72$, $f_{g} < 0.01$).

Sensitivity

Myogenic contrast (OR-OT). Although both protocols quantitatively reduced EMG contamination in the myogenic ROI (Fig. 10, red points), their sensitivity was poor (Supplementary Table 4, Supplementary Figs. 16–17).

Negatively covarying contrast (OT-CR). The Maximal-EMG/Maximal-NNNM pairing showed adequate sensitivity (Supplementary Table 4), whereas the Minimal-EMG/Intermediate-NNNM pairing evidenced poor sensitivity. Across protocols, voxels in the myogenic ROI showing weaker EMG contamination were more resistant to correction (Fig. 10).
Positively covarying contrast (OR-CT). Again, the Maximal-EMG/Maximal-NNNM pairing demonstrated adequate sensitivity, whereas the Minimal-EMG/Intermediate-NNNM pairing showed poor sensitivity (Supplementary Table 4). Inspection of the myogenic ROI voxels that yielded peak errors indicated that the Minimal-EMG/Intermediate-NNNM pairing tended to undercorrect the data (negative t-scores), whereas the Maximal-EMG/Maximal-NNNM pairing tended to overcorrect the data (positive t-scores), albeit to a lesser degree. This is depicted in Figs. 8 and 10.

Negatively covarying contrast (OT-CR). Both protocols tended to attenuate neurogenic activity, yielding questionable or worse specificity (Supplementary Table 5 and Fig. 10).

Positively covarying contrast (OR-CT). Both pairings exhibited acceptable or better specificity (Supplementary Table 5). Inspection of the neurogenic ROI voxels that yielded peak errors indicated that the Minimal-EMG/Intermediate-NNNM pairing led to undercorrection (negative t-scores), whereas the Maximal-EMG/Maximal-NNNM pairing led to overcorrection (positive t-scores). This is depicted in Fig. 10.

Correction artifact

Myogenic contrast (OR-OT) in the neurogenic ROI. The Maximal-EMG/Maximal-NNNM pairing yielded an acceptable amount of correction-induced artifact when EMG was present (Supplementary Table 6 and Fig. 10, blue points), whereas the Minimal-EMG/Intermediate-NNNM pairing showed poor performance.

Neurogenic contrast (OR-CR) in the myogenic ROI. Both pairings showed an acceptable level of correction artifact when EMG was absent (Supplementary Table 6 and Fig. 10, red points).

Discussion

Given the substantial inferential threat posed by EMG contamination, there is a pressing need for valid correction tools. In recent years, ICA has rapidly become a popular tool for correcting EMG artifact, despite limited work assessing its validity for this purpose.
Accordingly, we quantitatively tested the sensitivity and specificity of ICA-based EMG correction using a naturalistic dataset in which neurogenic (opening and closing the eyes, i.e., the "Berger maneuver") and myogenic activity (scripted muscle tensing and quiescence) were independently manipulated. This design allowed investigation of ICA-based EMG correction under conditions in which changes in neurogenic and myogenic activity covaried, as they do in many experimental settings.

Low-intensity clenching of the face increased EEG spectral power across much of the scalp, extending as far as the fronto-central electrodes. Smaller pockets of myogenic activity were also present at the posterior edge of the electrode array (Fig. 1). In contrast, the Berger maneuver altered power across the entire array, peaking at midline parietal sites. And while the effect-size of the anterior myogenic effect was only one-third that associated with the posterior neurogenic effect, it was sufficient to alter the magnitude of neurogenic effects. For instance, when changes in EMG and EEG negatively covaried, as they would be expected to do in many experiments, neurogenic effects associated with the Berger maneuver were attenuated at electrodes across the array, including posterior sites far removed from the area of peak myogenic activity (Figs. 1 and 4).

Consistent with these findings, uncorrected EMG artifact increased alpha-band current density across large regions (~30%) of the intracerebral source-space located near the facial muscles (i.e., frontopolar and ventral prefrontal cortex, insula, temporal poles; Fig. 9). This effect reflects the fact that LORETA-Key and many other distributed modeling packages restrict the source space to the cerebral cortex. Consequently, activity generated in the cranial muscles tends to be explained by dipoles fitted to the proximal region of cortex. The Berger maneuver was associated with attenuated current density across the posterior cortex (Fig. 9). Furthermore, covariation in neurogenic and myogenic activity significantly altered the magnitude of neurogenic effects in anterior regions of the brain (Fig. 9). Similar, albeit less dramatic, distortions were found in posterior regions, particularly for negatively covarying activity.

<table>
<thead>
<tr>
<th>EMG correction</th>
<th>NNNM filtering</th>
<th>Myogenic</th>
<th>Neurogenic</th>
<th>Negatively-covarying</th>
<th>Positively-covarying</th>
<th>Lowest rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal</td>
<td>Minimal</td>
<td>7/+</td>
<td>+/+</td>
<td>7/+</td>
<td>++/++</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
<td>7/+</td>
<td>+/++</td>
<td>+/+</td>
<td>++/++</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximal</td>
<td>7/+</td>
<td>++/++</td>
<td>+/++</td>
<td>++/++</td>
<td>7</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Minimal</td>
<td>+/+</td>
<td>+/+/+</td>
<td>+/+/+</td>
<td>++/++</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
<td>+/+</td>
<td>++/++</td>
<td>++/++</td>
<td>++/++</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximal</td>
<td>+/+</td>
<td>++/++</td>
<td>++/++</td>
<td>++/++</td>
<td></td>
</tr>
<tr>
<td>Maximal</td>
<td>Minimal</td>
<td>+/+</td>
<td>++/++</td>
<td>++/++</td>
<td>++/++</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
<td>+/+/+</td>
<td>++/++</td>
<td>++/++</td>
<td>++/++</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Maximal</td>
<td>+/+/+</td>
<td>++/++</td>
<td>++/++</td>
<td>++/++</td>
<td></td>
</tr>
</tbody>
</table>

a Results from the myogenic contrast (corrected OR-OT) in the myogenic and neurogenic ROI.
b The corrected OR-CR vs. Uncorrected OR-CR contrast.
c Corrected OT-CR vs. uncorrected OR-CR.
d Corrected OR-CT vs. uncorrected OR-CR.
e Ratings from each ROI: − poor, ? questionable, + adequate, ++ excellent (see Methods).
From these data, independent components were extracted and classified for each participant. Neurogenic components typically accounted for the most variance in scalp electrical activity (median: 25%), whereas those classified as entirely or predominantly myogenic each accounted for another 8–10%. Participants showed pronounced variability in the amount of variance accounted for by components exhibiting characteristics of both myogenic and neurogenic activity (interquartile range for “myogenic-dominant” components: 3–33%; see Fig. 6). Descriptively, the choice of which components to exclude had a marked impact on the amount of variance retained for analysis (Fig. 7). Nine different protocols for determining which components to discard were examined, reflecting the factorial pairing of three for removing non-neurogenic/non-myogenic (NNNM) components with three for removing EMG. NNNM filtering removed between one-quarter and one-third of the total variance and EMG correction removed between a tenth and one-third of the total. Together, these two filters led to the exclusion of as little as one-third to as much as three-quarters of the variance in scalp electrical activity.

Not surprisingly, the validity of these protocols was also quite variable. In the present study, sensitivity (i.e., attenuation of myogenic effects), specificity (i.e., preservation of neurogenic effects), and correction artifacts (i.e., generation of effects in the absence of artifact) were each quantitatively assessed using ROIs corresponding to areas of peak myogenic and neurogenic activation. Results indicated that most of the nine protocols did a reasonable job removing EMG artifact, evidenced by adequate or excellent sensitivity (Table 1). Unfortunately, many of these pairings also altered neurogenic activity, evidenced by inadequate specificity or excessive correction-induced artifact (Table 1). In fact, only two pairings—Maximal-EMG correction combined with Minimal- or Maximal-NNNM filtering—showed adequate or excellent performance across all three measures. Furthermore, the Maximal-EMG/Maximal-NNNM pairing tended to outperform the Maximal-EMG/Minimal-NNNM pairing. None of the nine procedures consistently displayed excellent performance.

On the scalp, the most sensitive and specific procedure for removing EMG artifact from the alpha band was among the strictest, indexed by the amount of variance discarded (median: 71%). That is, the Maximal-EMG/Maximal-NNNM pairing entailed rejecting any component containing myogenic signal—including those where myogenic activity was less prominent than neurogenic activity—in addition to those indexing gross or ocular artifacts, noise, or unclassifiable low-variance signals. The only other procedure that consistently showed adequate performance—the Maximal-EMG/ Minimal-NNNM pairing—was similarly strict (median variance discarded: 63%), differing only in the retention of noise and unclassifiable components.

In contrast to the scalp analyses, Maximal-EMG correction paired with Maximal-NNNM filtering was associated with inadequate performance in the intracerebral source-space (Fig. 10 and Supplementary Tables 4–6). In particular, it failed to adequately remove EMG when neural activity was fixed and overcorrected neurogenic activity when EMG was absent. Poor sensitivity was also evident for the other procedures examined in the source-space (i.e., Minimal-EMG/Intermediate-NNNM).

Prospects for ICA-based EMG correction

As noted in the Introduction, many studies have documented, with varying degrees of rigor, the utility of infomax ICA for attenuating various physical and biological artifacts. Nevertheless, several studies have found ICA to exhibit worse performance than alternative source separation algorithms for ocular (Wallstrom et al., 2004; Romero et al., 2008) and EMG artifacts (Crespo-Garcia et al., 2008; Fitzgibbon et al., 2007). More recently, Debener et al. (2007) showed that ICA displays low specificity for ballistocardiogram artifacts, evidenced by attenuation of event-related neurogenic activity, under some circumstances (Debener et al., 2008). Paralleling Debener and colleagues’ research, the present findings suggest that ICA is a valid means of correcting EMG in some, but not all, cases. On the scalp, some of the ICA-based protocols we tested displayed adequate sensitivity and specificity, whereas other did not. Furthermore, in the intracerebral source-space, even those protocols that showed the most promising performance on the scalp failed.

The inadequate performance of ICA in the source-space likely reflects two factors. First, source modeling, implemented here using the LORETA algorithm and a three-shell head model, makes use of data from all electrodes in the array, not just those in the scalp ROIs. Less-than-perfect EMG correction at even a modest number of electrodes in- or outside of the scalp ROIs could, therefore, exert a substantial influence on the EEG source model. As such, source modeling can be viewed as providing a “global” check on the quality of EMG correction performed on the scalp, complimenting the more “local” ROI analyses.
Fig. 9. Contrasts of interest in the source-space after applying the Intermediate-NNNM filter. (A) Myogenic contrast (OR-OT), (B) neurogenic contrast (OR-CR), (C) error induced by negatively-covarying artifact prior to EMG correction ([OT-CR] - [OR-CR]), and (D) error induced by positively-covarying artifact ([OR-CT] - [OR-CR]) prior to EMG correction.
Second, several lines of evidence suggest that ICA failed to adequately separate myogenic from neurogenic sources on the scalp, causing under-correction (low sensitivity) and over-correction (low specificity) in the source-space. In particular, the amount of variance accounted for by “mixed” components, those displaying characteristics of both neurogenic and myogenic activity, was equivalent (e.g., myogenic-dominant vs. pure myogenic components, Fig. 6). If ICA had cleanly separated the two sources, one would instead expect the mixed components to be infrequent and to account for little variance.

Quantitatively, none of the nine protocols consistently showed excellent performance on the scalp. That is, a reasonably small number of “worst-case” electrodes located inside of the scalp ROIs (Fig. 3) evinced under- or overcorrection (Supplementary Tables 1–3).
Qualitatively, visual inspection of the topographic maps created following EMG correction indicated significant distortions outside of the scalp ROIs. For instance, significant residual artifact was present at the edge of the array for the corrected myogenic contrast depicted in Fig. 2. Likewise, evidence of overcorrection (blue regions) and, to a lesser degree, undercorrection (red regions) was present outside of the defined ROIs when neurogenic and myogenic activity covaried (Figs. 4, 5).

It is worth emphasizing that these observations cannot be attributed to limitations in the procedure for classifying or rejecting components. The manual classification protocol was detailed (see Supplementary Method), and raters were extensively trained and highly reliable in its application ($\alpha = 0.98$). In contrast to prior studies, the impact of systematically varying criteria for rejecting both non-myogenic and myogenic components was examined. Of the nine protocols examined, only four consistently showed questionable or better performance on the scalp, and none did so in the source-space. Systematic biases in the classification or rejection of components cannot explain the combination of low sensitivity and specificity in the source-space, whereas a failure to cleanly separate myogenic and neurogenic sources does.

Two factors could plausibly account for the inability of infomax ICA to fully separate myogenic from neurogenic systems. Systematically testing these hypotheses represents a key challenge for future research. First, inadequate separation might reflect non-optimal specification of the number of components to extract (“model order”). Like most other high-resolution studies, a combination of PCA and ICA was used to extract fewer components, 64, than the limit imposed by the number of electrodes in the array, 128. As detailed in Footnote 3, the number of components was subjectively optimized using a trial-and-error approach. Model order was also constrained to be identical across participants. Notably, specifying too many components or too few (i.e., over- or under-fitting) might explain poor source separation (Naeem et al., 2009; Ryali et al., 2009). This possibility could be tested using an information theoretic approach to objectively identify the optimal model order for each participant, as is typical in the functional magnetic resonance imaging literature (Beckmann and Smith, 2004; Calhoun et al., 2001; Li et al., 2007) and has occasionally been done in the EEG literature (Moraux and Iannetti, 2009; see also Supplementary Method and Results). Alternatively, the deflationary approach implemented in the fastica package (http://www.cis.hut.fi/projects/ica/fastica) could be used (Mantini et al., 2008). Stepwise ICA fitting, in which model order and sources are estimated simultaneously, is another possible approach (Hesse and James, 2004). Procedures for optimizing model order on an individual basis would also facilitate the development of criteria for discarding problematic participants (e.g., those with an unusually large number of components). And, from a more practical perspective, application of such procedures would reduce the need to classify large numbers of closely related or uninformative components.

Second, inadequate separation could reflect EMG violating key assumptions of the infomax ICA algorithm (Bell and Sejnowski, 1995; James and Hesse, 2005). For instance, infomax assumes that the activation of each component is sparse and non-Gaussian. The lengthy blocks of neurogenic and myogenic activation in the present study—and in many studies of emotion (e.g., Coan and Allen, 2003; Davidson et al., 1990; Shackman et al., 2006)—may not adequately satisfy this assumption. Infomax also assumes that sources are mutually temporally independent. To the degree that neurogenic and myogenic activity are too closely coupled in the time-domain they would violate this assumption (but cf. Ohla et al., 2009). Future studies could test the degree to which second-order blind source separation algorithms, that do not require strong assumptions, such as Second Order Blind Identification (SOBI) or Algorithm for Multiple Unknown Signals Extraction (AMUSE), produce better separation (Joyce et al., 2004; Romero et al., 2008; Tang et al., 2005). It might also be fruitful to investigate the utility of signal-space projection methods (Nolte and Curio, 1999; Tesche et al., 1995; Uusitalo and Ilmoniemi, 1997) or ICA performed in the frequency-domain, which may improve the separation of sources with overlapping spectral profiles (Annemuller et al., 2003, 2004; Lee et al., 2008). Finally, it will be important to examine the degree to which these conclusions generalize to paradigms characterized by punctuate bursts of time-locked activation, as is typical of ERSP studies.

GLM-based EMG correction

The present results would seem to partially contradict prior reports (McMenamin et al., 2009; Shackman et al., 2009) in which the use of one form of GLM-based EMG correction, “epoch-wise regression,” was recommended for scalp analyses. Epoch-wise regression removes epoch-to-epoch variance in alpha-band activity predicted by contemporaneous EMG-band activity (e.g., 70–80 Hz) separately for each electrode and participant. To facilitate a more direct comparison of GLM- and ICA-based correction techniques, the validity of epoch-wise regression was re-examined in the present report using the identical methods used for testing ICA (see Supplementary Method and Results). Consistent with McMenamin et al (2009), epoch-wise regression exhibited adequate performance across nearly all validation tests on the scalp (Supplementary Tables 7–9), with the lone exception of poor specificity in the presence of positively covarying neurogenic activity. The performance discrepancy between these reports is likely to be a consequence of using much larger neurogenic ROIs for the present analyses (i.e., 23 vs. 7 electrodes). Larger ROIs were employed with the aim of indexing the impact of EMG artifact and EMG correction in regions characterized by less extreme signals (myogenic and neurogenic) with the idea that such signals would be more representative of real-world changes in spectral activity. By contrast, the small, highly focused neurogenic ROI used by McMenamin et al (2009) was blind to distortions outside of areas of peak neurogenic activity.

Future challenges

Four limitations of the present study represent additional avenues for future research. First, the impact of EMG artifact and EMG correction on individual differences in state or trait brain electrical activity remains unknown. In particular, the degree to which either ICA- or GLM-based correction techniques preserve hemispheric asymmetries in tonic (“resting”) frontality, a neural marker of individual differences in emotional reactivity (Shackman et al., in press; Coan et al., 2006) and affective disorders (Thibodeau et al., 2006) remains untested.

Second, our conclusions derive from analyses of the alpha band (8–13 Hz) during extended blocks of myogenic activity. While this represents a reasonable analog to studies using blocked manipulations of emotion (e.g., threat of shock, emotional films), the degree to which these conclusions generalize to event-related designs or other frequency bands is unclear. Still, it seems likely that the quality of performance for the neighboring theta band (4–8 Hz) would be similar to that observed for alpha. In contrast, we anticipate that the quality of correction would be lower for bands, such as gamma (>30 Hz), that lie closer to peak EMG activity. Indeed, exploratory analyses indicated poor performance for ICA in the 70–80 Hz range (Footnote 5). Finally, the degree to which the present conclusions generalize to stronger or weaker EMG contamination is unknown (Fitzgibbon et al., 2007).

Third, source modeling was not incorporated into the component classification protocol. Classifications were instead based on visual inspection of time-series, power spectrum, and topography (see Supplementary Method and Results). While this is a conventional...
approach, it is possible that inspection of component dipoles would have facilitated more accurate classifications, particularly in the case of “mixed” components (e.g., Myogenic dominant). More specifically, dipoles could be modeled using the dipfit2 or bestfit plug-ins for EEGLAB. Components characterized by dipoles at the edge of or outside of the brain could then be classified as artifactual (myogenic or otherwise; Onton and Makeig, 2006; Milne et al., 2009).

Fourth, like nearly all prior validation studies, the tests of specificity are founded on the assumption that myogenic activity was absent when participants were instructed to relax (Shackman et al., 2009). If, in fact, modest amounts of EMG were present during this condition, estimates of specificity would be artifactually reduced. We consider this a minor concern; the location of the neurogenic ROI (−Pz) for testing specificity on the scalp (Fig. 3) should minimize contributions from the anterior, lateral, and posterior muscle groups (Supplementary Fig. 1) when myogenic activity is weak. Nevertheless, it would be informative to validate ICA and other EMG correction techniques using data obtained during neuromuscular blockade (Whitham et al., 2007).

Recommendations and conclusions

Consideration of these observations and the extant literature yields several recommendations. First, if widespread EMG artifacts are suspected, Maximal-EMG correction protocol with Maximal-NNNM filtering should be employed. This method exhibited the best combination of sensitivity and specificity across all tests (Table 1). It also outperformed GLM-based EMG correction, which never demonstrated excellent performance in any of our validation tests (Supplementary Tables 7–5). Second, given its merely adequate sensitivity and inconsistent specificity, we no longer recommend the use of GLM-based EMG correction techniques (cf. McMenamin et al., 2009) for studies characterized by widespread EMG artifact. Third, for investigations where specificity is a smaller concern than sensitivity, any of the Intermediate- or Maximal-EMG correction ICA-based protocols represent reasonable choices. Fourth, the use of distributed modeling techniques, such as LORETA, to estimate the intracerebral sources of spectral EEG in studies with prominent myogenic activity is not recommended. It remains to be seen whether it is reasonable to do so using other approaches, such as dipoles or beamformers (Michel et al., 2004; Nazarpour et al., 2008). Fifth, the results of the present study and several others (e.g., Whitham et al., 2007; Yuval-Greenberg et al., 2009a,b; Shackman et al., 2009) indicate that findings in the upper frequency bands (i.e., beta: 14–13 Hz; gamma: >30 Hz), should be interpreted with extreme caution, particularly when they occur in the vicinity of scalp muscles. At minimum, plots depicting the scalp topography and spectral character of the results should be presented in sufficient detail to allow readers to independently assess whether the phenomenon in question is neurogenic (Shackman, in press). And while ICA cannot be viewed as a panacea for EMG contamination, careful application of ICA or related techniques for attenuating EMG artifact represents a useful means of rejecting the most dubious results.

Recent years have witnessed a resurgence of interest in using scalp-recorded and source-modeled EEG to answer fundamental questions about the neural implementation of cognitive and affective processes (Makeig et al., 2004; Pizzagalli, 2007). The continued development and careful validation of more sophisticated techniques, such as frequency-domain ICA, for separating myogenic from neurogenic signals will have substantial benefits for this important endeavor.

Acknowledgments

The first two authors (BWM and AJS) contributed equally to this research. We thank Alanna Clare, Donna Cole, Isa Dolski, Andre Mouraux, and Aaron Teche for assistance and three anonymous reviewers for thoughtful comments. This work was supported by the NIMH (5P0-MH069315 and R37/R01-MH43454 to RJD; BWM was supported by T32-HD007151).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.neuroimage.2009.10.1010.

References


Electromyogenic artifacts and electroencephalographic inferences revisited

Brenton W. McMenamin a,⁎,1, Alexander J. Shackman b,⁎,1, Lawrence L. Greischar c, Richard J. Davidson d,⁎

a Center for Cognitive Sciences and Department of Psychology, University of Minnesota—Twin Cities, USA
b Wisconsin Psychiatric Institute and Clinics, Departments of Psychology and Psychiatry, University of Wisconsin—Madison, USA
c Laboratory for Affective Neuroscience and Waisman Laboratory for Brain Imaging and Behavior, University of Wisconsin—Madison, USA
d Laboratory for Affective Neuroscience, Waisman Laboratory for Brain Imaging and Behavior, HealthEmotions Research Institute, and Center for Investigating Healthy Minds, Departments of Psychology and Psychiatry, University of Wisconsin—Madison, USA

Abstract

Recent years have witnessed a renewed interest in using oscillatory brain electrical activity to understand the neural bases of cognition and emotion. Electrical signals originating from pericranial muscles represent a profound threat to the validity of such research. Recently, McMenamin et al. (2010) examined whether independent component analysis (ICA) provides a sensitive and specific means of correcting electromyogenic (EMG) artifacts. This report sparked the accompanying commentary Olbrich et al. (2010), and here we revisit the question of how EMG can alter inferences drawn from the EEG and what can be done to minimize its pernicious effects. Accordingly, we briefly summarize salient features of the EMG problem and review recent research investigating the utility of ICA for correcting EMG and other artifacts. We then directly address the key concerns articulated by Olbrich and provide a critique of their efforts at validating ICA. We conclude by identifying key areas for future methodological work and offer some practical recommendations for intelligently addressing EMG artifact.

Electromyogenic artifacts and electroencephalographic inferences revisited

Recent years have witnessed a renewed interest in using neural oscillations to understand the substrates of mental function and dysfunction (Uhlhaas and Singer, 2010). Electrical activity generated by the pericranial musculature, electromyogenic (EMG) artifact, is one of the most profound threats to the validity of such studies. The danger is intrinsic to the cardinal features of EMG, particularly its high amplitude, broad spectral and anatomical distributions, and sensitivity to psychologically interesting processes. Consequently, even subtle EMG artifact can generate spurious effects and can mask or otherwise alter genuine ones across virtually the entire spectrum of the electroencephalogram (EEG).

A number of tools for EMG correction have been developed, however, the pace of algorithm development and dissemination has outstripped work to rigorously assess the sensitivity and specificity of such tools. Collectively, these issues motivated several recent methodological publications by our group (McMenamin et al., 2010, 2009; Shackman, 2010; Shackman et al., 2009). In particular, in McMenamin et al. (2010), we examined the validity of the extended Infomax independent component analysis (ICA) algorithm (Jung et al., 2000a,b; Onton et al., 2006).

It was this report that sparked the accompanying commentary by Olbrich et al., 2010. Our response is organized as follows. We begin by briefly summarizing the EMG problem and then outline recent research, including our own, assessing the utility of ICA for correcting artifacts. Next, we directly address the key concerns articulated by Olbrich and critique their efforts at validating ICA. We conclude by identifying key areas for future methodological work and offer some practical recommendations for dealing with EMG artifact.

Nature of the EMG problem

The difficulty encountered when addressing EMG artifact can be attributed to three key factors: a) its spatial and spectral distribution, b) its exquisite sensitivity to a variety of psychologically interesting processes, and c) its lack of stereotypy. The EMG signal has remarkable spatial extent and spectral breadth. Although the EMG power spectrum peaks at relatively high frequencies (~100 Hz), it is sufficiently broad to overlap with all EEG frequencies of interest. Concharova et al. (2003) report reliable myogenic effects as low as 2 Hz. They also demonstrated that EMG can be detected anywhere on the scalp due to volume conduction of activity generated by muscles across the head, face and neck. Comparably widespread effects were observed by McMenamin (Fig. 1).

These problems are further complicated by the fact that EMG is sensitive to a variety of experimental manipulations. Facial EMG, in
particular, is sensitive to numerous cognitive and affective processes, including cognitive load (Cohen et al., 1992; Waterink and van Boxtel, 1994), facial mimicry (Dimberg et al., 2000), vocalization (Booker and Donald, 1980), and induced emotional states (Borden et al., 1991; Coan et al., 2001; Bradley et al., 2001). Such effects are not limited to the face: activity generated by the muscles of the neck, for instance, has been shown to closely track performance motivation (Roesch and Olson, 2007). A consequence of such effects is that changes in neurogenic and myogenic activities are often confounded, allowing muscle activity to masquerade as EEG or fundamentally alter the magnitude or topography of genuine neurogenic effects. Indeed, McMenamin showed that when changes in neurogenic and myogenic activities negatively covaried, alpha-blocking associated with eye-opening was attenuated at locations across the scalp and source-space, including posterior locations far removed from peak myogenic activity. These effects did not reflect an artificially extreme degree of EMG contamination. In fact, the statistical effect-size for the alpha-blocking contrast was three times larger than the myogenic contrast. Together, these observations underscore that even low-intensity myogenic activity represents a serious risk to validity.

Lastly, EMG exhibits a poorly stereotyped response, making removal difficult. EMG arises from the activity of spatially distributed, functionally independent muscle groups, with distinct topographic and spectral signatures. For instance, frontalis activity peaks around 25 Hz, whereas temporalis has a lower peak (~20 Hz) and broad plateau around 40–80 Hz (Goncharova et al., 2003). The spectral composition of myogenic activity can vary with contraction intensity (Goncharova et al., 2003) and fatigue (Chung et al., 2002). This is compounded by the fact that the relative contributions of each muscle can vary substantially across elicitors and individuals (Tassinary et al., 2007). Consistent with this, McMenamin demonstrated that even carefully instructed myogenic activity is characterized by marked individual differences (Table 1). Across participants (n = 17),
anywhere from 3% to 67% of the 64 independent components that were extracted were classified as ‘pure’ myogenic activity. Likewise, myogenic components accounted for as little as 1.5% and as much as 78.5% of the variance in EEG activity (see Figs. 6–7 in McMenamin). Given such individual differences in the spectral and topographic profile of myogenic activity, canonical spatial or spectral filters or templates of the kind that have been fruitfully applied to the correction of ocular artifact (e.g., Viola et al., 2009) probably are not suitable for correcting typical EMG artifacts.

Validating the use of Infomax ICA for EMG correction

The Infomax ICA algorithm has rapidly become one of the most prominent techniques for removing EMG and other contaminants from the EEG. Conceptually, ICA-based correction entails three steps. First, ICA is used to perform an unsupervised decomposition of the EEG into temporally independent components (Onton et al., 2006; Onton and Makeig, 2006). Next, components are classified, and those categorized as artifact are discarded (for a detailed classification protocol, see the Supplement to McMenamin). Classification is typically performed manually, although algorithmic techniques have been developed (Shackman et al., 2009). Finally, the ‘artifact-free’ time-series is reconstructed from the remaining components. ICA’s ability to reconstruct the time-series is a key advantage over artifact correction techniques that cannot do so, such as those based on the general linear model (McMenamin et al., 2009). Although ICA shows great promise for correcting EMG and other artifacts (e.g., Jung et al., 2000a,b), attempts to assess its validity have been limited (reviewed in McMenamin and in Shackman et al., 2009). Moreover, few studies have gauged the impact of ICA-based EMG correction on source modeling (localization), an increasingly popular technique for maximizing the anatomical information yielded by the EEG (Pizzagalli, 2007).

Accordingly, McMenamin quantitatively assessed the sensitivity and specificity of ICA on the scalp and in the cerebral source-space using a dataset in which neurogenic and myogenic activities were independently manipulated. Critically, this design allowed us to examine ICA-based EMG correction under conditions in which neurogenic and myogenic activities covaried, as they do in many experimental settings. The degree to which artifact correction produced spurious effects in the absence of artifact was also assessed. Finally, given earlier work suggesting the importance of varying the procedures for discarding non-myogenic artifacts (Shackman et al., 2009), McMenamin tested whether varying these procedures affected the quality of EMG correction. Critically, our validation analyses made use of both mean difference and mean equivalence tests (Seaman and Serlin, 1998) — the latter is essential for rigorously demonstrating the sensitivity and specificity of a particular correction technique.

Results revealed that some, but not all, of the correction procedures exhibited adequate sensitivity and specificity on the scalp. The most sensitive and specific procedure was among the most strict of the nine examined, entailing the rejection of any component showing evidence of EMG, in addition to those reflecting gross or ocular artifacts, noise, or unclassifiable low-variance signals. None of the procedures consistently showed excellent performance on the scalp. That is, a modest number of ‘worst-case’ electrodes evinced under- or over-correction. This residual error on the scalp resulted in poor sensitivity and specificity in the cerebral source-space. This conclusion was not entirely unexpected, given other work highlighting the potential limitations of Infomax ICA (Crespo-Garcia et al., 2008; Fitzgibbon et al., 2007; Wallstrom et al., 2004; Romero et al., 2008; Castellanos and Makarov, 2006; Debener et al., 2008; Hyvärinen et al., 2010; Klemm et al., 2009; Lindsen and Bhattacharya, 2010; Vanderperren et al., 2010).

Table 1

<table>
<thead>
<tr>
<th>Subject</th>
<th>Unclassifiable Noise</th>
<th>Myo &gt; Neuro</th>
<th>Neurogenic</th>
<th>Myogenic</th>
<th>Low variance</th>
<th>Ocular</th>
<th>Neuro &gt; Myo</th>
<th>Gross</th>
<th>Sum (Est. model order)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>29</td>
<td>2</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>64</td>
<td>51 (51)</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>18</td>
<td>8</td>
<td>5</td>
<td>15</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>64 (37)</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>20</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>13</td>
<td>2</td>
<td>0</td>
<td>64 (41)</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>64 (35)</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>18</td>
<td>10</td>
<td>14</td>
<td>8</td>
<td>18</td>
<td>3</td>
<td>0</td>
<td>64 (29)</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>16</td>
<td>11</td>
<td>10</td>
<td>13</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>64 (53)</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>64 (44)</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>8</td>
<td>11</td>
<td>14</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>64 (37)</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>64 (34)</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>13</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>64 (34)</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>3</td>
<td>8</td>
<td>21</td>
<td>13</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>64 (31)</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>5</td>
<td>12</td>
<td>15</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>64 (53)</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
<td>18</td>
<td>6</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>64 (47)</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>7</td>
<td>14</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>64 (37)</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>10</td>
<td>0</td>
<td>43</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>64 (40)</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>12</td>
<td>8</td>
<td>10</td>
<td>8</td>
<td>22</td>
<td>2</td>
<td>1</td>
<td>64 (40)</td>
</tr>
<tr>
<td>17</td>
<td>10</td>
<td>11</td>
<td>5</td>
<td>21</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>64 (43)</td>
</tr>
<tr>
<td>Median</td>
<td>12</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>64 (40)</td>
</tr>
<tr>
<td>Min–max</td>
<td>1–22</td>
<td>1–29</td>
<td>0–14</td>
<td>2–43</td>
<td>0–19</td>
<td>2–22</td>
<td>0–4</td>
<td>0–3</td>
<td>64–23 (53)</td>
</tr>
</tbody>
</table>

Preliminary inspection of the ICA results indicated that low-variance components were dominated by noise, making them difficult to reliably classify and leading raters to devote an undue amount of time to their consideration. Accordingly, components accounting for <0.2% of scalp variance were categorized as Low Variance, excepting cases where unambiguous classification was possible. Using the protocol detailed in the Supplement to McMenamin, remaining components were classified as Neurogenic, Myogenic, a combination of the two (Neuro > Myo or Myo > Neuro), or artifact (Gross or Ocular). Components that meet the minimum variance criterion, but proved impossible to unambiguously classify were considered Noise. Classifications were made by two raters based on inspection of component time-series, power spectra, topography, and where necessary the raw time-series. Interrater reliability was excellent, $\alpha = 0.98$ (see the Supplement to McMenamin). Final classification was by consensus. Post hoc analyses used a Bayesian model order estimation procedure (Beckmann and Smith, 2002, 2004; Rajan and Rayner, 1997) to estimate the minimum number of components, Model order, necessary to adequately describe our data. Results indicated that the 64-component extraction used by McMenamin was sufficient to avoid underfitting, but moderately overfit most participants.

Two procedures for filtering classified components were examined by McMenamin, reflecting the factorial crossing of procedures for discarding myogenic vs. non-myogenic artifacts (see Table 1 and the Supplement to McMenamin). Three procedures for discarding myogenic components were assessed: Minimal-EMG: myogenic components; Intermediate-EMG: myogenic and neurogenic components; Maximal-EMG: myogenic, myogenic–neurogenic (‘Myogenic-Dominant’ heterogeneous) components; Minimal-NNNM: gross or ocular components; Intermediate-NNNM: gross, ocular or noise components; Maximal-NNNM: gross, ocular, noise or low variance components. Application of the different filtering procedures led to marked differences in the percentage of scalp variance that was discarded (range: 26%–73%; see McMenamin, Figs. 6–7).
Rebuttal of Olbrich

We next address the three key concerns articulated by Olbrich. Although they raise several conceptually and methodologically important points, none of their specific concerns fundamentally change the implications of our research.

Too few components?

Olbrich suggests that the inadequate performance of ICA-based EMG correction reported by McMenamin might reflect the procedures used for extracting independent components, rather than an intrinsic limitation of the Infomax ICA algorithm. In particular, they expressed concern that an insufficient number of components may have been extracted to adequately separate neurogenic and myogenic sources. There are two reasons for rejecting this possibility. First, as noted by McMenamin (their footnote 5), preliminary visual inspection of the 128 components extracted from the native electrode array indicated overfitting, evidenced by the fission of artifactual signals across components (e.g., ocular and cardiac artifacts loaded onto an excessive number of components; see also Viola et al., 2009). Based on these exploratory analyses, we ultimately elected to use principal components analysis (PCA) to extract 64 components for validation testing. Qualitatively, reducing the number of components attenuated the amount of ‘splitting’ or ‘leakage’ of artifacts across multiple components. Second, using information theoretic procedures, McMenamin demonstrated that extracting 64 components always exceeded the Bayesian estimate of the minimum number required to describe the data (Table 1), providing quantitative evidence against overfitting.

Olbrich also suggested that it might have been more appropriate to compute ICA separately for each condition (i.e., quadrupling the number of components available for separating sources). There are several potential problems with this suggestion, aside from the strong possibility of overfitting the data. First, everything else being equal, reducing the amount of data used to decompose the EEG will tend to degrade the quality of the source separation (Makeig and Onton, 2009). Put simply, given fewer exemplars of a particular artifact, ICA will be less able to cleanly separate it from other sources. A second problem with this suggestion is that generating separate ICA-based artifact filters for each condition has the potential to confound errors introduced by artifact correction with differences in neurogenic activity. Artificial correlations between condition and correction errors, like correlations between condition and EMG artifact, pose a direct threat to inferential validity. Fortunately, it is easy to eliminate this artificial threat simply by performing component extraction and classification on the complete dataset for each participant (i.e., all conditions considered simultaneously).

Classification biases?

Olbrich suggested that our results might reflect problems with component classification, rather than a failing of Infomax ICA. But this concern appears to be entirely theoretical. To our knowledge, a reasonably well-validated algorithm for classifying components simply does not exist. Instead, prior work to develop classification algorithms has largely relied on simple heuristics (e.g., EMG components display more activity >40 Hz than <20 Hz) or training datasets that were pre-classified by ‘expert’ raters (e.g., Mammone and Morabito, 2008). By contrast, we provided a detailed classification protocol that relied on a comprehensive inspection of each component, including its time-series, spectra, and topography (see the Supplement to McMenamin). This classification protocol proved highly reliable across two independent human raters (\( \alpha = .93 -.98 \)). Notably, we also systematically examined the influence of 9 different procedures for rejecting myogenic and non-myogenic artifacts (see Footnote 2). Systematic biases in the classification or rejection of components cannot explain the combination of low sensitivity and low specificity in the source-space. Had we been too strict, rejecting all but the most prototypical neurogenic components, we would expect to find excellent sensitivity and poor specificity. Conversely, had we been too liberal, rejecting only the most egregious artifacts, we would expect to find the reverse pattern. The bottom line is that systematic procedural biases cannot readily account for our results, whereas a failure to cleanly separate myogenic and neurogenic sources does.

Confounding myogenic and neurogenic activities?

Finally, Olbrich raises the possibility that our results may reflect the fact that one of our simplifying assumptions, namely that neurogenic activity is consistent across periods of muscle tensing and quiescence, is wrong. The key contrast for testing the sensitivity of ICA-correction technique in McMenamin was comparing the eyes-open, tense condition (OT) after EMG correction to the uncorrected eyes-open, relaxed condition (OR). Significant differences in this contrast were interpreted as the presence of residual myogenic artifact—an interpretation that hinges on the assumption there is no systematic variation in neurogenic signals during the tensing of facial muscles. Olbrich et al. 2010 cite several ways in which this assumption may fail, and we agree that basic neurophysiology dictates that there will be greater neuronal activity in regions innervating the pericranial musculature during the Tense condition. Empirically, however, the observed effects in the corrected OT–OR contrast are far more consistent with residual artifact.

In particular, McMenamin did not observe effects in the lateral precentral gyrus for the OT–OR contrast (Fig. 1), so these regions were not incorporated into the ‘myogenic’ region of interest (ROI) that was relied on for testing the sensitivity of ICA-based EMG correction. Primary motor regions were also not incorporated into the ‘neurogenic’ ROI that were used to test specificity. Thus, neither validation test was likely to have been confounded by differences in neurogenic activity associated with the tensing manipulation.3

A further argument against Olbrich’s interpretation stems from the remarkable similarity between the topography of the corrected and uncorrected myogenic contrasts (OT vs. OR; Figs. 1–2 in McMenamin). The distribution of the myogenic contrast is nearly identical before and after ICA-based correction. This implies one of two possibilities. Either ICA-based correction failed to wholly correct the artifact, leaving a substantial residual, or it unmasked a neurogenic signal that was temporally and spatially coincident with EMG. The frontal topography renders the latter suggestion implausible. Collectively, these observations indicate that neural activity associated with muscle tensing had little effect on our conclusions.

Critique of Olbrich’s simulation

Quantitatively validating any EMG correction tool requires data in which the presence and absence of EMG is definitive or can be reasonably assumed. McMenamin generated such a dataset by instructing participants to tense or relax facial muscles. An alternative approach is to build a synthetic dataset in which the experimenter mathematically ‘injects’ artifact into otherwise artifact-free data. Simple 3 Similar logic allows us to rule out a significant contribution from between condition differences in gross arousal—as one might expect if muscle tensing was associated with greater arousal or cognitive workload than quiescence. Olbrich et al. (2009) recently used source modeling to show that such differences are associated with altered alpha-band (8–12 Hz) cortical current density in posterior regions of the cingulate, occipital, and temporal cortices. None of these regions contributed to the ROI we used to evaluate the sensitivity of ICA-based artifact correction (i.e., attenuation of myogenic activity). Such regions did overlap with the ROI used for evaluating specificity (i.e., the preservation of neurogenic activity; see Fig. 9 in McMenamin), but differences in alpha-band activity associated with tensing could not have influenced the primary test of specificity, which examined the impact of ICA-based correction on the neurogenic contrast (eyes-open/muscles-relaxed vs. eyes-closed/muscles-relaxed) in the absence of EMG artifact.
simulations can provide proof-of-principle, but may not be very informative about real-world performance given the substantial variability that is a hallmark of the pericranial EMG (see above and Table 1) and some other kinds of artifact (e.g., Gwin et al., 2010). A related concern is that the assumptions underlying simple generative models (e.g., the degree of temporal and spatial correlation with neurogenic signals) probably do not characterize real EMG contamination, leading to low external validity (Grouiller et al., 2007; Hoffmann and Falkenstein, 2009). Of course, it is possible to create more complex simulations that can yield a more detailed understanding of the conditions under which a particular correction algorithm is valid (e.g., Delorme et al., 2007; Fitzgibbon et al., 2007).

But the simulation conducted by Olbrich was not of this kind. Instead, they simulated EMG contamination by superimposing EMG from a single extracranial muscle on the artifact-free EEG at three adjacent channels. The magnitude, spectral characteristics, and topography of simulated artifact was held constant across participants (n = 10). No attempt was made to examine conditions in which neurogenic and myogenic signals covared. It is not clear what protocol was used for classifying components or even whether rater(s) were blind to the simulation protocol. Analyses were restricted to testing the null, a significant limitation that could have been circumvented using the sorts of equivalence tests employed by McMenamin. This concern is magnified by the strict threshold used for examining mean differences. Given these limitations, it is unlikely that the conclusions drawn by Olbrich from this simulation will accurately predict the real-world performance of ICA-based EMG correction.

Future challenges

Several factors could plausibly account for the inability of Infomax ICA to fully separate myogenic from neurogenic sources. Testing these hypotheses represents a profitable avenue for future research. First, inadequate separation might reflect the extraction of too many components. Consistent with this possibility, Bayesian estimates of "model order", the minimum number of components necessary to describe each participant's EEG, indicated a moderate degree of overfitting that varied substantially across individuals (Table 1). Overfitting could explain poor source separation (Naem et al., 2009; Ryali et al., 2009). This possibility could be more systematically investigated using information theoretic approaches (Beckmann and Smith, 2004; Calhoun et al., 2001; Li et al., 2007; Moraux and Iannetti, 2009), deflationary approaches (http://www.cis.hut.fi/projects/ica/fastic; Mantini et al., 2008), or stepwise ICA (Hesse and James, 2004) to objectively identify the optimal model order for each participant. Doing so might also facilitate the development of criteria for discerning problematic participants (e.g., those with an unusual number of components).

Second, source modeling was not incorporated into the component classification protocol. Classification was instead based on the visual inspection of time-series, power spectrum, and topography. While this is a conventional approach, it is possible that inspection of component dipoles would have facilitated more accurate classifications, particularly in the case of 'mixed' components (e.g., Myogenic dominant). This possibility could be evaluated by modeling dipoles for each independent component. Those characterized by dipoles at the edge or outside of the brain could, in combination with other criteria, then be classified as artifactual (myogenic or otherwise; Onton and Makeig, 2006, 2009; Milne et al., 2009) and the impact on sensitivity and specificity assessed.

Third, inadequate separation might reflect EMG violating key assumptions of the Infomax algorithm (Bell and Sejnowski, 1995; James and Hesse, 2005; Makeig and Onton, 2009). The lengthy blocks of neurogenic and myogenic activation in the present study – and in many studies of emotion (e.g., Coan et al., 2001; Davidson et al., 1990; Shackman et al., 2006) – may not adequately satisfy the assumption that component activation is non-Gaussian. Furthermore, if neurogenic and myogenic activities are too closely coupled in the time-domain they would violate the assumption that sources are mutually temporally independent (but cf. Ohla et al., 2009). Future studies should test whether second-order separation algorithms, which do not make such assumptions, yield better separation (Joyce et al., 2004; Romero et al., 2008; Tang et al., 2005). It might also prove fruitful to investigate signal-space projection methods (Nolte and Curio, 1999; Tesche et al., 1995; Uusitalo and Ilmoniemi, 1997) or source separation in the frequency-domain (Anemuller et al., 2003; Hyvärinen et al., 2010; Lee et al., 2008), which may improve the separation of sources with overlapping spectra.

A fourth promising direction is to validate correction techniques using data obtained in the presence and absence of neuromuscular blockade (Whitham et al., 2008, 2007). Doing so would avoid the assumptions necessitated by simulated, scripted, and ad hoc datasets (Shackman et al., 2009).

Recommendations and conclusions

Although ICA cannot be viewed as a panacea for EMG contamination, careful application is a useful means of rejecting the most dubious results on the scalp. Nevertheless, in cases where myogenic activity is plausible, we cannot recommend the use of source modeling techniques for hypothesis testing. Likewise, the results of McMenamin and others indicate that findings in the upper frequency bands (i.e., beta, gamma) should be interpreted with extreme caution, particularly when they occur in the vicinity of scalp muscles. At minimum, scalp topography plots should be presented (Shackman, 2010; Shackman et al., 2010). We recommend that investigators adequately describe the procedures used for classifying and filtering artifactual components and, where relevant, quantitatively assess inter-rater reliability.

There is increasing interest in using scalp-recorded and source-localized EEG to answer fundamental questions about how the mind arises from and interacts with the brain (Makeig et al., 2004; Pizzagalli, 2007). The development and careful validation of novel tools for separating myogenic from neurogenic signals will have substantial benefits for this endeavor.

Acknowledgments

We thank Jeff Maxwell, David Bachhuber, and Adam Koppenhaver for their contributions to this research and three anonymous reviewers for the critical feedback. The two first authors contributed equally to this report, which was supported by the NIMH (PS0–MH069315 and R27/R01-MH43454 to RJD; BWM was supported by T32-HD007151). Authors acknowledge no conflicts of interest.

References


