Received: 31 May 2017 | Revised: 14 November 2017 | Accepted: 14 November 2017

DOI: 10.1111/psyp.13044

ORIGINAL ARTICLE

WILEY PSYCHOPHYSIOLOGY SPR SOCIETY FOR PSYCHOPHYSIOLOGICAL RESEARCH

Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials

Hannah I. Volpert-Esmond[1] | **Edgar C. Merkle**[1] | **Meredith P. Levsen**[1] | **Tiffany A. Ito**[2] | **Bruce D. Bartholow**[1]

[1] Department of Psychological Sciences, University of Missouri, Columbia, Missouri, USA

[2] Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, Colorado, USA

Correspondence

Hannah I. Volpert-Esmond, Department of Psychological Sciences, University of Missouri, 210 McAlester Hall, Columbia, MO 65211, USA.
Email: hannah.volpert@mail.missouri.edu

Funding information

National Science Foundation (grant BCS 0847872) (to B. D. B. and T. A. I.), National Institute on Alcohol Abuse and Alcoholism (grant R01 AA020970) (to B. D. B.), University of Missouri Life Sciences Graduate Fellowship (to H. I. V.-E.)

Abstract

EEG data, and specifically the ERP, provide psychologists with the power to examine quickly occurring cognitive processes at the native temporal resolution at which they occur. Despite the advantages conferred by ERPs to examine processes at different points in time, ERP researchers commonly ignore the trial-to-trial temporal dimension by collapsing across trials of similar types (i.e., the signal averaging approach) because of constraints imposed by repeated measures ANOVA. Here, we present the advantages of using multilevel modeling (MLM) to examine trial-level data to investigate change in neurocognitive processes across the course of an experiment. Two examples are presented to illustrate the usefulness of this technique. The first demonstrates decreasing differentiation in N170 amplitude to faces of different races across the course of a race categorization task. The second demonstrates attenuation of the ERN as participants commit more errors within a task designed to measure implicit racial bias. Although the examples presented here are within the realm of social psychology, the use of MLM to analyze trial-level EEG data has the potential to contribute to a number of different theoretical domains within psychology.

KEYWORDS

analysis/statistical methods, ERPs, error processing, face processing

1 | INTRODUCTION

The ERP is particularly well suited to examining psychological processes that occur quickly and often covertly following the presentation of a stimulus or a behavioral response within a particular task. Because the signal in each individual trial is quite small, the EEG is typically recorded over a large number of trials, sometimes over the course of hours. In the signal-averaging approach, EEG activity is then epoched around each stimulus or response and averaged across trials of the same type to isolate activity specific to the event (see Luck, 2014). This averaging process increases the signal-to-noise ratio by eliminating ongoing unrelated neural activity or noise from other sources, leaving activity locked to the stimulus or response in the waveform, given enough trials are included (see Meyer, Riesel, & Proudfit, 2013). The amplitude of a deflection in the averaged waveform, or ERP component, is inferred to represent the extent to which a particular psychological process is engaged at a particular time. To test differences in ERP components across participants or experimental conditions, the mean amplitude is quantified within the time interval around the deflection and analyzed using repeated measures analysis of variance (rANOVA; Jennings, 1987).

Creating averaged waveforms in this way makes an important assumption, namely, that the EEG activity directly associated with the event is constant over trials (Luck, 2014). This effectively ignores how a particular process may change from trial to trial and condenses the rich data set to one data point per condition per subject (Vossen, Van Breukelen, Hermens, Van Os, & Lousberg, 2011). This approach has the benefit of eliminating noise from the waveform before quantifying the amplitude within the time interval of interest. However, it assumes that no change in the signal of any kind is occurring over the course of the experiment. This assumption is likely not appropriate, especially across long studies where subjects may experience fatigue, learning, or familiarization with the stimuli themselves. Previously, researchers were limited to the signal-averaging approach because of the limitations of the statistical technique being used (rANOVA). However, with the advent of the application of multilevel modeling (MLM) to psychophysiological research, researchers can now use trial-level data to examine change in cognitive processes across the course of an experiment.

Multilevel models (alternatively called hierarchical linear models, mixed effects models, or mixed regression) have emerged as one of the most flexible and appropriate methods for repeated measures designs. MLM is an extension of regression that allows for simultaneous estimation of fixed effects (effects that generalize to a population) and random effects (effects specific to the sample). In their application to ERP data, multilevel models typically assume that individuals deviate randomly from the average of all individuals in their baseline response (intercept) and/or that individuals differ in the effect of a particular predictor on their response (slope).

MLM has a number of strengths that make it particularly well suited for application to ERP research. For example, it can account for a number of unique sources of variability. Commonly, one source of variability is individuals (subjects) who give repeated measurements across the course of the study. Psychophysiologists may additionally want to model variance specific to each electrode, either independently from subjects (i.e., in a cross-classified model; Baayen, Davidson, & Bates, 2008; Judd, Westfall, & Kenny, 2012) or nested within subjects (i.e., in a hierarchical model). By fitting the model to data from all electrodes of interest and including electrode as a grouping variable, variance attributed to electrodes is estimated and fixed effects are interpretable as the effect at the "average" electrode. By partitioning more sources of variance from the error term, MLM increases power to detect fixed effects (Gelman & Hill, 2007; Vossen et al., 2011), which allows researchers to examine trial-level data without first eliminating noise by averaging across trials.

Multilevel models are additionally robust to unbalanced data (i.e., missing observations). Unbalanced data can occur if one electrode is particularly noisy or if a subject has too few valid trials within a particular trial type to create a reliable average waveform using the signal-averaging approach. In this case, researchers constrained by rANOVA must discard a given subject's entire record or use mean imputation to try to reconstruct missing observations, which has other undesirable effects (Schafer & Graham, 2002). MLM eliminates the need to balance data, which is especially useful when investigating trial-level data, since the number of discarded trials due to artifacts can vary across subjects and conditions (Tibon & Levy, 2015).

Lastly, the hierarchical nature of MLM allows for both categorical and continuous predictors at any level of the model. This allows researchers to include trial or time as a continuous variable while simultaneously testing other categorical effects at the trial level (e.g., stimulus condition) and subject level (e.g., gender) simultaneously. The ability to employ trial-level data in this way provides a much more detailed picture of how neurophysiological responses—and, thus, the psychological processes that those responses reflect—change over the course of an experimental session as a function of learning, fatigue, or any number of other factors that can influence performance. In turn, this information can be critically important for investigating the effects of experimental manipulations on neurocognitive processes and how they are related to changes in response behavior. For example, does affective priming habituate or become more pronounced through repetition, and is this predicted by changes in categorization processes indexed by ERPs? How does fatigue over the course of a task affect cognitive control processes, and what implications might this have for the influence of automatic processes? How do learned expectations or changing probability information change neurocognitive responses to errors? The examination of trial-level data with MLM opens the door to investigating and statistically testing these types of hypotheses.

Despite these advantages and the adoption of MLM in other fields (e.g., Duncan, Jones, & Moon, 1998; Gueorguieva & Krystal, 2004; Lee & Bryk, 1989), use of MLM in ERP research remains rare (though its use is increasing; e.g., Alday, Schlesewsky, & Bornkessel-Schlesewsky, 2014; Bailey, Bartholow, Saults, & Lust, 2014; Hilgard, Weinberg, Hajcak Proudfit, & Bartholow, 2014; Nieuwland, 2016; Saliasi, Geerligs, Lorist, & Maurits, 2013; Tritt, Peterson, Page-Gould, & Inzlicht, 2016; Wierda, van Rijn, Taatgen, & Martens, 2010). Additionally, the vast majority of ERP studies in which MLM has been used still quantify components' amplitudes from waveforms produced by the signal-averaging approach. The current article presents two examples to illustrate the advantages of using trial-level data to investigate how ERP components change over the course of an experiment.

2 | STUDY 1

Due to the importance of accurately and quickly processing faces during social interactions, certain regions of the brain appear to be specialized for the processing of human faces (i.e., the fusiform face area; see Corrigan et al., 2009). Arising from activity in these areas, the N170 ERP component is observed over occipital-temporal areas of the scalp and responds selectively to faces within just ~170 ms.

Extensive research has linked the N170 most directly to structural encoding of a face (Rossion & Jacques, 2011). Prominent face perception models (e.g., Bruce & Young, 1986) theorize that structural encoding must precede subsequent identification and categorization processes, which is consistent with a number of studies that fail to show a distinction in N170 amplitude according to race (Caldara et al., 2003; Caldara, Rossion, Bovet, & Hauert, 2004; Chen, Pan, Wang, Xiao, & Zhao, 2013; Ofan, Rubin, & Amodio, 2011). Other studies, however, have found differences in N170 amplitude according to race (e.g., Brebner, Krigolson, Handy, Quadflieg, & Turk, 2011; He, Johnson, Dovidio, & McCarthy, 2009; Herrmann et al., 2007; Walker, Silvert, Hewstone, & Nobre, 2008; Wiese, Stahl, & Schweinberger, 2009), gender (Wolff, Kemter, Schweinberger, & Wiese, 2014), and even membership in arbitrarily created "minimal" groups (Ratner & Amodio, 2013), although the direction of these effects is mixed. On the basis of these findings, authors have suggested that, contrary to previous theorizing, differences in N170 amplitude elicited by members of different social categories may reflect motivation or perceiver goals related to distinguishing in-group and out-group faces (Ito & Senholzi, 2013; Ratner & Amodio, 2013; Senholzi & Ito, 2013), and that this may occur during structural encoding (Freeman, Ambady, & Holcomb, 2010).

The inconsistency in the existing literature challenges our understanding of when social category cues are incorporated in face processing and how top-down factors, such as motivation to make group-based distinctions, influence early face processing. Examining trial-level data using MLM can help test possible explanations. If motivation to distinguish faces by race diminishes as the participant tires over the course of a task, and if differentiation in the N170 to faces of different races depends on motivation to categorize by race, one would expect race-based differentiation in N170 amplitude to decrease over the course of the task. Although participant motivation was not measured, the current study illustrates the usefulness of trial-level data and MLM to test for change in N170 amplitude to different race faces as time-on-task increases (and as participants' motivation presumably decreases).

2.1 | Method

2.1.1 | Participants

Sixty-five adults (34 women, 31 men), ages 18–48 ($M = 20.4$), participated in exchange for monetary compensation or credit toward a research requirement. Sixty identified as White, 2 identified as Asian, and 3 identified as more than one race. None identified as African American.

2.1.2 | Measures and procedure

Data were originally collected for a project investigating the time course and influence of bottom-up visual manipulations on race perception, as reported in Volpert-Esmond, Merkle, and Bartholow (2017). EEG data were recorded as participants viewed grayscale photographs of White and Black men's faces with neutral expressions and categorized them by race. During each trial, a fixation cross was presented (duration varied randomly across trials to be 500, 700, or 900 ms), followed by a Black or White male face (270 ms), which was then masked with a visual noise pattern (530 ms). Participants were instructed to categorize the race of the face on each trial as White or Black by pressing one of two buttons as quickly as possible (response mapping was randomized across participants). If participants failed to respond within 800 ms following face onset, text reading "Too slow!" was displayed for 1,000 ms. The ITI was 600 ms. Participants completed 8 practice trials followed by 256 experimental trials (128 presentations of each race), separated into two blocks. More methodological details and additional manipulations and tasks can be found in Volpert-Esmond et al. (2017).

EEG data were collected using 20 tin electrodes in a stretch-Lycra cap and placed according to the extended International 10–20 system (i.e., the 10-10 system; American Clinical Neurophysiology Society, 2006).[1] Scalp electrodes were referenced online to the right mastoid. Signals were amplified with a Neuroscan Synamps amplifier (Compumedics, Charlotte, NC), filtered online at .10–40 Hz at a sampling rate of 1000 Hz, and rereferenced offline to an electrode placed on the tip of the nose (e.g., Caldara et al., 2003; Eimer, 2000; Senholzi & Ito, 2013). Impedances were kept below 15 KΩ. Blinks were corrected from the EEG signal offline using a regression-based procedure (Semlitsch, Anderer, Schuster, & Presslich, 1986). Trials containing voltage deflections of ± 75 microvolts (µV) were discarded, as well as trials undetected by the automatic artifact rejection procedure that contained large muscle artifacts, as determined by visual inspection. Grand averages revealed a negative deflection at temporal lateral electrodes peaking around 165 ms and maximal at TP7, consistent with previous characterizations of the N170 (Rossion & Jacques, 2011). First, a signal-averaging approach was used to examine differences in the N170 to faces of different races on an experiment-wide level: EEG activity following the presentation of each face was averaged together separately for each race to create average waveforms for each subject for each race condition. The N170 was quantified as the mean amplitude at electrodes

[1] Electrodes included Fp1, Fp2, Fz, FCz, FC3, FC4, Cz, C3, C4, CPz, CP3, CP4, TP7, TP8, Pz, P3, P4, P7, P8, and Oz, plus four electrodes placed above and below the left eye and on the outer canthi of each eye, as well as on each mastoid and the tip of the nose.

TABLE 1 Intraclass correlations associated with subject and electrode for the signal averaging approach and trial level approach in both studies

	Signal averaging approach		Trial level approach	
	ICC	Variance	ICC	Variance
Study 1				
Subject	.588	*4.90*	.048	*5.15*
Electrode	.046	*0.38*	.005	*0.48*
Residual		*3.05*		*101.89*
Study 2				
Subject	.479	*8.18*	.116	*8.29*
Electrode	.118	*2.02*	.028	*2.03*
Residual		*6.88*		*61.04*

Note. Variances (italicized) were estimated from intercept-only models (i.e., models without predictors). ICCs for subject and electrode were then calculated from the variances and represent the proportion of total variance in the dependent variable accounted for by each grouping variable (i.e., provide an index of between-subject and between-electrode variability). Large subject ICCs in the signal averaging approach are consistent with previous work reporting high between-subject variability in visually evoked ERPs (e.g., Gaspar, Rousselet, & Pernet, 2011). See supporting information for further discussion.
ICC = intraclass correlation.

P7, P8, TP7, and TP8 between 135 ms and 195 ms poststimulus onset (30 ms before and after the peak at TP7) for each waveform. Then, for trial-level analysis, the N170 was quantified in the same window for each separate trial for each subject.

2.1.3 | Model specification

The R package *lme4* (Bates, Mächler, Bolker, & Walker, 2015) was used to fit multilevel models for data analysis. All models for both studies used an unstructured covariance matrix and allowed for covariances between random slopes and intercepts. Subjects and electrodes were used as crossed random factors (i.e., a cross-classified model; Judd et al., 2012). Intraclass correlations associated with each random factor in each model can be found in Table 1. To determine which slopes and intercepts should be included as random effects, we used model-specification procedures described by Bates, Kliegl, Vasishth, and Baayen (2015). This procedure involved starting with a maximal model and then removing random slopes based on the magnitude of the correlations between random effects. Estimated random effect variances and correlations can be found in the online supporting information. Satterthwaite approximations were used to estimate degrees of freedom and to obtain two-tailed p values; in situations where the degrees of freedom were above 200, we report the results as z statistics. Data analysis and code can be found at <https://github.com/hiv8r3/MLM-ERP>. Identical model specification procedures were used for both examples.

Three models were used in Study 1: Model A demonstrates the typical signal averaging approach, whereas Model B and C use a trial-level approach. The dependent variable for Model A was mean N170 amplitude quantified from each subject's averaged waveforms (i.e., two observations per subject). As determined by model specification procedures, the intercept and slope of race were allowed to vary by subject; the intercept was also allowed to vary by electrode. Race was added as a Level 1 predictor (effect coded; Black = −1, White = 1).

For Model B and C, the dependent variable was mean N170 amplitude quantified from the raw waveform recorded during each trial (i.e., however many observations per subject as there were valid trials for that subject). In both models, race was again a Level 1 predictor (effect coded as in Model A). Trial was also added as a predictor, which was rescaled to have a range of 10 so that betas associated with trial would be large enough to be interpretable. Rescaling trial in this way has no impact on the estimated random effects or the significance of the regression weights. The same random effects structure was used as in Model A (i.e., random intercept and slope of race by subject, random intercept by electrode), as the model would not converge when a random slope for trial was included.

In Model B, the trial variable was centered at the beginning of the experiment (ranged from 0 to 10), whereas in Model C, trial was centered at the end of the experiment (ranged from −10 to 0). This approach (Aiken & West, 1991) allows the fixed effect of race to be estimated for the beginning and end of the experiment, respectively. The estimated fixed effects of any predictors involving trial (i.e., main effect of trial, interaction between race and trial) are identical for Model B and C.

2.2 | Results

Model A revealed a significant effect of race, $b = .24$, $t(64) = 2.59$, $p = .012$, such that Black faces elicited larger (more negative) N170s than White faces. Thus, from the signal averaging approach, we would conclude that there are mean level differences between N170s elicited by Black faces and White faces. However, the results from Model B and C show a more dynamic picture. As evident in Figure 1, the amplitude of the N170 was overall larger (more negative-going) to Black faces compared to White faces, but this difference appears to decrease across the course of the experiment. Indeed, Model B and C estimated a significant Race × Trial interaction, $b = -.04$, $z = -2.5$, $p = .013$, suggesting the difference in N170 amplitude elicited by White and Black faces changed significantly over the course of the experiment. Examination of the fixed effect of race estimated

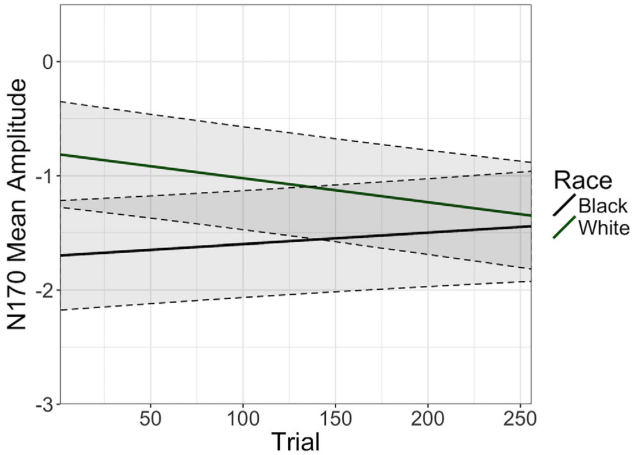

FIGURE 1 The slopes associated with change in mean amplitude of the N170 across the course of the task are plotted separately for Black and White male faces. Simple slopes and intercepts are obtained from Model B. Shaded areas reflect ± 1 standard error in model predictions

separately in Model B and Model C revealed a significant difference at the beginning of the experiment, $b = .44$, $t(181) = 3.8$, $p < .001$, but not at the end of the experiment, $b = .05$, $z = .38$, $p = .703$. In other words, the difference in N170s elicited by White and Black faces evident at the beginning of the task changed significantly over the course of the task and was no longer apparent by the end of the task.

Lastly, ERP waveforms were additionally created for the first and last 25% of trials, from which mean amplitudes (including standard error) were calculated to show compatibility between the new method advocated here and existing methods (Figure 2). Patterns of results show a very similar pattern as that found with MLM.

2.3 | Discussion

With the increased power of MLM and a large sample size, we found a main effect of race using the typical signal-averaging approach, such that Black faces elicited larger (more negative) N170s than White faces, consistent with other studies (e.g., Brebner et al., 2011; Walker et al., 2008). However, the signal-averaging approach obscures any change that occurs over the course of the experiment. Trial-level data revealed that, while the amplitude of the N170 was significantly larger to Black compared to White faces at the beginning of the task, this differentiation had effectively disappeared by the end of the task. This pattern is consistent with the possibility that differentiation in N170 amplitude to faces of different races depends on motivation to categorize faces by race, and that motivation to adhere to task demands (categorize faces by race) diminishes over the course of the task. However, participant motivation was not measured over the course of the task, and there remain other possible explanations for this pattern, including increasing familiarity of the faces, ease of processing, familiarity with the task, fatigue, or a combination of these.

Additionally, based on these results, one might expect to see a negative correlation between the number of trials in a task and the effect size of race-related differences in N170 amplitude reported in previous studies. However, this does not appear to be the case, in that there is no consistent association across studies between numbers of trials and the presence (e.g., Brebner et al., 2011; He et al., 2009; Herrmann et al., 2007; Senholzi & Ito, 2013; Walker et al., 2008; Wiese et al., 2009) or absence (e.g., Caldara et al., 2003, 2004; Chen et al., 2013; Ofan et al., 2011) of race differences in the N170, likely because existing studies differ on numerous other dimensions (e.g., task parameters, trial timing, and task demands) beyond number of trials. In the current Study 1, number of trials may serve as an index for decreasing motivation, but this index may not be consistent across different studies with different task parameters. It is possible that a meta-analysis examining a number of potential moderators (including number of trials) could identify the specific factors contributing to race (or ingroup–outgroup) effects in N170 amplitude. However, such an analysis is beyond the scope of the current report, the goals of which were to illustrate the potential utility of individual trial-level data for characterizing change in psychological processes related to face perception across an experimental session.

3 | STUDY 2

Our second example presents an extensively studied response-locked ERP component known as the error-related negativity (ERN), a negative deflection in the ERP waveform that occurs immediately following an incorrect response (Gehring, Liu, Orr, & Carp, 2011). The ERN is typically maximal in the frontocentral midline region of the scalp and originates from activity in the dorsal anterior cingulate cortex (dACC; van Veen & Carter, 2002). Initially assumed to reflect the activity of a neural error-monitoring system sensitive to internal conflict (see Botvinick & Cohen, 2014; Holroyd & Coles, 2002), the ERN (and its neural generators in the dACC) is now conceptualized as part of a broader neural salience network that is necessary for making performance adjustments when control wavers (Ham, Leff, Boissezon, Joffe, & Sharp, 2013; Hoffstaedter et al., 2014).

One area of ongoing discussion concerns the relationship between ERN amplitude and the number of errors an individual makes over the course of a task. Several studies have shown that ERN amplitude is positively correlated with the number of errors committed, such that individuals with smaller (less negative) ERNs commit more errors (Hajcak, McDonald, & Simons, 2003; Pieters et al., 2007; Riesel, Weinberg, Moran, & Hajcak, 2013; Santesso, Segalowitz, &

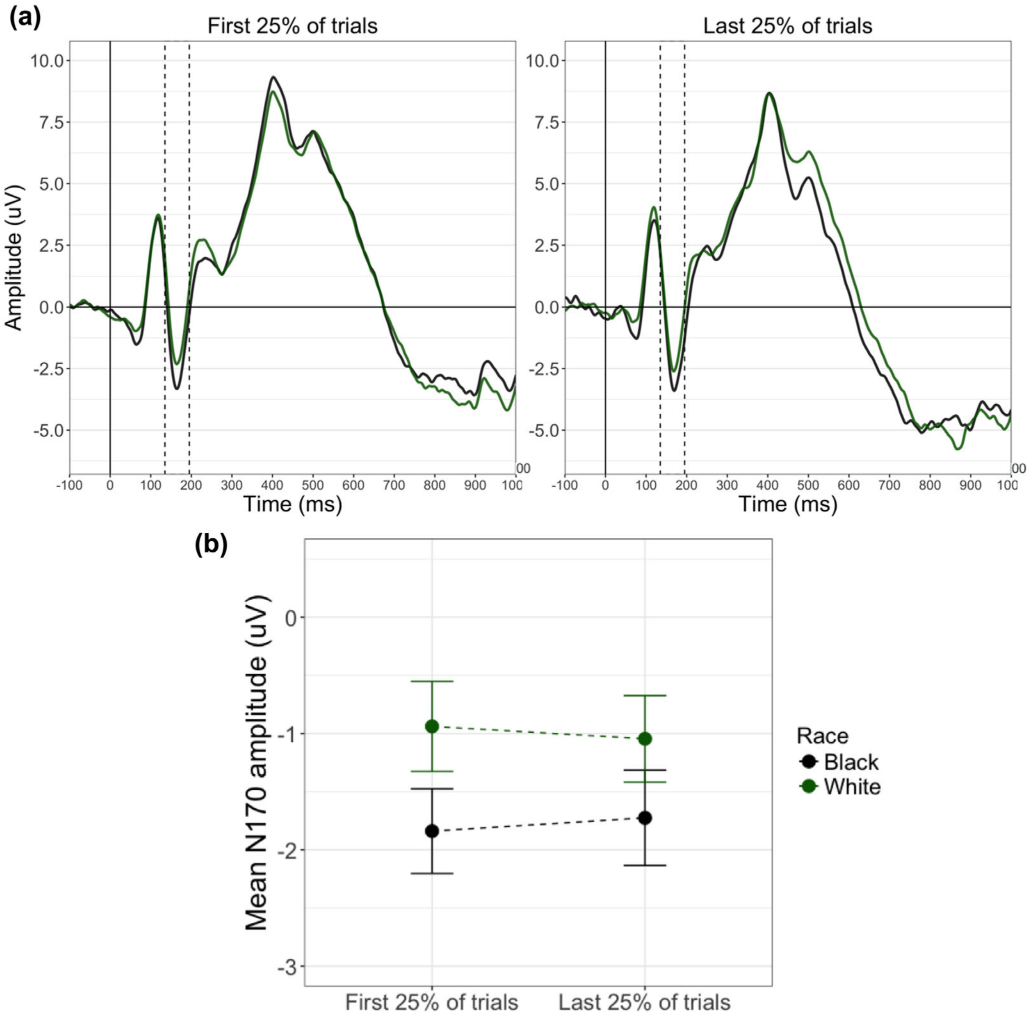

FIGURE 2 (a) Grand-averaged waveforms were formed using an average of P7, P8, TP7, and TP8 during the first and last 25% of trials. Negative amplitudes are plotted downward. Dashed lines indicate the interval in which N170 mean amplitude was quantified (135–195 ms). (b) Mean amplitude of the N170 quantified from averaged waveforms for first and last 25% of trials. Error bars indicate standard error

Schmidt, 2005). However, there are two possible interpretations for this relationship: (a) that the ERN habituates over time because errors become less salient as an individual makes more of them, resulting in smaller ERNs when many error trials are averaged together; or (b) that ERN amplitude is stable over time and indicative of an individual difference, such that individuals who have larger ERNs make fewer errors because of the successful implementation of cognitive control.

To investigate these possibilities, we used MLM and trial-level data to look at change in the ERN elicited by errors committed in the weapons identification task (WIT; Payne, 2001), an implicit racial bias task. Although the relationship between ERN amplitude and number of errors committed has primarily been examined in the context of cognitive tasks outside the realm of social psychology, this example investigates change in ERN amplitude as more errors are committed over the course of the task as a possible explanation for the positive correlation between ERN amplitude and number of errors committed reported in previous research.

3.1 | Method

3.1.1 | Participants

For this example, we examined EEG data collected from a subset of participants from a larger study ($N = 485$) first reported in Ito et al. (2015). The subset included 134 young adults (79 men, 55 women) at two universities (University of Missouri, $n = 74$, and University of Colorado, $n = 60$) who completed the WIT (identical versions across sites). Participants' ages ranged from 18 to 32 years ($M = 19.8$). One hundred-twenty identified as White, 4 identified as Asian, 3 identified as Black, and 5 identified as more than one race; 9 participants identified as Hispanic.

3.1.2 | Materials and procedure

In the WIT, participants classified images of objects as either handguns or household tools, each of which was primed by a grayscale image of an African American or European American male face. The task included a practice block of 30 trials, followed by a test block of 384 experimental trials. On each trial, participants saw a visual pattern mask (a scrambled black and white pattern, 500 ms), followed by a White or Black male face (i.e., prime, 200 ms), followed by a gun or tool (i.e., target, 200 ms), and then a second visual mask (300 ms). Participants were instructed to classify the target as either a gun or tool as quickly and accurately as possible; if they responded slower than 500 ms, a "Too slow!" message was presented to encourage faster responses. Trials were separated by a 1,000-ms intertrial interval.

EEG data were collected using 29 tin electrodes in a stretch-Lycra cap and placed according to the extended International 10–20 system (i.e., the 10-10 system; American Clinical Neurophysiology Society, 2006).[2] Scalp electrodes were referenced online to the right mastoid and rereferenced offline to an average mastoid reference. Signals were amplified with a Neuroscan Synamps2 amplifier (Compumedics, Charlotte, NC), filtered online at .10–40 Hz at a sampling rate of 1000 Hz. Impedances were kept below 10 KΩ. Blinks were corrected from the EEG signal and trials with artifacts were rejected using the same procedure reported above. For response-locked epochs, the data were further filtered at 1–15 Hz (96 db roll-off). Grand averages of the response-locked averages revealed a negative deflection at frontocentral sites peaking around 75 ms and maximal at Fz on error trials (see Figure 3), consistent with previous characterizations of the ERN (e.g., Olvet & Hajcak, 2008). Response-locked amplitudes on correct trials are also shown in Figure 3 for comparison but were not of theoretical interest and are not considered further. The ERN was quantified as the mean amplitude 25–125 ms postresponse at electrodes Fz, FCz, Cz, F3, F4, FC3, FC4, C3, and C4 for each error trial for each subject.

To examine how ERN amplitude changes as a function of how many errors have been committed, we numbered all error trials sequentially for each subject (i.e., the first error every participant committed was labeled as Error Number 1, regardless of whether it occurred at Trial 10 or Trial 100). As in Study 1, subjects and electrodes were included as crossed random factors (cross-classified model). The intercept and slopes of race and object were allowed to vary by subject; the intercept was also allowed to vary by electrode. The dependent variable was mean ERN amplitude quantified following each error committed by each subject (i.e., trial-level data), ordered sequentially. Race of the prime (Black, White), the type of object (gun, tool), and error number were included as Level 1 predictors of ERN amplitude.[3] Fixed effects of race and object are interpreted at the first error committed by each subject in this model.

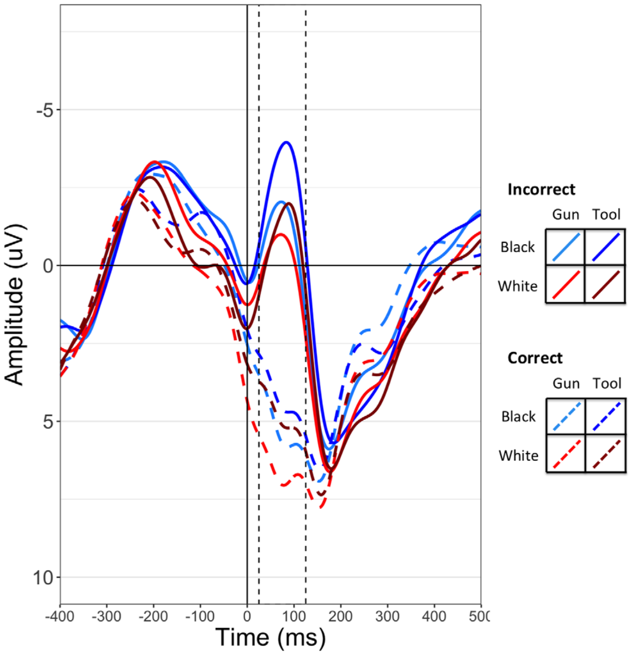

FIGURE 3 Grand-averaged waveforms depicting the ERN for the first and last 25% of trials at frontocentral electrodes (averaged over Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4) as a function of trial type, separately for correct and incorrect trials. Negative amplitudes are plotted upward, as is conventional when viewing the ERN

3.2 | Results

First, we tested the bivariate correlation between each individual's ERN amplitude and the number of errors they committed, as in previous studies (e.g., Hajcak et al., 2003). This correlation was not significant, $r = .08$, $p = .339$. However, when trials were separated according to experimental conditions, a significant correlation emerged between ERN amplitude and number of errors in Black–tool trials, $r = .24$, $p = .004$, such that subjects who more often misclassified a tool as a gun following Black faces exhibited a more dampened ERN response overall in those trials. Correlations in each of the other conditions were nonsignificant: Black–gun trials, $r = -.03$, $p = .745$; White–tool trials, $r = .10$, $p = .230$; White–gun trials, $r = -.05$, $p = .559$.

Several significant fixed effects emerged from the trial-level analysis. Interpreted at the first error committed by

[2] Electrodes included Fp1, Fp2, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, O2, and O1, as well as electrodes placed above and below the left eye, on the outer canthi of each eye, and on each mastoid.

[3] We ran a similar model with original trial number, rather than error number, as the continuous variable, such that time on task was maintained between error commissions (i.e., an error committed on Trial 10 was not equivalent to an error committed on Trial 50, even if they were both the first error committed by a participant). This resulted in a very similar pattern of results as reported with error number as a predictor (see online supporting information).

each subject, the main effect of race was significant, such that errors committed following Black primes elicited larger ERNs than errors committed following White primes (Table 2). The main effect of object was also significant, such that mistakenly classifying tools as guns elicited larger ERNs than mistakenly classifying guns as tools. These main effects were qualified by a significant Race × Object interaction, where the difference in ERNs elicited by wrongly classifying objects following Black faces was larger than the difference in ERNs to different objects following White faces.

Additionally, the effect of error number was significant, such that ERN amplitude became more positive (smaller) as more errors were committed. This main effect was qualified by an Error Number × Object interaction, such that the positive association between error number and ERN amplitude was stronger for errors following tools than errors following guns. Further examination of the simple slopes (Figure 4, Table 3) revealed a significantly positive slope in the relationship between error number and ERN amplitude for all conditions except the Black–gun condition (i.e., when participants miscategorized guns as tools following Black faces), although the three-way Error Number × Race × Object interaction was only marginally significant.

ERP waveforms created for the first and last 25% of trials (and mean ERN amplitudes) are shown in Figure 5. Results show a very similar pattern as that found with MLM. However, the advocated MLM approach is preferable, especially as very few error trials are included in each waveform, as evidenced by remaining noise in each waveform.

TABLE 2 Fixed effects of multilevel model on the mean amplitude of the ERN

	b	*p*
Race	.639 (.097)	.000*
Object	-.605 (.119)	.000*
Race × Object	.290 (.091)	.002*
Error number	.076 (.015)	.000*
Error Number × Race	.001 (.015)	.929
Error Number × Object	.033 (.015)	.023*
Error Number × Race × Object	-.028 (.015)	.056

Note. Unstandardized betas are presented; standard errors of the estimates are in parentheses. Satterthwaite approximations were used to estimate degrees of freedom to calculate p values. Race and object were effect coded (Black = −1, White = 1; gun = −1, tool = 1). Error number is rescaled to range between 0 and 10.
*$p < .05$.

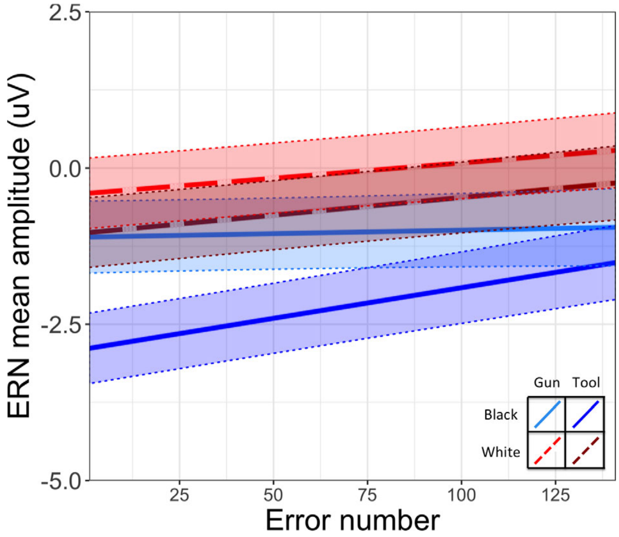

FIGURE 4 The slopes associated with change in mean amplitude of the ERN across the course of the task are plotted separately for different trial types. Simple slopes and intercepts are obtained from the model. Shaded areas reflect ± 1 standard error in model predictions

3.3 | Discussion

In this example, we tested within-subject change in ERN amplitude across the course of an implicit bias task. This approach revealed that early errors elicited larger (more negative) ERNs than errors committed later in the experiment, although this was not the case when miscategorizing guns following Black faces. This is the first demonstration of attenuation in the ERN as the number of errors committed increases over the course of an experiment, which has been theorized but not tested due to limitations inherent in traditional statistical and methodological approaches. As error number is correlated with time on task, and analyses conducted investigating the effect of time on task produced very similar results, it is unclear whether attenuation of the ERN is due to decreasing error salience as more errors are committed or decreasing motivation as a task continues (especially since ERP tasks can be quite long in duration), and what implications this might have for reactive cognitive control (e.g., Von Gunten, Volpert-Esmond, & Bartholow, 2017). However, evidence that ERN amplitude decreases (becomes

TABLE 3 Unstandardized coefficients and confidence intervals for the simple slope of error number on ERN mean amplitude as a function of trial type

	Black primes	White primes
Object		
Gun	.016 [-.051, .082]	.069* [.012, .126]
Tool	.138* [.087, .190]	.080* [.021, .139]

Note. Numbers in brackets are the 95% confidence interval around the estimate. Error number has been rescaled to range between 0 and 10.
*Estimates for which the 95% confidence interval does not cross 0.

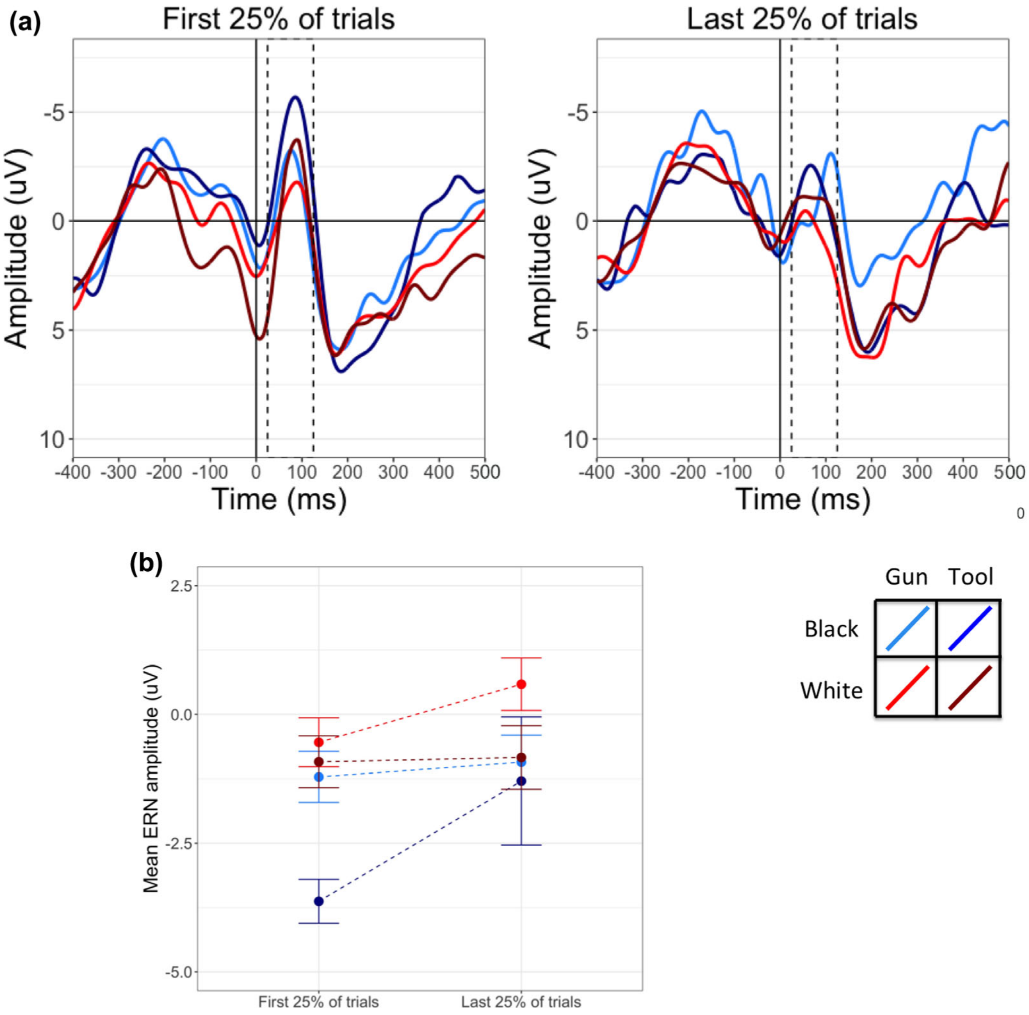

FIGURE 5 (a) Grand-averaged waveforms depicting the ERN for the first and last 25% of trials at frontocentral electrodes (averaged over Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4) as a function of trial type (error trials only). (b) Mean amplitude of ERN quantified from averaged waveforms (25–125 ms) for first and last 25% of trials. Error bars indicate standard error

more positive) as more errors are committed suggests that correlations between small ERN amplitude and worse performance in speeded computerized tasks are not the result of traitlike individual differences in conflict monitoring, but instead a dynamic relationship such that conflict monitoring diminishes as more errors are committed.

4 | GENERAL DISCUSSION

As illustrated by these two examples, the use of trial-level data in specifying multilevel models allows researchers to investigate dynamic and changing psychological processes indexed by ERPs. Although these two examples are primarily within the realm of social cognition, these principles apply to other fields in psychology that are interested in processes that may change across time. With the large number of trials that ERP paradigms typically use, ERP research is ideal for investigating change across trials within a single experiment. As with any new tool, the increase in power and utility of MLMs comes at the cost of devoting time to develop the expertise necessary to implement them appropriately (i.e., to choose appropriate random effects and covariance structures). Fortunately, several articles and tutorials have recently been published detailing the use of MLMs for ERPs and other psychophysiological data (Kristjansson, Kircher, & Webb, 2007; Page-Gould, 2017; Tremblay & Newman, 2015; Vossen et al., 2011), as well as more generally for experimental behavioral data (Baayen et al., 2008; Gueorguieva & Krystal, 2004; Quené & van den Bergh, 2004, 2008). Because MLM approaches use the rich trial-by-trial information of the full data set produced from an ERP experiment, they have yet unrealized potential for broadening our understanding of sociocognitive processes as they unfold, both over the course of a trial and over the course of an entire experiment. We believe the benefits—including the ability to handle missing observations, partition multiple sources of variance simultaneously, and examine trial-level effects—far outweigh the costs. Through these examples, we hope to demonstrate the usefulness of this technique and encourage other researchers to investigate trial-level change in their own paradigms.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regressions: Testing and interpreting interactions*. Newbury Park, CA: Sage Publications.

Alday, P. M., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2014). Towards a computational model of actor-based language comprehension. *Neuroinformatics*, *12*(1), 143–179. https://doi.org/10.1007/s12021-013-9198-x

American Clinical Neurophysiology Society. (2006). Guideline 5: Guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology*, *23*(2), 107–110.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bailey, K., Bartholow, B. D., Saults, J. S., & Lust, S. A. (2014). Give me just a little more time: Effects of alcohol on the failure and recovery of cognitive control. *Journal of Abnormal Psychology*, *123*(1), 152–167. https://doi.org/10.1037/a0035662

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. *ArXiv*:*1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*(6), 1249–1285. https://doi.org/10.1111/cogs.12126

Brebner, J. L., Krigolson, O., Handy, T. C., Quadflieg, S., & Turk, D. J. (2011). The importance of skin color and facial structure in perceiving and remembering others: An electrophysiological study. *Brain Research*, *1388*, 123–133. https://doi.org/10.1016/j.brainres.2011.02.090

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Caldara, R., Rossion, B., Bovet, P., & Hauert, C. A. (2004). Event-related potentials and time course of the 'other-race' face classification advantage. *NeuroReport*, *15*(5), 905–910.

Caldara, R., Thut, G., Servoir, P., Michel, C. M., Bovet, P., & Renault, B. (2003). Face versus non-face object perception and the 'other-race' effect: A spatio-temporal event-related potential study. *Clinical Neurophysiology*, *114*(3), 515–528. https://doi.org/10.1016/S1388-2457(02)00407-8

Chen, Y., Pan, F., Wang, H., Xiao, S., & Zhao, L. (2013). Electrophysiological correlates of processing own- and other-race faces. *Brain Topography*, *26*(4), 606–615. https://doi.org/10.1007/s10548-013-0286-x

Corrigan, N. M., Richards, T., Webb, S. J., Murias, M., Merkle, K., Kleinhans, N. M., … Dawson, G. (2009). An investigation of the relationship between fMRI and ERP source localized measurements of brain activity during face processing. *Brain Topography*, *22*(2), 83–96. https://doi.org/10.1007/s10548-009-0086-5

Duncan, C., Jones, K., & Moon, G. (1998). Context, composition and heterogeneity: Using multilevel models in health research. *Social Science & Medicine*, *46*(1), 97–117. https://doi.org/10.1016/S0277-9536(97)00148-2

Eimer, M. (2000). The face-specific N170 component reflects late stages in the structural encoding of faces. *NeuroReport*, *11*(10), 2319–2324.

Freeman, J. B., Ambady, N., & Holcomb, P. J. (2010). The face-sensitive N170 encodes social category information. *NeuroReport*, *21*(1), 24–28. https://doi.org/10.1097/WNR.0b013e3283320d54

Gaspar, C. M., Rousselet, G. A., & Pernet, C. R. (2011). Reliability of ERP and single-trial analyses. *NeuroImage*, *58*(2), 620–629. https://doi.org/10.1016/j.neuroimage.2011.06.052

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2011). The error-related negativity (ERN/Ne). In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components* (pp. 231–291). Oxford, UK: Oxford University Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gueorguieva, R., & Krystal, J. H. (2004). Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. *Archives of General Psychiatry*, *61*(3), 310–317. https://doi.org/10.1001/archpsyc.61.3.310

Hajcak, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology*, *40*(6), 895–903. https://doi.org/10.1111/1469-8986.00107

Ham, T., Leff, A., de Boissezon, X., Joffe, A., & Sharp, D. J. (2013). Cognitive control and the salience network: An investigation of error processing and effective connectivity. *Journal of Neuroscience*, *33*(16), 7091–7098. https://doi.org/10.1523/JNEUROSCI.4692-12.2013

He, Y., Johnson, M. K., Dovidio, J. F., & McCarthy, G. (2009). The relation between race-related implicit associations and scalp-recorded neural activity evoked by faces from different races. *Social Neuroscience*, *4*(5), 426–442. https://doi.org/10.1080/17470910902949184

Herrmann, M. J., Schreppel, T., Jäger, D., Koehler, S., Ehlis, A.-C., & Fallgatter, A. J. (2007). The other-race effect for face perception: An event-related potential study. *Journal of Neural Transmission*, *114*(7), 951–957. https://doi.org/10.1007/s00702-007-0624-9

Hilgard, J., Weinberg, A., Hajcak Proudfit, G., & Bartholow, B. D. (2014). The negativity bias in affective picture processing depends on top-down and bottom-up motivational significance. *Emotion*, *14*(5), 940–949. https://doi.org/10.1037/a0036791

Hoffstaedter, F., Grefkes, C., Caspers, S., Roski, C., Palomero-Gallagher, N., Laird, A. R., … Eickhoff, S. B. (2014). The role of anterior midcingulate cortex in cognitive motor control. *Human Brain Mapping*, *35*(6), 2741–2753. https://doi.org/10.1002/hbm.22363

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218. https://doi.org/10.1037/a0038557

Ito, T. A., & Senholzi, K. B. (2013). Us versus them: Understanding the process of race perception with event-related brain potentials. *Visual Cognition*, *21*(9–10), 1096–1120. https://doi.org/10.1080/13506285.2013.821430

Jennings, J. R. (1987). Editorial policy on analyses of variance with repeated measures. *Psychophysiology*, *24*(4), 474–475. https://doi.org/10.1111/j.1469-8986.1987.tb00320.x

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kristjansson, S. D., Kircher, J. C., & Webb, A. K. (2007). Multilevel models for repeated measures research designs in psychophysiology: An introduction to growth curve modeling. *Psychophysiology*, *44*(5), 728–736. https://doi.org/10.1111/j.1469-8986.2007.00544.x

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education*, *62*(3), 172–192. https://doi.org/10.2307/2112866

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: MIT Press.

Meyer, A., Riesel, A., & Proudfit, G. H. (2013). Reliability of the ERN across multiple tasks as a function of increasing errors. *Psychophysiology*, *50*(12), 1220–1225. https://doi.org/10.1111/psyp.12132

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 316–334. https://doi.org/10.1037/xlm0000173

Ofan, R. H., Rubin, N., & Amodio, D. M. (2011). Seeing race: N170 responses to race and their relation to automatic racial attitudes and controlled processing. *Journal of Cognitive Neuroscience*, *23*(10), 3153–3161. https://doi.org/10.1162/jocn_a_00014

Olvet, D. M., & Hajcak, G. (2008). The error-related negativity (ERN) and psychopathology: Toward an endophenotype. *Clinical Psychology Review*, *28*(8), 1343–1354. https://doi.org/10.1016/j.cpr.2008.07.003

Page-Gould, E. (2017). Multilevel modeling. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *The handbook of psychophysiology* (pp. 662–678). Cambridge, UK: Cambridge University Press.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192. https://doi.org/10.1037/0022-3514.81.2.181

Pieters, G. L. M., de Bruijn, E. R. A., Maas, Y., Hulstijn, W., Vandereycken, W., Peuskens, J., & Sabbe, B. G. (2007). Action monitoring and perfectionism in anorexia nervosa. *Brain and Cognition*, *63*(1), 42–50. https://doi.org/10.1016/j.bandc.2006.07.009

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1), 103–121. https://doi.org/10.1016/j.specom.2004.02.004

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002

Ratner, K. G., & Amodio, D. M. (2013). Seeing "us vs. them": Minimal group effects on the neural encoding of faces. *Journal of Experimental Social Psychology*, *49*(2), 298–301. https://doi.org/10.1016/j.jesp.2012.10.017

Riesel, A., Weinberg, A., Moran, T., & Hajcak, G. (2013). Time course of error-potentiated startle and its relationship to error-related brain activity. *Journal of Psychophysiology*, *27*(2), 51–59. https://doi.org/10.1027/0269-8803/a000093

Rossion, J., & Jacques, C. (2011). The N170: Understanding the time course of face perception in the human brain. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components* (pp. 115–142). Oxford, UK: Oxford University Press.

Saliasi, E., Geerligs, L., Lorist, M. M., & Maurits, N. M. (2013). The relationship between P3 amplitude and working memory performance differs in young and older adults. *PLOS One*, *8*(5), e63701. https://doi.org/10.1371/journal.pone.0063701

Santesso, D. L., Segalowitz, S. J., & Schmidt, L. A. (2005). ERP correlates of error monitoring in 10-year olds are related to socialization. *Biological Psychology*, *70*(2), 79–87. https://doi.org/10.1016/j.biopsycho.2004.12.004

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Semlitsch, H. V., Anderer, P., Schuster, P., & Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology*, *23*(6), 695–703. https://doi.org/10.1111/j.1469-8986.1986.tb00696.x

Senholzi, K. B., & Ito, T. A. (2013). Structural face encoding: How task affects the N170's sensitivity to race. *Social Cognitive and Affective Neuroscience*, *8*(8), 937–942. https://doi.org/10.1093/scan/nss091

Tibon, R., & Levy, D. A. (2015). Striking a balance: Analyzing unbalanced event-related potential data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00555

Tremblay, A., & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. *Psychophysiology*, *52*(1), 124–139. https://doi.org/10.1111/psyp.12299

Tritt, S. M., Peterson, J. B., Page-Gould, E., & Inzlicht, M. (2016). Ideological reactivity: Political conservatism and brain responsivity to emotional and neutral stimuli. *Emotion*, *16*(8), 1172–1185. https://doi.org/10.1037/emo0000150

van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of*

Cognitive Neuroscience, *14*(4), 593–602. https://doi.org/10.1162/08989290260045837

Volpert-Esmond, H. I., Merkle, E. C., & Bartholow, B. D. (2017). The iterative nature of person construal: Evidence from event-related potentials. *Social Cognitive and Affective Neuroscience*, *12* (7), 1097–1107. https://doi.org/10.1093/scan/nsx048

Von Gunten, C. D., Volpert-Esmond, H. I., & Bartholow, B. D. (2017). Temporal dynamics of reactive cognitive control as revealed by event-related brain potentials. *Psychophysiology*. Advance online publication. https://doi.org/10.1111/psyp.13007

Vossen, H., Van Breukelen, G., Hermens, H., Van Os, J., & Lousberg, R. (2011). More potential in statistical analyses of event-related potentials: A mixed regression approach. *International Journal of Methods in Psychiatric Research*, *20*(3), e56–e68. https://doi.org/10.1002/mpr.348

Walker, P. M., Silvert, L., Hewstone, M., & Nobre, A. C. (2008). Social contact and other-race face processing in the human brain. *Social Cognitive and Affective Neuroscience*, *3*(1), 16–25. https://doi.org/10.1093/scan/nsm035

Wierda, S. M., van Rijn, H., Taatgen, N. A., & Martens, S. (2010). Distracting the mind improves performance: An ERP study. *PLOS One*, *5*(11), e15024. https://doi.org/10.1371/journal.pone.0015024

Wiese, H., Stahl, J., & Schweinberger, S. R. (2009). Configural processing of other-race faces is delayed but not decreased. *Biological Psychology*, *81*(2), 103–109. https://doi.org/10.1016/j.biopsycho.2009.03.002

Wolff, N., Kemter, K., Schweinberger, S. R., & Wiese, H. (2014). What drives social in-group biases in face recognition memory? ERP evidence from the own-gender bias. *Social Cognitive and Affective Neuroscience*, *9*(5), 580–590. https://doi.org/10.1093/scan/nst024

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Appendix S1
Table S1
Table S2

How to cite this article: Volpert-Esmond HI, Merkle EC, Levsen MP, Ito TA, Bartholow BD. Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. *Psychophysiology*. 2018;55;e13044. https://doi.org/10.1111/psyp.13044