International Journal of Psychophysiology 162 (2021) 145–156

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Using multilevel models for the analysis of event-related potentials☆

Hannah I. Volpert-Esmond [a,*], Elizabeth Page-Gould [b], Bruce D. Bartholow [c]

[a] Department of Psychology, University of Texas at El Paso, El Paso, TX 79968, USA
[b] Department of Psychology, University of Toronto, Toronto, ON M5S, Canada
[c] Department of Psychological Sciences, University of Missouri, Columbia, MO 65211, USA

ARTICLE INFO

Keywords:
Event-related potentials
Multilevel modeling

ABSTRACT

Multilevel modeling (MLM) is becoming increasingly accessible and popular in the analysis of event-related potentials (ERPs). In this article, we review the benefits of MLM for analyzing psychophysiological data, which often contains repeated observations within participants, and introduce some of the decision-making points in the analytic process, including how to set up the data set, specify the model, conduct hypothesis tests, and visualize the model estimates. We highlight how the use of MLM can extend the types of theoretical questions that can be answered using ERPs, including investigations of how ERPs vary meaningfully across trials within a testing session. We also address reporting practices and provide tools to calculate effect sizes and simulate power curves. Ultimately, we hope this review contributes to emerging best practices for the use of MLM with psychophysiological data.

Psychophysiologists have long recognized that the multivariate and densely repeated-measures nature of their data call for special approaches to data analysis (e.g., Games, 1976; Keselman and Rogan, 1980; Vasey and Thayer, 1987; Wilson, 1967). Over the past half-century most researchers have continued to use traditional approaches for analysis of psychophysiological data, including the use of repeated measures ANOVA to test differences in mean amplitude or latency of traditionally quantified event-related potential (ERP) components (see Jennings and Allen, 2017; Luck, 2014). In the early years of the field, this practice was likely driven by the lack of available alternatives or feasible means to carry them out. However, since statistical software packages for conducting complex data analyses—and desktop-type computers with which to run them—became available in the early 1980s, new analytic approaches have been developed that require more intensive computational resources for fitting models, including sophisticated approaches that do not rely on quantifying a particular ERP component at a single moment in time (e.g., Kiebel and Friston, 2004; Litvak et al., 2011; Pernet et al., 2011a). However, it is beyond the scope of this article to describe all data analytic advancements for ERPs. Instead, we focus on one popular approach—multilevel modeling (MLM), alternatively called hierarchical linear modeling, mixed linear modeling, mixed effects modeling, and mixed effect regression—that has emerged for analyzing traditionally quantified ERP components.

MLM is appropriate for any data that is structured such that observations are recorded within naturally occurring groups. In the realm of ERPs, multiple observations are grouped within individuals. A number of previous articles have advocated for the use of MLM with psychophysiological data, including ERPs (see Bagiella et al., 2000; Boisgontier and Cheval, 2016; Goedert et al., 2013; Kristjansson et al., 2007; Krueger and Tian, 2004; Page-Gould, 2017; Tibon and Levy, 2015; Tremblay and Newman, 2015; Volpert-Esmond et al., 2018; Vossen et al., 2011). The purpose of this article is to provide a gentle orientation to psychophysiologists who are interested in learning more about how to apply MLMs to their ERP data, and to provide suggestions for best practices to increase the reproducibility of these analyses and orient researchers to available resources to make the best analytical choices.

1. Description of MLM and its advantages

Multilevel modeling is an extension of the General Linear Model (GLM) that estimates both *fixed* effects, as the GLM does, and *random* effects. Fixed effects refer to effects that are expected to generalize

☆ HVE's contribution was supported by the National Institute on Minority Health and Health Disparities (F31 MD012751). EPG's contribution was supported by Canada Research Chairs (CRC 152583), the Social Sciences and Humanities Research Council of Canada (Insight Grant 140649), and the Ontario Ministry of Research and Innovation (Early Research Award 152655). BDB's contribution was supported by the National Institute on Alcohol Abuse and Alcoholism (R01 AA025451).

* Corresponding author.
E-mail address: hivolpertes@utep.edu (H.I. Volpert-Esmond).

https://doi.org/10.1016/j.ijpsycho.2021.02.006
Received 28 August 2020; Received in revised form 1 February 2021; Accepted 3 February 2021
Available online 15 February 2021

across the population and include the estimated effects of the specified predictor or independent variables (IV) on the outcome or dependent variable (DV). Fixed effects estimated with MLM, including betas, degrees of freedom, and associated *p*-values, are interpreted in a similar way as fixed effects estimated within a GLM. Unique to the MLM relative to the GLM are the random effects, which allow researchers to specify natural grouping variables (or "random factors") in the data that result in non-independence of observations. In the case of ERP data, random factors will likely include participants and channels; however, other random factors are possible, including items or the stimuli used to elicit the ERP signal. The intercept of the random factor can be allowed to be random, meaning that a unique intercept will be estimated for each unit of that random factor (e.g., if participants are specified as a random factor, a different intercept can be estimated for each participant). Additionally, the slope associated with a particular predictor variable can be allowed to be random for each unit of the random factor (e.g., the effect of a particular predictor is estimated separately for each participant). The MLM will provide an estimate of the variance of the random intercept or slope, thereby providing an estimate of how much variability in the intercept or slope exists within a particular random factor.[1]

rANOVA is essentially a special case of MLM and, thus, a multilevel model can be specified in a way that reproduces the results of a rANOVA. However, because it is the general case, MLM is much more flexible and allows for experimental or analytic designs that rANOVA cannot accommodate. For example, whereas rANOVA handles participants as the single random factor that results in dependence of observations, MLM can include multiple random factors in a variety of structures. This is particularly useful in ERP studies because repeated measurement within other factors produces dependence of observations, namely, electrodes/channels. In rANOVA, channel is often included as a predictor, which can result in unwieldy higher-order interactions that are difficult to interpret, especially as the number of channels increases (Luck, 2014). Instead, in MLM, in addition to specifying participants as a random factor, we can include channel as a random factor and estimate fixed effects of interest at the "average" channel.

Additionally, by specifying multiple random factors, MLM can test questions about the relative amounts of variance explained by different random factors using a special case of MLM called *covariance component models*, alternatively called *cross-classified models* (Dempster et al., 1981; Goldstein, 1987; Rasbash and Goldstein, 1994). For example, consider a stimulus set of emotional faces in which each target person makes a series of expressions that vary by arousal and valence. Covariance component models could be used to determine whether variance in P300 amplitude elicited by these faces is determined more by the targets or the participants (i.e., do P300s vary more as a function of which perceiver they are recorded from or which target they are elicited by?). Using unique ERP waveforms for each perceiver and target combination, we can specify perceivers and targets as crossed random factors and compare the variance in the random intercept for each group. More variance in one random factor or another suggests that either the perceiver or the target accounts for more variance in P300 amplitude. Thus, MLM expands the types of theoretical questions that can be answered using ERPs.

A second advantage of MLM is the flexibility it allows in the assumptions made about the variance and covariance between the observations in the dataset. In the early years of the field (e.g., Games, 1976; Keselman and Rogan, 1980; Wilson, 1967), researchers were particularly concerned about the possibility that the use of rANOVA for the kinds of successive measurements commonly obtained in psychophysiological studies often violate core assumptions underlying the use of rANOVA, especially the assumption of *sphericity* (i.e., that the variance of all pairwise differences between repeated measurements is constant). As noted by numerous researchers tackling this issue (e.g., Blair and Karniski, 1993; Jennings and Wood, 1976; Keselman and Rogan, 1980; Vasey and Thayer, 1987), the assumption of sphericity is unrealistic when applied to psychophysiological data. Other solutions have been proposed within the context of rANOVA, including well-known adjustments to the degrees of freedom of a test—and therefore the observed *p*-value—based on the degree of non-sphericity it introduces (e.g., Greenhouse and Geisser, 1959; Huynh and Feldt, 1970, 1980) and multivariate tests such as Hoteling's T^2 test (Mardia, 1975). MLM handles this issue by allowing models to be specified in a way that does not assume sphericity, thus making an adjustment unnecessary. Specifically, because MLMs are estimated with maximum likelihood methods, the assumed covariance structure of the data can be specified in a number of ways, including as an autoregressive covariance matrix, a compound symmetry covariance matrix (satisfies conditions of sphericity but more restrictive), or an unstructured covariance matrix, which makes no assumptions of equivalence among elements of the variance-covariance matrix (for a more in-depth discussion of variance-covariance structures, see Arnau et al., 2010; Page-Gould, 2017; Singer and Willett, 2003). In the case of violations of sphericity, MLMs with unstructured covariance matrices (and thus no assumption of sphericity) outperform rANOVA in containing the Type 1 error rate (Haverkamp and Beauducel, 2017), and thus may be particularly appropriate for analyzing ERP data. Additionally, another positive benefit of using maximum likelihood estimation is robustness to missing data (e.g., Enders and Tofighi, 2007; Graham, 2009; Krueger and Tian, 2004).

Lastly, in contrast with rANOVA, MLM allows for both continuous and categorical IVs. Continuous IVs can include observation-level variables, such as the hue of a particular stimulus if stimuli vary along a continuum of color, or a person-level variable, such as self-reported depression symptoms. Depending on how the random and fixed effects are specified, researchers can investigate questions such as how continuous individual differences influence the effect of a particular experimental manipulation within a single model (i.e., a cross-level interaction), rather than using difference or residual scores to produce a single ERP observation per participant and examine how it correlates with the individual difference variable of interest. Lastly, this feature of MLMs also allows researchers to investigate single-trial ERPs with time or trial included as a continuous variable in the model to look at change in ERPs over time, which we will address in later sections.

To introduce readers to the application of MLMs to ERP data, we will first use an example dataset with the error-related negativity and correct-response negativity (ERN/CRN) quantified from signal averaged waveforms to illustrate the steps of the analytic process. Then, we will discuss several extensions that are possible with MLM that rANOVA cannot accommodate.

2. Using MLMs with signal averaged ERP waveforms: an example

In the example data set, seventy-three college student participants (all African American; 22 male, 49 female, 2 trans/non-binary) completed a flanker task while EEG was recorded using 33 tin electrodes.[2] All scalp electrodes were referenced online to the right mastoid; an average mastoid reference was derived offline. Signals were

[1] It is worth noting that other approaches exist to model clustered data, and that some approaches do not involve specifying random factors (e.g., generalized estimating equations; McNeish et al., 2017). These approaches may be more useful when researchers are not interested in the random effects, as a GEE will provide similar inferences as an MLM.

[2] EEG was recorded at FP1, FP2, Fz, F1, F2, F3, F4, FCz, FC3, FC4, Cz, C1, C2, C3, C4, CPz, CP3, CP4, Pz, P1, P2, P3, P4, POz, PO5, PO6, PO7, PO8, Oz, TP7, TP8, T5/P7, and T6/P8. Additional electrodes were placed above and below the left eye and on the outer canthus of each eye (to record blinks and saccades) and over each mastoid.

amplified with a Neuroscan Synamps amplifier (Compumedics, Inc.), filtered on-line with a bandpass of 0.05–40 Hz at a sampling rate of 500 Hz. Electrode impedances were kept below 10 KΩ. Ocular artifacts (i.e., blinks) were corrected from the EEG signal using a regression-based procedure (Semlitsch et al., 1986). On each trial of the flanker task, participants saw a horizontal string of five arrowheads facing to the left or right, in which the central arrowhead matched (congruent condition) or did not match (incongruent condition) the direction of the four flanker arrowheads. Participants completed 200 trials total and were allowed to rest every 50 trials. On each trial, participants first saw a fixation cross (jittered: 1400 ms, 1500 ms, 1600 ms), followed by the string of arrows (100 ms). Participants had 800 ms from the onset of the stimulus to identify the direction of the target (central) arrowhead with their right or left index fingers using a game controller. If they did not respond within the 800 ms response deadline, a 'TOO SLOW' message was presented on the screen before the next trial.

Following baseline correction (baseline window: −300 to −100 ms prior to the response), error trials and correct trials were averaged separately to create two averaged waveforms per participant (see Fig. 1). Trials where no response was made and trials containing deflections ±75 μV were not included. Only participants with more than 6 artifact-free error trials were included in the analysis (Olvet & Hajcak, 2009), resulting in a sample of 60 participants (19 male, 39 female, 2 trans/non-binary) for the analysis. We quantified the ERN/CRN as the mean amplitude from 0 to 100 ms following incorrect/correct button presses, respectively, at channels Fz, F1, F2, F3, F4, FCz, FC3, FC4, Cz, C1, C2, C3, and C4. In addition to the flanker task, participants completed several self-report measures, including symptoms of anxiety and depression using the GAD-7 (Spitzer et al., 2006) and PHQ-9 (Kroenke & Spitzer, 2002), respectively.

As with any other analytic strategy, the first thing to do is determine the theoretical hypothesis to test, including the predictor and outcome variables in the model. A large body of literature describes the ERN as a negative-going deflection that is larger following incorrect relative to correct responses (Holroyd & Coles, 2002; Olvet & Hajcak, 2009; Yeung et al., 2004). To confirm this pattern in the example dataset, we first need to set up our data in long format, which means each observation is in a unique row with columns identifying each variable associated with each observation (e.g., participant number, response type, channel). This is in contrast to wide format, which is typically used in rANOVA, where each participant is in a unique row and columns represent both different variables and repeated observations (see Fig. 2 for an illustration). We can also include individual differences variables as unique columns (e.g., anxiety and depression scores). Since each participant has only one value for each individual difference variable, that value is repeated for every row associated with a particular participant when the data is in long format.

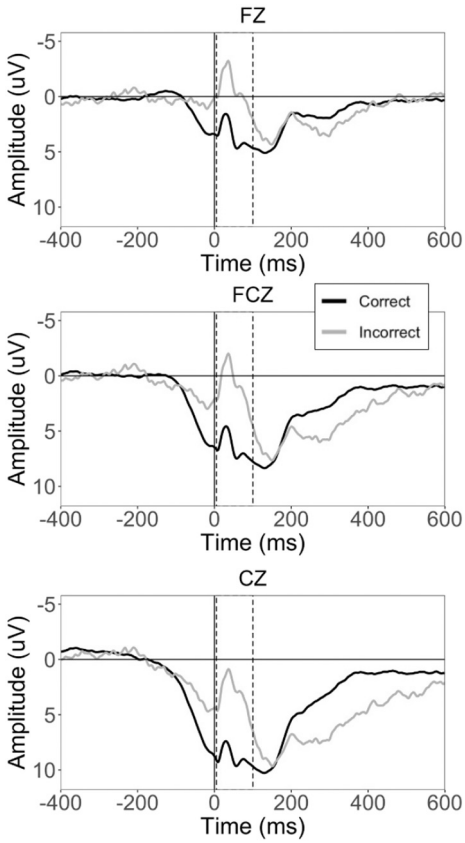

Fig. 1. Averaged waveforms from example dataset (ERN-CRN).

2.1. Setting up the model

To fit the model, we will use the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) packages in R. All R code is downloadable at [https://github.com/hivolpertes/MLMbestpractices]. First, we need to determine the hypothesis we want to test, and thus the outcome variable and the fixed effects or predictor variables to include in the model. In this example, we want to test differences in the mean amplitude of the ERN/CRN following incorrect/correct responses. Thus, mean amplitude of the ERN/CRN is the outcome variable and response (e.g., incorrect, correct) is the predictor (or fixed effect).

Then, we must specify which random factors and structure to use, which reflects the hierarchical nature of the data. This includes which grouping variables (alternatively called random factors) to include and which slopes and intercepts you allow to vary for each random factor. ERP studies using averaged waveforms often have multiple observations for each channel and for each participant and thus, the most common random factors are participants and channels. Participants and channels can either be specified as independent factors (i.e., cross-classified model) or channels can be nested within participants (i.e., hierarchical model). A hierarchical model assumes that lower-level units (in this case, channels) belong to one and only one higher-level unit (in this case, participants). This might be the case if you expect the placement of the cap on each participant to vary, such that Fz measured for one participant is substantially different from Fz measured for another participant. In contrast, a cross-classified model assumes that lower-level units do not belong to one and only one higher level unit.

Additional random factors can be selected depending on the data set and theoretical hypothesis being tested, such as stimulus items. Importantly, any random factor should contain *enough* units or clusters that observations are clustered within, although the threshold of what is *enough* is debated and depends on what estimated parameter you are most interested in (Gelman and Hill, 2007; Huang, 2018; McNeish and Stapleton, 2016a, 2016b; Snijders and Bosker, 1999). In general, the fewer the units or number of clusters within a random factor, the poorer the estimation of the variance associated with the random factor (Maas and Hox, 2005). A common rule of thumb is the 30/30 rule (30 units or clusters with 30 observations within each cluster). However, when examining fixed effects, others recommend a minimum threshold of 10 (Snijders and Bosker, 1993), some suggest a minimal threshold of 5 when examining fixed effects but 10–100 when examining random effects (McNeish and Stapleton, 2016a), and others suggest that having fewer than 5 units within a random factor does no harm but also does not differentiate the multilevel model from a classical regression model when examining the fixed effects (e.g., Gelman and Hill, 2007). Since most ERP studies using MLM are primarily interested in the fixed effects, we recommend including measurements from at least 5 units or clusters within a random factor (e.g., at least 5 channels in order to use channel as a random factor). In the current example, we have repeated measurements within channels (13 channels total) and participants (60

Wide format

ParticipantID	Fz_Incorrect	Fz_Correct	FCz_Incorrect	FCz_Correct	Cz_Incorrect	Cz_Correct	Anx_score	Dep_score
1	-1.057	5.084	1.244	8.938	6.585	12.737	5	2
2	-2.210	3.944	-3.568	8.545	-1.522	11.534	12	9
3	-1.896	0.481	1.816	4.425	4.036	6.832	9	3

Long format

ParticipantID	Electrode	Response	MeanAmp	Anx_score	Dep_score
1	Fz	Incorrect	-1.057	5	2
1	FCz	Incorrect	1.244	5	2
1	Cz	Incorrect	6.585	5	2
1	Fz	Correct	5.084	5	2
1	FCz	Correct	8.938	5	2
1	Cz	Correct	12.737	5	2
2	Fz	Incorrect	-2.210	12	9
2	FCz	Incorrect	-3.568	12	9
2	Cz	Incorrect	-1.522	12	9
2	Fz	Correct	3.944	12	9
2	FCz	Correct	8.545	12	9
2	Cz	Correct	11.534	12	9
3	Fz	Incorrect	-1.896	9	3
3	FCz	Incorrect	1.816	9	3
3	Cz	Incorrect	4.036	9	3
3	Fz	Correct	0.481	9	3
3	FCz	Correct	4.425	9	3
3	Cz	Correct	6.832	9	3

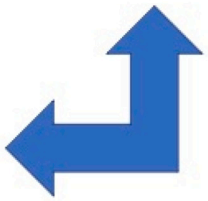

Fig. 2. Illustration of wide and long data formats.

participants total), so both are used as random factors. Since the same channels are being used for all participants, and we expect channels measured for one participant to be the same as for another participant, we will use a cross-classified model for this example.

Now that we have our random factors, we can think about which variables correspond to each level of the model. Level 1 variables correspond to individual observations, such as response type, other experimental manipulations, or aspects of the stimuli or trials that are included as predictors. Level 2 variables correspond to one level above that. In a cross-classified model where participants and channels are crossed random factors (and on the same "level," so to speak), variables corresponding to either participants or channels are Level 2 variables. In a hierarchical model where channels are nested within participants, variables corresponding to channels are Level 2 variables and variables corresponding to participants are Level 3 variables.

Once you have chosen your random factors and decided to use a hierarchical or cross-classified model, you must decide which slopes and intercepts will vary by random factor. In general, allowing the effect of a variable to vary by a random factor (i.e., including it as a random slope) will not affect the estimate of the fixed effect for that variable, because the fixed effect is essentially the average of the random slopes. However, including a random slope will generally expand the standard error of the fixed estimate, thus increasing the associated *p*-value (Barr et al., 2013; Gelman and Hill, 2007). In other words, including a random slope (especially when there is a lot of group-related variance) controls the Type 1 error rate of the test of the fixed effect more tightly and provides a more conservative (and, some argue, more appropriate) test (Heisig and Schaeffer, 2019). When choosing which effects to include as random slopes, you can use either a theory-driven approach or an empirical or data-driven approach.

2.1.1. *Theory-driven approach*

A linear model is a formal representation of a hypothesis, which extends to how you believe units within a random factor differ from one another. If you think that people differ in terms of the outcome variable (e.g., average amplitude of a given ERP component), then you will want to estimate random intercepts for each participant. If you think that the effect of a particular predictor on the outcome will differ across people in either size or direction, then you will want to estimate a random slope for that particular predictor by participant. Similar justifications can be made for including random intercepts and slopes for other random factors, such as channels. However, given that adjacent electrodes are theorized to measure similar brain activity, the effect of a predictor is not often expected to differ across channels and random slopes are not often used in this case. Thus, one way to make decisions about random effects specification is based on past empirical data or theory.

2.1.2. *Empirical approach*

Of course, your theory may be wrong (or limited). Within the last decade, researchers in psycholinguistics began calling for researchers to follow a data-driven procedure where the maximal random effects are specified for every model ("maximal model;" Barr et al., 2013). In a maximal model, all Level-1 predictors[3] are specified as random slopes. However, others have noted that using maximal models can result in significant loss of power (Matuschek et al., 2017). Additionally, as noted by Barr et al. (2013), the maximal model is frequently too complex to properly converge. When the maximal model is too complex to converge, parameter estimates are incorrect and models must be simplified. Thus, the maximal model may not always be appropriate and parsimonious models may be preferable. To determine the most appropriate parsimonious models, a number of strategies are used, including comparing nested models using likelihood ratio tests (for a comparison of strategies for model selection, see Seedorff et al., 2019).

[3] This applies only to Level-1 predictors, as Level-2 predictors cannot be included as a random slopes within a Level-2 random factor, because they are invariant within Level-2 units. In other words, if participants are being used as a random factor, Level-2 predictors (like depression or anxiety scores) will only have one observation for each person, so the effect of these variables on the outcome cannot be estimated separately for each person. In order to estimate a different random slope for each unit in a random factor, you need at least two observations per unit.

Regardless of whether you use a theory-driven or empirical approach to specify random effects, we believe that best practices involve *at minimum* including random slopes for the main fixed effects of interest to properly control for Type 1 error, as intercept-only models are frequently too liberal and may result in spurious findings (Bell et al., 2019). Once you have accounted for the main fixed effects of interest, you can make decisions whether to include more complex interactions as random slopes using a data-driven approach.

Last, after having determined the fixed effects and random effects, you should choose the type of variance-covariance matrix to use, which specifies assumptions about how observations within and across units in a random factor (e.g., within and across participants) vary and covary. Some variance-covariance matrices involve more stringent assumptions, such as a compound symmetry variance-covariance matrix, which more closely approximates a rANOVA. We suggest using an unstructured variance-covariance matrix, which removes the sphericity assumption of rANOVA, as the assumption of sphericity is unrealistic when applied to psychophysiological data (e.g., Blair and Karniski, 1993; Jennings and Wood, 1976; Keselman and Rogan, 1980; Vasey and Thayer, 1987). By default, the lme4 package in R uses an unstructured variance-covariance matrix, although SAS by default uses a VC variance-covariance matrix (for more information on variance-covariance matrixes, see Haverkamp and Beauducel, 2017; Page-Gould, 2017).

As mentioned before, in our example, we are testing the effect of Response Type (RespType) on the mean amplitude of the ERN/CRN. The model includes two crossed random factors (Participant, Channel), which we are estimating using an unstructured variance-covariance matrix. Using a theory-driven approach to determine the random effects, we included 1) a random intercept by participant, 2) response type as a random slope by participant, and 3) a random intercept by channel. The full model is described using Wilkinson notation as:

$$\text{MeanAmp}\sim\text{RespType} + (\text{RespType}|\text{Participant}) + (1|\text{Channel})$$

Our interpretation of the fixed effects depends on how Response Type is coded, similarly to interpreting fixed effects from a single-level regression model. When Response Type is dummy-coded (correct = 0, incorrect = 1), the estimate of the intercept is $b = 5.926$, 95% CIs [4.51, 7.34] and the estimate of the effect of Response Type is $b = -4.294$, 95% CIs [−5.20, −3.39]. From these estimates, we can calculate the estimated marginal means for each group: the estimated marginal mean in the correct condition is 5.926 μV (the estimate of the intercept, since correct is coded as 0) and the estimated marginal mean in the incorrect condition is 1.632 μV (the estimate of the intercept minus the estimate of Response Type). When Response Type is effect-coded (correct = −1, incorrect = 1), the estimate of the intercept is $b = 3.779$, 95% CIs [2.42, 5.13] and the estimate of the effect of Response Type is $b = -2.147$, 95% CIs [−2.60, −1.69], which means that across all trials, the estimated marginal mean is 3.779 μV. Then, we can calculate the estimated marginal means for each condition by adding or subtracting the estimate of Response Type to the intercept, which gives us the equivalent marginal means for correct and incorrect trials as the dummy-coded model. Using unstandardized estimates in this way gives us a sense of the magnitude of the difference between conditions in a meaningful unit (μV). When examining latency as the outcome variable, estimates similarly can be interpreted in whichever meaningful unit the outcome variable was measured on (such as milliseconds).

Researchers using the null hypothesis significance testing (NHST) approach will additionally want to know if the effect of Response Type is statistically different from zero. Compared to single-level regression, determination of degrees of freedom (and thus, the *p*-value associated with a test of a fixed effect) is much more complicated in MLM. A number of possibilities exist for testing the significance of a fixed effect, including likelihood ratio tests of nested models, applying the z distribution to the Wald *t* values, Markov-chain Monte Carlo (MCMC) sampling, parametric bootstrapping, and different approximations for denominator degrees of freedom. We recommend the Satterthwaite approximation for denominator degrees of freedom, partly because it more appropriately controls Type 1 error and is less dependent on sample size than other methods, especially for REML-fitted models (Luke, 2017) and because of the ease of implementation—Satterthwaite approximation is the default for SAS and can be applied in R using the lmerTest package (Kuznetsova et al., 2017) in conjunction with the lme4 package (Bates et al., 2015). All examples presented in this paper use the Satterthwaite approximation when reporting *p* values.

Critics of NHST suggest that whether the effect of a particular predictor is different from zero is not always informative—instead, it may be more useful to understand the proportion of variance explained by the fixed effects (and therefore make judgements of the meaningfulness of the effect). In a single level regression or GLM, readers are familiar with R^2 as the variance explained by all of the fixed effects included in the model. However, in multilevel models, the variance explained is a little more complex, since there are now multiple residual terms. Thus, several methods of calculating a pseudo-R^2 have been proposed (e.g., Edwards et al., 2008; Johnson, 2014; Nakagawa et al., 2017; Nakagawa and Schielzeth, 2013; Snijders and Bosker, 1999). Importantly, there is a distinction between the marginal R^2, which is the proportion of the total variance explained by the fixed effects, and the conditional R^2, which is the proportion of the variance explained by both fixed and random effects. Either the marginal or conditional R^2 can then be converted to other effect sizes that may be more common in your particular research literature. For example, the model R^2 can be used to compute Cohen's f^2 (Cohen, 1992) using:

$$f^2 = \frac{R^2}{1 - R^2}$$

To estimate the variance explained by a particular predictor (i.e., to obtain an estimate of the local effect size), several methods exist. One method is to estimate Cohen's f^2 for each local effect by estimating R^2 for two nested models:

$$f^2 = \frac{R_2^2 - R_1^2}{1 - R_2^2}$$

where R_2^2 represents the variance explained by a model with the effect of interest (the full model) and R_1^2 represents the variance explained by a model without the effect of interest (the restricted model). Cohen's f^2 for a local effect can easily be directly calculated using this method in SAS (Selya et al., 2012) and in R by fitting each model separately and estimating the pseudo-R^2 as mentioned previously using the r.squaredGLMM() function in the MuMIn packag (Bartoń, 2020) or the r2beta() function in the r2glmm package (Jaeger, 2017).

An alternative method is to calculate a *partial* R^2 statistic for each predictor, R_β^2 (Edwards et al., 2008). One R_β^2 statistic can be calculated for each predictor using the ANOVA output of an MLM model to get the *F*-statistic, the numerator (effect) degrees of freedom, and the denominator (residual or error) degrees of freedom that correspond to each predictor.

$$R_\beta^2 = \frac{\left(\frac{df_{numerator}}{df_{denominator}}\right)F}{1 + \left(\frac{df_{numerator}}{df_{denominator}}\right)F}$$

In the realm of ERPs, it remains unclear how large of an effect is meaningful, as meaningful differences in amplitude may vary depending on the ERP component of interest and the variance in the outcome is related to a number of factors, including the noisiness of the data and how many trials are included in each averaged waveform. Thus,

descriptions of effect size in future ERP studies are essential to triangulate what may be a meaningful effect size in the study of ERPs.

3. Visualizing data

In addition to statistical tests, visualizing data is an important component of understanding statistical results. As most ERP studies are interested in the effect of categorical predictors, a common approach using rANOVA is to use bar or line graphs to depict mean amplitude averaged across participants and channels in each condition. However, depicting averages from the data does not account for the multilevel structure of the data, nor does it depict how much variability in the effect exists across people. When using multilevel modeling, we can plot 1) the fixed effects estimates to summarize patterns across the whole sample, 2) the variance within each grouping variable (e.g., how participants vary from each other), or both. To plot mean differences across experimental conditions and still account for the multilevel structure, we can plot the model-estimated means from the fixed effects (alternatively called estimated marginal means). For both bar and line charts, this should include the values of the outcome variable that are predicted from your model for each condition (i.e., estimated means) and the standard errors of these model-estimated means as error bars. Estimated means can be calculated using a user-friendly, online tool available at http://www.quantpsy.org/interact/ (Preacher et al., 2006) or the emmeans package in R (Lenth, 2020) and then plotted as in Fig. 3.

However, one benefit of MLM is being able to estimate unique effects for each unit in a random factor (e.g., participant) by including random slopes in the model. To visually represent the variance in a particular effect, plot the *best linear unbiased predictions* (BLUPs) estimated for each participant using a "spaghetti plot". Spaghetti plots illustrate the variance in the effect, which we can see in differences in the slopes of the lines. If all lines are relatively parallel, there is little variance in the effect of Response Type across participants (which will be reflected in a small estimate of variance of the random slope of Response Type by participant), whereas lots of intersecting lines that are not parallel suggest a large amount of variance in the random slope. We suggest plotting each line with an opacity level below 100% to make each line easier to see and consider making each line its own color, if color visualizations are an option for your publishing outlet of choice (see Fig. 4 for an example). As we can see in this example, most participants show the same pattern as the fixed effect (more negative ERN amplitude in the incorrect condition compared to the correct condition), but some slopes are flatter than others, and some participants even show an effect in the opposite direction.

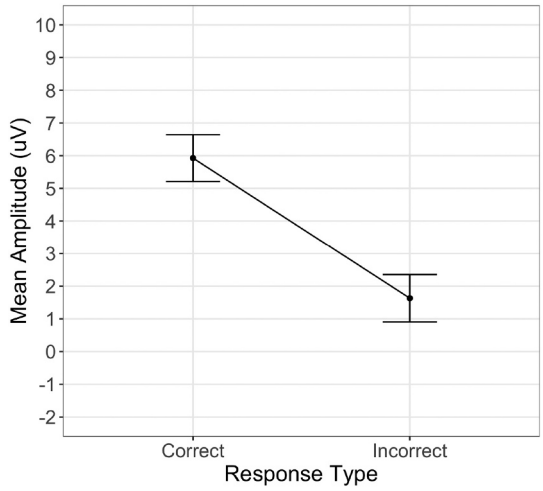

Fig. 3. Model-estimated means plot illustrating fixed effect of response type.

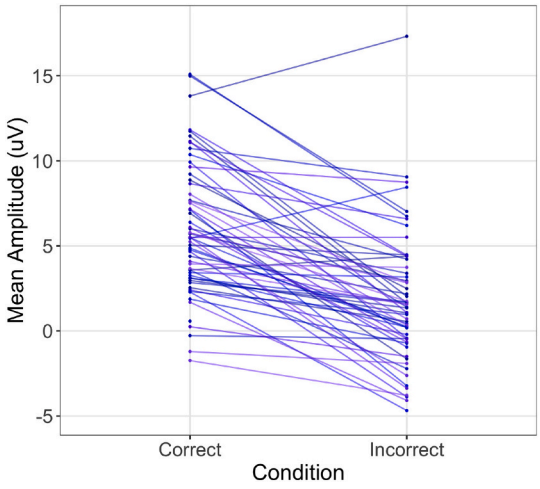

Fig. 4. Spaghetti plot illustrating variance in effect of Condition across participants (i.e., random slopes).

Of course, one can also plot both the estimated means for each condition and variance across individuals by overlaying the two plots. We suggest plotting the "average" effect (i.e., the fixed effect) in a slightly thicker width or different color to make it stand out (e.g., see Fig. 5).

4. Extended applications of MLM

One of the major benefits of MLM that rANOVA cannot accommodate is including continuous variables in the model. One example of this is testing how individual difference variables moderate the effect of the manipulated predictor. Past work has shown a link between trait anxiety and the size of the ERN/CRN, such that those who are more anxious show a more pronounced negativity following errors (Hajcak et al., 2003; Weinberg et al., 2010; Meyer, 2017). To test the effect of trait anxiety on the size of the ERN/CRN using MLM, we can simply include trait anxiety as a predictor in the model (Response Type is effect coded; Correct = −1, Incorrect = 1):

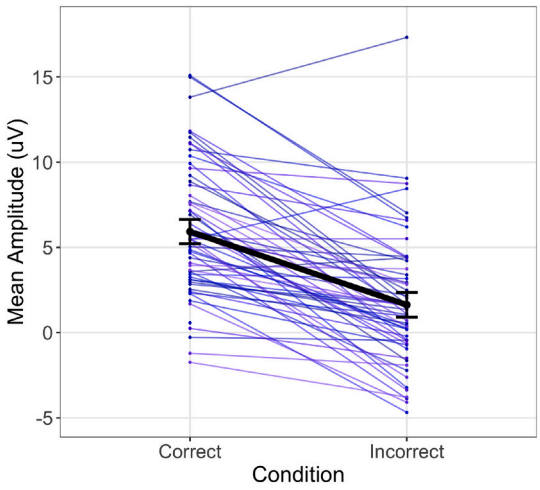

Fig. 5. Spaghetti plot with model estimated means overlaid.
Note. The thick black line represents the average relationship estimated by the fixed effect and the thinner, multicolored lines represent the specific relationships estimated for each person (the random slopes).

$$\text{MeanAmp} \sim \text{RespType} * \text{Anx} + (\text{RespType}|\text{Participant}) + (1|\text{Channel})$$

As mentioned in an earlier footnote, we would not include anxiety as a random slope by participant because there is only one observation per participant (and is thus invariant). In this model, the effect of Response Type remains significant, $b = -1.87$, 95% CIs [−2.61, −1.12], $t(58.0) = -4.92$, $p < .001$, such that mean amplitude is more negative following incorrect responses than correct responses. The effect of trait anxiety is marginally significant, $b = 0.19$, 95% CIs [−0.01, 0.39], $t(58.0) = 1.91$, $p = .061$. Most importantly, to examine whether trait anxiety moderates reactivity to errors, we would look at the Response Type x Anxiety interaction. In this sample, the interaction is not significant, $b = -0.05$, 95% CIs [−0.15, 0.05], $t(58.0) = -0.93$, $p = .356$. The interaction provides a similar test as correlating trait anxiety scores with a difference score of the ERN and CRN (e.g., ΔERN). Previous research has shown differences in ΔERN between anxious and control groups (Ladouceur et al., 2006; Pasion and Barbosa, 2019; Weinberg et al., 2010, 2012, 2015) and significant relationships between symptoms of generalized anxiety disorder and ΔERN (Bress et al., 2015; Klawohn et al., 2020), such that more anxious participants show a more negative ERN relative to the CRN, although other studies have not found consistent significant correlations between self-reported anxiety and ΔERN (e.g., Meyer et al., 2012).

Another application that MLM allows for is the investigation of ERP responses to specific stimuli or events from individual trials, allowing researchers to investigate how ERP signals meaningfully change over the course of different trials or meaningfully differ in response to specific instantiations of stimulus presentations. As mentioned previously, prior to data analysis researchers typically average all responses elicited by stimuli of the same type or experimental condition (i.e., signal averaging; Luck, 2014), which results in a data structure in which each participant has a single observation per channel for each experimental condition. This technique is effective for isolating physiological responses to events of interest (i.e., increasing signal-to-noise ratio) but makes assumptions that might not be tenable, including that the signal is constant across trials, and that any trial-to-trial variation is solely the result of noise, and therefore meaningless. A number of factors, including habituation, fatigue, sensitization, or momentary lapses in attention can result in meaningful variation (i.e., not merely noise) in ERPs across trials, thereby undermining the validity of signal averaging in some situations.

A number of approaches to analyzing single trial ERPs have been proposed (Blankertz et al., 2011; Coles et al., 1985; Debener et al., 2005; Gaspar et al., 2011; Jung et al., 2001; Pernet et al., 2011b; Philiastides et al., 2006; Quiroga and Garcia, 2003; Ratcliff et al., 2009; Regtvoort et al., 2006; Rousselet et al., 2011; Sassenhagen et al., 2014). Multilevel modeling provides an extremely useful additional tool for researchers interested in trial-level variation in ERPs. Note, however, that because noise is not first being removed from the waveforms using the signal averaging approach, it is important that the EEG data are as clean as possible when a trial-level approach is used. Researchers should spend additional time and effort during the data collection process to ensure the highest quality data possible to reduce noise in the data and increase the ability of multilevel models to detect fixed effects of interest from individual trials.

To examine the linear effect of time on change in psychological processes, researchers can include time or trial number as an additional fixed predictor in the model (e.g., Berry et al., 2019; Brush et al., 2018; Volpert-Esmond et al., 2018). As an example, the model may be specified as:

$$\text{DV} \sim \text{IV} + \text{Trial} + (\text{IV}|\text{Participant}) + (1|\text{Channel})$$

Note that the inclusion of Trial in this way can only capture long-range trends such as habituation and fatigue. Quadratic and other non-linear effects can be included as additional predictors, although little research has been done in this area and polynomial fitting comes with its own set of challenges (Kristjansson et al., 2007; Tremblay and Newman, 2015). By examining the fixed effect of time, or interactions between time and other fixed predictors, researchers can infer large-scale change in the amplitude or latency of ERP components over the course of an experiment, as well as different rates of change for different experimental conditions. Additionally, the variable indexing time can be included as a random slope by participant so that researchers can examine how the effect of time (including processes such as habituation or learning) differs across participants. To get estimates of individual differences in the rate of change in ERPs, researchers can extract the BLUPs, which are participant-specific estimates of the effect of time. However, including time as a random slope often results in non-convergence issues, which must be addressed before interpreting the BLUPs. Last, using MLM with single-trial ERPs opens the door to using ERP amplitude or latency as a predictor of other trial-level variables (such as reaction time or other downstream ERP components; Volpert-Esmond and Bartholow, 2020; Von Gunten et al., 2018).

Including continuous variables introduces increased complexity surrounding issues of centering variables that are unique to MLM. In typical single-level OLS regression, researchers often center and/or standardize continuous variables in order to interpret all other fixed effects as the effect observed at the mean of the centered variable. We suggest taking a similar approach to all continuous Level 2 variables (e.g., individual difference variables). However, in multilevel data, continuous Level 1 variables can either be centered across the entire data set (grand-mean centering) or centered within each level of the grouping variables (group-mean centering). The type of centering one chooses can significantly impact the interpretation of the fixed effects. There are a number of other resources discussing centering (e.g., Brauer and Curtin, 2018; Enders and Tofighi, 2007; Kreft et al., 1995; Paccagnella, 2006; Page-Gould, 2017) and contrast coding (Schad et al., 2020) within multilevel data.

One particular case of centering that may be of interest may be disaggregating between- and within-participant effects of a continuous predictor (e.g., Curran and Bauer, 2011). This is particularly relevant when using single-trial ERPs as a continuous predictor of some other outcome, such as a behavioral response within the same trial or an ERP response on a subsequent trial. In the absence of disaggregation, the relationship between single-trial ERPs and reaction time (for example) conflates the between-person effect (i.e., Do people with particularly large ERP responses also respond faster to stimuli in a particular task?) and the within-person effect (i.e., Does a larger ERP response on a particular trial, relative to a person's average ERP response, facilitate a faster reaction time?). Depending on the theoretical question, researchers may be more interested in one relationship than the other. To disaggregate within- and between-person effects, the researcher can effectively separate the predictor variable of interest into two separate predictors. The first predictor—each participants' mean—is entered as a Level-2 (person level) predictor and represents the between-person effect. The second predictor—the participant-centered variable—is entered as a Level-1 predictor and represents the within-person effect.

5. Reporting practices

Because of the complexity surrounding MLMs, researchers have a number of degrees of freedom with respect to how MLMs are estimated and reported, including what covariance structure to use, which variables to include as fixed and random effects, how to test for interactions, how to center or effect-code variables, etc. Because of the flexibility of these models, it is imperative to provide enough information for an independent party to replicate the analysis and evaluate its suitability for the dataset at hand. Of course, providing code in an online repository such as Open Science Framework or GitHub is preferable. But we encourage researchers to include all essential information in the manuscript as well. At minimum, the entire model, including variance-covariance structure and random effects should be described (Meteyard

and Davies, 2020). To most effectively communicate the structure of each model used, we suggest using Wilkinson notation, which specifies the DV, IVs, and random effects. For example,

$$DV \sim IV1 + IV2 + (1 + IV1|Participant) + (1|Participant : Channel)$$

specifies that two predictors were included, but not their interaction; that the intercept and the effect of the first predictor was allowed to vary by participant (i.e., IV1 was included as a random slope by participant); and that the intercept was allowed to vary by channel nested within subject. Alternatively, the following model specifies participants and channels as crossed random factors:

$$DV \sim IV1 + IV2 + (1 + IV1|Participant) + (1|Channel)$$

R users will recognize that Wilkinson notation is used in the lme4 package to specify models (and is also used in Matlab), thus providing less of a barrier than formal mathematical notation. The statistical software used to fit the models should additionally be reported, along with any changes to the default specifications (e.g., which covariance structure is specified). More extensive recommendations about reporting practices regarding model selection, model output, etc., can be found in Meteyard and Davies (2020).

In addition to reporting the structure of the models using Wilkinson notation, we suggest reporting the intraclass correlation coefficient (ICC) for each random factor, which can be calculated from the variances estimated in the random effects:

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2}$$

where τ^2 is the between-cluster variance (the variance associated with the random factor) and σ^2 is the residual variance (Lorah, 2018). This gives you the proportion of total variation in the data that is accounted for by a particular random factor where higher ICCs represent more variance between units within that random factor (Gelman and Hill, 2007). Since the complexity of the model affects the calculation of ICC, you should use variance estimates from an *intercept-only model* (i.e., a model with no fixed predictors):

$$DV \sim 1 + (1|Participant) + (1|Channel)$$

When including more than one random factor (e.g., including participants and channels in a cross-classified model), one would include the variance of all groups in the denominator. As an example, let's look at sources of variance in the mean amplitude of the P2 ERP component elicited by Black and White male faces during a race categorization task.[4] To calculate the ICC associated with subject, we would look at the output for the random effects from the following intercept-only model,[5] first using the signal averaged data:

$$Model : P2amp \sim 1 + (1|Participant) + (1|Channel)$$

Random effects output:

Random factors	Name	Variance	Std. dev.
Participant	(Intercept)	8.314	2.883
Channel	(Intercept)	0.141	0.376
Residual		2.306	1.5184

[4] Data were previously published in Volpert-Esmond et al. (2017). Although the original study manipulated where participants fixated on the face, data used here include only trials presented so that participants fixated in a typical location (i.e., between the eyes). The sample includes 65 participants and the average number of trials included per participant was 107.7 (min = 54, max = 127). Data from 7 channels were used (C3, C4, CP3, CP4, CPz, Cz, Pz).

[5] This is an example of a cross-classified model, where subject and channel are included as separate grouping variables, rather than channel being nested within subjects in a typical hierarchical model. Calculating ICCs for groups nested within each other is similar (i.e., estimates of variance for all groups plus the residual variance is used in the denominator).

$$ICC_{par} = \frac{8.314}{8.314 + 0.141 + 2.306} = 0.77$$

$$ICC_{elec} = \frac{0.141}{8.314 + 0.141 + 2.306} = 0.01$$

In other words, variance between people accounts for 77% of the total variance, suggesting there is a lot of between-person variability in ERPs, whereas variance between channels accounts for 1% of the total variance, suggesting is there not a lot of variability between channels, which is expected given similarities in waveforms at adjacent channels. In contrast, when using trial level data, the ICC associated with subject is 0.09, suggesting between-person variability only accounts for 9% of the total variance. Because of the amount of within-person variance from trial to trial, between-person variance accounts for much less of the total variance when using single-trial ERPs instead of signal-averaged ERPs.

6. Estimating power

Another barrier in transitioning to using MLM is the daunting prospect of having to do a power analysis. Evaluating the power of a hypothesis test, which is defined as the probability that the test will correctly reject the null hypothesis when the null hypothesis is false, is important in assessing how likely a particular result is true and able to be replicated. Additionally, ERP studies are often underpowered to find small effects (Clayson et al., 2019). Given that estimating an effect of zero—or estimating effects completely at random—is more accurate at determining the true population mean than using sample means derived from poorly powered studies (Davis-Stober et al., 2018), and that EEG studies are time-intensive and costly to run, an a priori power analysis can inform a researcher whether they have the resources to conduct a study that is well-powered enough to be informative. Additionally, according to recent guidelines for best practices in reporting of ERP studies (Keil et al., 2014), researchers should always report the achieved power of a particular design. Many tools are available to estimate power for typical single-level designs (e.g., Faul et al., 2009; Murphy et al., 2014), although discussion is still ongoing about the most appropriate ways to conduct and use a power analysis (Anderson et al., 2017; Cribbie et al., 2019; Albers & Lakens, 2018; Lakens & Evers, 2014).

In multilevel designs, how power relates to sample size is more complicated, as both the number of groups (e.g., the number of individuals who participate during the study) and the number of observations per group (e.g., the number of trials or observations per individual) can vary. In multilevel models, power is affected by group sample size, observation sample size within each group, the ICC associated with group, whether you are testing a Level 1 (observation-level) or Level 2 (participant-level) effect, and numerous other parameters of the model (Arend and Schäfer, 2019). As a general rule of thumb, increasing the number of Level 2 units (e.g., the number of people participating in the study) has a larger effect on power to detect fixed effects than increasing the number of Level 1 units (e.g., the number of experimental conditions or trials within each participant; Maas and Hox, 2005; Snijders, 2005). For a more specific approximation of the sample size (at both Level 1 and Level 2) needed to achieve the desired level of power for a particular test, most researchers use a simulation approach to power using Monte Carlo simulations. This approach repeatedly simulates data from the hypothetical distribution that we expect our sampled data to come from and then fits the same multilevel model to each data set. Power is estimated by how often the true effect is detected.

To set up a power simulation, you need to make assumptions about the true treatment effect and also specify all the other parameters that characterize the study, including the size of the fixed effect of interest, ICCs of any random grouping variables, variances of random intercepts and slopes, correlations between random intercepts and slope, etc. Because of the large number of parameters needed to simulate an

appropriate data set, it is often easier to conduct a power simulation on a set of pilot data, although parameters can be assumed and simulated without pilot data (Gelman and Hill, 2007). The simr package in R (Green and MacLeod, 2016) has emerged as a popular tool for power simulations using multilevel models. The package allows users to input a sample data set (either a pilot or simulated data set) and calculate observed power for a desired effect, as well as produce power curves in which power is plotted as a function of a particular aspect of the design, such as number of participants, number of observations within each participant, or effect size. To provide an example of a power curve generated using simr, we use previously published data[6] looking at how the race of a face influences mean P2 amplitude:

$$\text{Model}: \text{P2amp} \sim \text{Race} + (\text{Race}|\text{Participant}) + (1|\text{Channel})$$

When using signal averaged data, the effect of race is significant, $b = -0.59$, $t(64.0) = -4.82$, $p < .001$, such that Black male faces elicit larger P2s than White male faces. A post-hoc power simulation indicates that the observed power for this effect with the given sample size (65 participants, 7 channels for each participant, and 2 observations for each channel) is 99.6%, suggesting this design is very well-powered to detect this effect. Fig. 6 shows a power curve demonstrating the decrease in power as the sample size decreases.

Thus, we achieve 80% power to detect an effect of this size with about 25 participants. However, using pilot data to estimate the true effect size may result in an underpowered follow-up study (Albers & Lakens, 2018; Anderson et al., 2017). Thus, we suggest either adjusting the anticipating effect size to be smaller than that achieved in a pilot study when planning a follow up study or producing a sensitivity power curve to identify what sample size would be needed to detect the smallest meaningful effect size. Fig. 7 demonstrates the decrease in power as the effect size decreases, indicating that with a sample of this size, we would be able to detect an effect as small as $b = -0.35$ with 80% power.

7. Limitations

MLM is not a panacea. As with any analytic approach, MLM comes with significant limitations. First, effectively using MLM involves gaining the expertise to organize data in the appropriate format, learning how to implement models in statistical software, making appropriate decisions for model specification, correctly interpreting the output, etc. Additionally, due to the continued evolution and development of knowledge about MLMs, there currently is a perceived lack of consensus and established, standardized procedures (Meteyard and Davies, 2020). Many resources are becoming available for researchers interested in learning this statistical approach, including workshops at prominent conferences (i.e., Society for Psychophysiological Research), standalone workshops hosted by societies, private organizations, and universities (e.g., APA Advanced Training Institutes, Statistical Horizons, University of Michigan, University of North Carolina, University of Connecticut, Arizona State University), and numerous tutorials and articles on applying MLM to both behavioral data (Arnau et al., 2010; Baayen et al., 2008; Brauer and Curtin, 2018; Gueorguieva and Krystal, 2004; Jaeger, 2008; Judd et al., 2012; Maas and Snijders, 2003; Quené and van den Bergh, 2004, 2008) and psychophysiological data (Bagiella et al., 2000; Kristjansson et al., 2007; Page-Gould, 2017; Tibon and Levy, 2015; Tremblay and Newman, 2015; Volpert-Esmond et al., 2018; Vossen et al., 2011). However, little is known about the effectiveness of this training and how it is implemented in practice (King et al., 2019). Moreover, the mere fact that these opportunities exist does not ensure that researchers will or can take advantage of them, and therefore this situation is far from ideal in terms of ensuring adequate quantitative methods training in the field—likely contributing to a significant gap in psychologists' quantitative training. Thus, learning how to appropriately apply MLM to ERP data may be a significant barrier.

In addition to the time cost of learning the approach, MLM is often quite computing-power intensive and models can take much longer than a typical rANOVA to run. In the case of the P2 example given previously, the first author ran these models on a MacBook Air with a 1.6 GHz Dual-Core Intel Core i5 processer with 8 GB of RAM. To test the effect of face race on P2 amplitude using signal-averaged data, it took only a few seconds to fit the model. To run the same model using trial-level data, it took less than 10 s to fit the model. However, as the data set becomes larger and the model becomes more complex, the time required to fit a MLM increases dramatically. For example, this more complex model testing the effect of target race, target gender, fixation, task, and participant race on trial-level P2 data recorded in two face processing

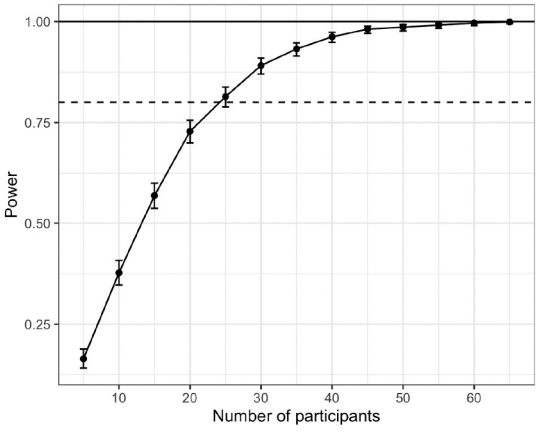

Fig. 6. Power to detect the fixed effect of race on P2 amplitude as a function of sample size.
Note. 14 observations are included per each participant (7 channels with 2 observations at each channel). The effect of race is set at $b = -0.59$ (the observed effect size in the data).

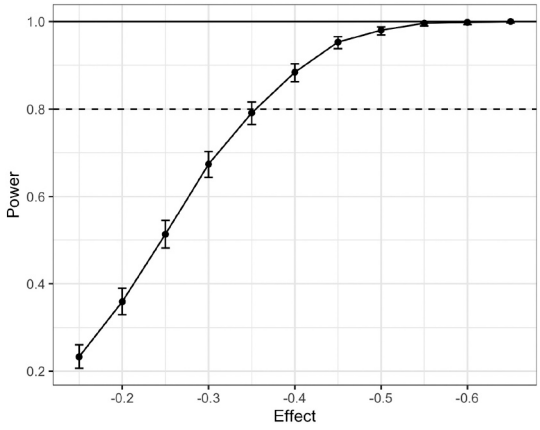

Fig. 7. Power to detect the fixed effect of race on P2 amplitude as a function of effect size.
Note. 14 observations are included for each participant (7 channels with 2 observations at each channel). The sample size is set at 65 participants.

[6] Same data as used in ICC example.

tasks (256 trials in each task)[7] took over 12 h to fit:

P2amp~TarRace*TarGender*Fix*Task*ParRace + (TarRace*TarGender*Fix*Task|Participant) + (1|Participant : Channel)

Similarly, because power simulations require fitting the same model to multiple simulated data sets, power simulations can take quite some time. To produce the power curve for the signal averaged data shown in Fig. 1 (varying the sample size), the power simulation took roughly 90 min to run. The power curve shown in Fig. 2 (varying the size of the effect) took roughly 3 h to run. To produce the same power curve depicted in Fig. 1 for trial-level data rather than signal-averaged data, the first author did not have enough processing power—when she attempted it, the simulation ran for one week straight (24 h/day) and was only a quarter of the way finished before the author stopped the simulation. Thus, processing power is essential for running MLMs, especially when using hundreds of trial-level observations, which is increased 20-fold when running power simulations. Given limitations in researchers' access to powerful computers or server clusters, this may be a significant limitation in the use of MLMs for ERP data, especially when using trial-level data.

Lastly, researchers may encounter estimation problems when running the model, the most common of which are convergence problems, very long estimation times, and singularity issues. Most estimation problems can be addressed with two guiding principles: *Simpler models* and *better fitting models* will have fewer estimation problems. Generally speaking, the more complex your random effects are, the more difficult the model parameters are to estimate. As mentioned earlier, one approach is to trim the most complex random effect (e.g., random slope for an interaction term), run the model again, and then progressively trim random effects in order of decreasing complexity until the model converges.

Additionally, the time that it will take to run an MLM is directly related to how well your model specification reflects reality. If it takes a very long time to run an MLM but you received a warning (e.g., the Hessian matrix was not positive-definite), look at the variance estimates in the random effects to see if any variances are zero or very close to zero. If so, then it means that the random effect does not vary much from group to group, and thus is an over-specification. Trim any random effects that have zero variance, and run the model again.

The process of troubleshooting estimation problems is done iteratively (i.e., remove one term, rerun the MLM), so that the model does not get oversimplified in the process. Given how many choices that can be made when handling these problems, it is a good idea to establish a common approach to troubleshooting estimation problems that you apply across all your research studies that use MLM and reporting your approach in your method sections.

8. Conclusion

Despite its limitations, MLM has great potential to advance the study of neurocognitive processing through advanced modeling of psychophysiological data. We have focused here on the use of MLM for ERP data, but of course MLM also can be used with other psychophysiological data that are structured similarly (e.g., Bourassa et al., 2016; Briollais et al., 2003; de Looff et al., 2019). Not only can MLM address limitations inherent to the use of rANOVA with such data, such as violations of assumptions leading to inflated Type 1 error rate and the need for listwise deletion when observations are missing, MLM can greatly expand the types of research questions that can be posed and tested with psychophysiology. For example, here we highlighted that MLM permits examination of change over time in physiological and behavioral responses using trial-level data, and also how trial-level data can be used to test for within- and between-trial associations among dependent measures (e.g., ERP amplitudes predicting reaction time; Volpert-Esmond and Bartholow, 2020). Even if researchers stick with a traditional signal averaging approach to signal processing, MLM affords more precise modeling of effects and more appropriate parsing of error variance (e.g., by estimating random slopes differing across individuals and independent variable conditions) than does rANOVA. Thus, we recommend that researchers invest the time—and, if necessary, the resources to bolster their computing power—to learn MLM, and discover the flexibility the technique affords for their psychophysiological research programs. Finally, we strongly recommend that researchers who adopt MLM pay close attention to the latest developments in the rapidly evolving literature on best practices in the use of the technique and its limitations.

[7] Data previously reported in Volpert-Esmond and Bartholow (2019). Data set includes 65 participants and the average number of trials included per participant was 201.5 (min = 131, max = 250) in the gender categorization task and 210.2 (min = 130, max = 252) in the race categorization task. Data from 11 channels were used (C1, C2, C3, C4, CP1, CP2, CP3, CP4, CPz, Cz, Pz).

References

Albers, C., Lakens, D., 2018. When power analyses based on pilot data are biased: inaccurate effect size estimators and follow-up bias. J. Exp. Soc. Psychol. 74, 187–195.

Anderson, S.F., Kelley, K., Maxwell, S.E., 2017. Sample-size planning for more accurate statistical power: a method adjusting sample effect sizes for publication bias and uncertainty. Psychol. Sci. 28 (11), 1547–1562.

Arend, M.G., Schäfer, T., 2019. Statistical power in two-level models: a tutorial based on Monte Carlo simulation. Psychol. Methods 24 (1), 1–19. https://doi.org/10.1037/met0000195.

Arnau, J., Bono, R., Balluerka, N., Gorostiaga, A., 2010. General linear mixed model for analysing longitudinal data in developmental research. Percept. Mot. Skills 110 (2), 547–566. https://doi.org/10.2466/pms.110.2.547-566.

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

Bagiella, E., Sloan, R.P., Heitjan, D.F., 2000. Mixed-effects models in psychophysiology. Psychophysiology 37 (1), 13–20. https://doi.org/10.1111/1469-8986.3710013.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68 (3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bartoń, K., 2020. MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bell, A., Fairbrother, M., Jones, K., 2019. Fixed and random effects models: making an informed choice. Qual. Quant. 53 (2), 1051–1074. https://doi.org/10.1007/s11135-018-0802-x.

Berry, M.P., Tanovic, E., Joormann, J., Sanislow, C.A., 2019. Relation of depression symptoms to sustained reward and loss sensitivity. Psychophysiology 56 (7), e13364. https://doi.org/10.1111/psyp.13364.

Blair, R.C., Karniski, W., 1993. An alternative method for significance testing of waveform difference potentials. Psychophysiology 30 (5), 518–524. https://doi.org/10.1111/j.1469-8986.1993.tb02075.x.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.-R., 2011. Single-trial analysis and classification of ERP components—a tutorial. NeuroImage 56 (2), 814–825. https://doi.org/10.1016/j.neuroimage.2010.06.048.

Boisgontier, M.P., Cheval, B., 2016. The anova to mixed model transition. Neurosci. Biobehav. Rev. 68, 1004–1005. https://doi.org/10.1016/j.neubiorev.2016.05.034.

Bourassa, K.J., Hasselmo, K., Sbarra, D.A., 2016. Heart rate variability moderates the association between separation-related psychological distress and blood pressure reactivity over time. Psychol. Sci. 27 (8), 1123–1135. https://doi.org/10.1177/0956797616651972.

Brauer, M., Curtin, J., 2018. Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. Psychol. Methods 23 (3), 389–411. https://doi.org/10.1037/met0000159.
Bress, J.N., Meyer, A., Hajcak, G., 2015. Differentiating anxiety and depression in children and adolescents: evidence from event-related brain potentials. J. Clin. Child Adolesc. Psychol. 44 (2), 238–249. https://doi.org/10.1080/15374416.2013.814544.
Briollais, L., Tzontcheva, A., Bull, S., 2003. Multilevel modeling for the analysis of longitudinal blood pressure data in the Framingham Heart Study pedigrees. BMC Genet. 4 (1), S19. https://doi.org/10.1186/1471-2156-4-S1-S19.
Brush, C.J., Ehmann, P.J., Hajcak, G., Selby, E.A., Alderman, B.L., 2018. Using multilevel modeling to examine blunted neural responses to reward in major depression. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 3 (12), 1032–1039. https://doi.org/10.1016/j.bpsc.2018.04.003.
Clayson, P.E., Carbine, K.A., Baldwin, S.A., Larson, M.J., 2019. Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. Psychophysiology 56 (11), e13437.
Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155.
Coles, M.G., Gratton, G., Bashore, T.R., Eriksen, C.W., Donchin, E., 1985. A psychophysiological investigation of the continuous flow model of human information processing. Journal of Experimental Psychology. Human Perception and Performance 11 (5), 529–553. https://doi.org/10.1037//0096-1523.11.5.529.
Cribbie, R., Beribisky, N., Alter, U., 2019. A multi-faceted mess: a review of statistical power analysis in psychology journal articles. PsyArxiv.
Curran, P.J., Bauer, D.J., 2011. The disaggregation of within-person and between-person effects in longitudinal models of change. Annu. Rev. Psychol. 62 (1), 583–619. https://doi.org/10.1146/annurev.psych.093008.100356.
Davis-Stober, C.P., Dana, J., Rouder, J.N., 2018. Estimation accuracy in the psychological sciences. PLoS One 13 (11), e0207239. https://doi.org/10.1371/journal.pone.0207239.
de Looff, P., Noordzij, M.L., Moerbeek, M., Nijman, H., Didden, R., Embregts, P., 2019. Changes in heart rate and skin conductance in the 30 min preceding aggressive behavior. Psychophysiology 56 (10), e13420. https://doi.org/10.1111/psyp.13420.
Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Cramon, D.Y. von, Engel, A.K., 2005. Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. J. Neurosci. 25 (50), 11730–11737. https://doi.org/10.1523/JNEUROSCI.3286-05.2005.
Dempster, A.P., Rubin, D.B., Tsutakawa, R.K., 1981. Estimation in covariance components models. J. Am. Stat. Assoc. 76 (374), 341–353. https://doi.org/10.1080/01621459.1981.10477653.
Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F., Schabenberger, O., 2008. An R2 statistic for fixed effects in the linear mixed model. Stat. Med. 27 (29), 6137–6157. https://doi.org/10.1002/sim.3429.
Enders, C.K., Tofighi, D., 2007. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. Psychol. Methods 12 (2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121.
Faul, F., Erdfelder, E., Buchner, A., Lang, A., 2009. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav. Res. Methods 41, 1149–1160.
Games, P.A., 1976. Programs for robust analyses in ANOVA's with repeated measures. Psychophysiology 13 (6), 603. https://doi.org/10.1111/j.1469-8986.1976.tb00890.x.
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.
Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
Goedert, K.M., Boston, R., Barrett, A.M., 2013. Advancing the science of spatial neglect rehabilitation: an improved statistical approach with mixed linear modeling. Front. Hum. Neurosci. 7 https://doi.org/10.3389/fnhum.2013.00211.
Goldstein, H., 1987. Multilevel covariance component models. Biometrika 74 (2), 430–431. https://doi.org/10.1093/biomet/74.2.430.
Graham, J.W., 2009. Missing data analysis: making it work in the real world. Annu. Rev. Psychol. 60 (1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530.
Green, P., MacLeod, C.J., 2016. SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods Ecol. Evol. 7 (4), 493–498. https://doi.org/10.1111/2041-210X.12504.
Greenhouse, S.W., Geisser, S., 1959. On methods in the analysis of profile data. Psychometrika 24 (2), 95–112. https://doi.org/10.1007/BF02289823.
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. Arch. Gen. 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.
Haverkamp, N., Beauducel, A., 2017. Violation of the sphericity assumption and its effect on type-I error rates in repeated measures ANOVA and multi-level linear models (MLM). Front. Psychol. 8 https://doi.org/10.3389/fpsyg.2017.01841.
Heisig, J.P., Schaeffer, M., 2019. Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. Eur. Sociol. Rev. 35 (2), 258–279. https://doi.org/10.1093/esr/jcy053.
Holroyd, C.B., Coles, M.G., 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. 109 (4), 679–709.
Huang, F., 2018. Multilevel Modeling Myths. Sch. Psychol. Q. 33 (3), 492–499. https://doi.org/10.1037/spq0000272.
Huynh, H., Feldt, L.S., 1970. Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. J. Am. Stat. Assoc. 65 (332), 1582–1589. https://doi.org/10.1080/01621459.1970.10481187.
Huynh, H., Feldt, L.S., 1980. Performance of traditional f tests in repeated measures designs under covariance heterogeneity. Communications in Statistics - Theory and Methods 9 (1), 61–74. https://doi.org/10.1080/03610928008827859.
Jaeger, T.F., 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. J. Mem. Lang. 59 (4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007.
Jaeger, B., 2017. r2glmm: Computes R Squared for Mixed (Multilevel) Models. R package version 0.1.2. https://CRAN.R-project.org/package=r2glmm.
Jennings, J.R., Allen, B., 2017. Methodology. In: Cacioppo, J.T., Tassinary, L.T., Berntson, G.G. (Eds.), Handbook of Psychophysiology, 4th ed. Cambridge University Press, pp. 583–627.
Jennings, J. Richard, Wood, C.C., 1976. The e-adjustment procedure for repeated-measures analyses of variance. Psychophysiology 13 (3), 277–278. https://doi.org/10.1111/j.1469-8986.1976.tb00116.x.
Johnson, P.C.D., 2014. Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. Methods Ecol. Evol. 5 (9), 944–946. https://doi.org/10.1111/2041-210X.12225.
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. J. Pers. Soc. Psychol. 103 (1), 54–69. https://doi.org/10.1037/a0028347.
Jung, T.P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., Sejnowski, T.J., 2001. Analysis and visualization of single-trial event-related potentials. Hum. Brain Mapp. 14 (3), 166–185. https://doi.org/10.1002/hbm.1050.
Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E.S., Luck, S.J., Yee, C.M., 2014. Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. Psychophysiology 51 (1), 1–21.
Keselman, H.J., Rogan, J.C., 1980. Repeated measures F tests and psychophysiological research: controlling the number of false positives. Psychophysiology 17 (5), 499–503. https://doi.org/10.1111/j.1469-8986.1980.tb00190.x.
Kiebel, S.J., Friston, K.J., 2004. Statistical parametric mapping for event-related potentials: I. Generic considerations. NeuroImage 22 (2), 492–502. https://doi.org/10.1016/j.neuroimage.2004.02.012.
King, K.M., Pullmann, M.D., Lyon, A.R., Dorsey, S., Lewis, C.C., 2019. Using implementation science to close the gap between the optimal and typical practice of quantitative methods in clinical science. J. Abnorm. Psychol. 128 (6), 547–562. https://doi.org/10.1037/abn0000417.
Klawohn, J., Meyer, A., Weinberg, A., Hajcak, G., 2020. Methodological choices in event-related potential (ERP) research and their impact on internal consistency reliability and individual differences: an examination of the error-related negativity (ERN) and anxiety. J. Abnorm. Psychol. 129 (1), 29–37. https://doi.org/10.1037/abn0000458.
Kreft, I.G.G., Leeuw, J. de, Aiken, L.S., 1995. The effect of different forms of centering in hierarchical linear models. Multivar. Behav. Res. 30 (1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1.
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.
Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr. Ann. 32 (9), 509–515.
Krueger, C., Tian, L., 2004. A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. Biological Research For Nursing 6 (2), 151–157. https://doi.org/10.1177/1099800404267682.
Kuznetsova, A., Brockhoff, P.B., Christensen, R.H., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Softw. 82 (13), 1–26. https://doi.org/10.18637/jss.v082.i13.
Ladouceur, C.D., Dahl, R.E., Birmaher, B., Axelson, D.A., Ryan, N.D., 2006. Increased error-related negativity (ERN) in childhood anxiety disorders: ERP and source localization. J. Child Psychol. Psychiatry 47 (10), 1073–1082. https://doi.org/10.1111/j.1469-7610.2006.01654.x.
Lakens, D., Evers, E.R., 2014. Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. Perspect. Psychol. Sci. 9 (3), 278–292.
Lenth, R., 2020. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.2-1. https://CRAN.R-project.org/package=emmeans.
Litvak, V., Mattout, J., Kiebel, S., Phillips, C., Henson, R., Kilner, J., Barnes, G., Oostenveld, R., Daunizeau, J., Flandin, G., Penny, W., Friston, K., 2011. EEG and MEG data analysis in SPM8. Computational Intelligence and Neuroscience 2011. https://doi.org/10.1155/2011/852961.
Lorah, J., 2018. Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. Large-Scale Assess. Educ. 6 (1), 1–11.
Luck, S.J., 2014. An Introduction to the Event-Related Potential Technique, 2nd ed. MIT Press.
Luke, S.G., 2017. Evaluating significance in linear mixed-effects models in R. Behav. Res. Methods 49 (4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y.
Maas, C.J.M., Hox, J.J., 2005. Sufficient sample sizes for multilevel modeling. Methodology 1 (3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86.

Maas, C.J.M., Snijders, T.A.B., 2003. The multilevel approach to repeated measures for complete and incomplete data. Qual. Quant. 37 (1), 71–89. https://doi.org/10.1023/A:1022545930672.
Mardia, K.V., 1975. Assessment of multinormality and the robustness of Hotelling's T2. Test. J. R. Stat. Soc.: Ser. C: Appl. Stat. 24 (2), 163–171. https://doi.org/10.2307/2346563.
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D., 2017. Balancing type I error and power in linear mixed models. J. Mem. Lang. 94, 305–315. https://doi.org/10.1016/j.jml.2017.01.001.
McNeish, D., Stapleton, L., 2016a. The effect of small sample size on two-level model estimates: a review and illustration. Educ. Psychol. Rev. 28 (2), 295–314. https://doi.org/10.1007/s10648-014-9287-x.
McNeish, D., Stapleton, L.M., 2016b. Modeling clustered data with very few clusters. Multivar. Behav. Res. 51 (4), 495–518. https://doi.org/10.1080/00273171.2016.1167008.
McNeish, D., Stapleton, L.M., Silverman, R.D., 2017. On the unnecessary ubiquity of hierarchical linear modeling. Psychol. Methods 22 (1), 114–140. https://doi.org/10.1037/met0000078.
Meteyard, L., Davies, R.A.I., 2020. Best practice guidance for linear mixed-effects models in psychological science. J. Mem. Lang. 112, 104092 https://doi.org/10.1016/j.jml.2020.104092.
Meyer, A., 2017. A biomarker of anxiety in children and adolescents: A review focusing on the error-related negativity (ERN) and anxiety across development. Dev. Cogn. Neurosci. 27, 58–68.
Meyer, A., Weinberg, A., Klein, D.N., Hajcak, G., 2012. The development of the error-related negativity (ERN) and its relationship with anxiety: evidence from 8 to 13 year-olds. Developmental Cognitive Neuroscience 2 (1), 152–161. https://doi.org/10.1016/j.dcn.2011.09.005.
Murphy, K., Myors, B., Wolach, A., 2014. Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, 4th ed. Routledge.
Nakagawa, S., Schielzeth, H., 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol. Evol. 4 (2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.
Nakagawa, S., Johnson, P.C.D., Schielzeth, H., 2017. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. J. R. Soc. Interface 14 (134), 20170213. https://doi.org/10.1098/rsif.2017.0213.
Olvet, D.M., Hajcak, G., 2009. The stability of error-related brain activity with increasing trials. Psychophysiology 46 (5), 957–961.
Paccagnella, O., 2006. Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. Eval. Rev. 30 (1), 66–85. https://doi.org/10.1177/0193841X05275649.
Page-Gould, E., 2017. Multilevel modeling. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), The Handbook of Psychophysiology. Cambridge University Press, pp. 662–678.
Pasion, R., Barbosa, F., 2019. ERN as a transdiagnostic marker of the internalizing-externalizing spectrum: a dissociable meta-analytic effect. Neurosci. Biobehav. Rev. 103, 133–149. https://doi.org/10.1016/j.neubiorev.2019.06.013.
Pernet, C.R., Chauveau, N., Gaspar, C., Rousselet, G.A., 2011a. LIMO EEG: a toolbox for hierarchical LInear MOdeling of ElectroEncephaloGraphic data. Computational Intelligence & Neuroscience 1–11. https://doi.org/10.1155/2011/831409.
Pernet, C.R., Sajda, P., Rousselet, G.A., 2011b. Single-trial analyses: why bother? Front. Psychol. 2 https://doi.org/10.3389/fpsyg.2011.00322.
Philiastides, M.G., Ratcliff, R., Sajda, P., 2006. Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. J. Neurosci. 26 (35), 8965–8975. https://doi.org/10.1523/JNEUROSCI.1655-06.2006.
Preacher, K.J., Curran, P.J., Bauer, D.J., 2006. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. J. Educ. Behav. Stat. 31 (4), 437–448. https://doi.org/10.3102/10769986031004437.
Quené, H., van den Bergh, H., 2004. On multi-level modeling of data from repeated measures designs: a tutorial. Speech Comm. 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. J. Mem. Lang. 59 (4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002.
Quiroga, R.Q., Garcia, H., 2003. Single-trial event-related potentials with wavelet denoising. Clin. Neurophysiol. 114 (2), 376–390. https://doi.org/10.1016/S1388-2457(02)00365-6.
Rasbash, J., Goldstein, H., 1994. Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. J. Educ. Behav. Stat. 19 (4), 337–350. https://doi.org/10.2307/1165397.
Ratcliff, R., Philiastides, M.G., Sajda, P., 2009. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proc. Natl. Acad. Sci. 106 (16), 6539–6544. https://doi.org/10.1073/pnas.0812589106.
Regtvoort, A.G.F.M., van Leeuwen, T.H., Stoel, R.D., van der Leij, A., 2006. Efficiency of visual information processing in children at-risk for dyslexia: habituation of single-trial ERPs. Brain Lang. 98 (3), 319–331. https://doi.org/10.1016/j.bandl.2006.06.006.
Rousselet, G.A., Gaspar, C.M., Wieczorek, K.P., Pernet, C.R., 2011. Modelling single-trial ERP reveals modulation of bottom-up face visual processing by top-down task constraints (in some subjects). Front. Psychol. 2 https://doi.org/10.3389/fpsyg.2011.00137.
Sassenhagen, J., Schlesewsky, M., Bornkessel-Schlesewsky, I., 2014. The P600-as-P3 hypothesis revisited: single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. Brain Lang. 137, 29–39. https://doi.org/10.1016/j.bandl.2014.07.010.
Schad, D.J., Vasishth, S., Hohenstein, S., Kliegl, R., 2020. How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. J. Mem. Lang. 110, 104038 https://doi.org/10.1016/j.jml.2019.104038.
Seedorff, M., Oleson, J., McMurray, B., 2019. Maybe Maximal: Good Enough Mixed Models Optimize Power While Controlling Type I Error. *PsyArXiv*. https://doi.org/10.31234/osf.io/xmhfr.
Selya, A.S., Rose, J.S., Dierker, L.C., Hedeker, D., Mermelstein, R.J., 2012. A practical guide to calculating Cohen's f2, a measure of local effect size, from PROC MIXED. Front. Psychol. 3 https://doi.org/10.3389/fpsyg.2012.00111.
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.
Singer, J., Willett, J., 2003. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford University Press.
Snijders, T., 2005. Power and sample size in multilevel linear models. In: Everitt, B.S., Howell, D.C. (Eds.), Encyclopedia of Statistics in Behavioral Science, vol. 3. Wiley, pp. 1570–1573.
Snijders, T.A.B., Bosker, R.J., 1993. Standard errors and sample sizes for two-level research. J. Educ. Stat. 18 (3), 237–259. https://doi.org/10.3102/10769986018003237.
Snijders, T., Bosker, R., 1999. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage.
Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch. Intern. Med. 166 (10), 1092–1097.
Tibon, R., Levy, D.A., 2015. Striking a balance: analyzing unbalanced event-related potential data. Front. Psychol. 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.
Vasey, M.W., Thayer, J.F., 1987. The continuing problem of false positives in repeated measures ANOVA in psychophysiology: a multivariate solution. Psychophysiology 24 (4), 479–486. https://doi.org/10.1111/j.1469-8986.1987.tb00324.x.
Volpert-Esmond, Hannah I., Bartholow, B.D., 2019. Explicit categorization goals affect attention-related processing of race and gender during person construal. J. Exp. Soc. Psychol. 85, 103839 https://doi.org/10.1016/j.jesp.2019.103839.
Volpert-Esmond, H.I., Bartholow, B.D., 2020. A functional coupling of brain and behavior during social categorization of faces. Personal. Soc. Psychol. Bull.
Volpert-Esmond, H.I., Merkle, E.C., Bartholow, B.D., 2017. The iterative nature of person construal: Evidence from event-related potentials. Soc. Cogn. Affect. Neurosci. 12 (7), 1097–1107.
Volpert-Esmond, H.I., Merkle, E.C., Levsen, M.P., Ito, T.A., Bartholow, B.D., 2018. Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. Psychophysiology 55 (5), e13044. https://doi.org/10.1111/psyp.13044.
Von Gunten, C.D., Volpert-Esmond, H.I., Bartholow, B.D., 2018. Temporal dynamics of reactive cognitive control as revealed by event-related brain potentials. Psychophysiology. 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. Int. J. Methods Psychiatr. Res. 20 (3), e56–e68. https://doi.org/10.1002/mpr.348.
Weinberg, A., Olvet, D.M., Hajcak, G., 2010. Increased error-related brain activity in generalized anxiety disorder. Biol. Psychol. 85 (3), 472–480. https://doi.org/10.1016/j.biopsycho.2010.09.011.
Weinberg, A., Klein, D.N., Hajcak, G., 2012. Increased error-related brain activity distinguishes generalized anxiety disorder with and without comorbid major depressive disorder. J. Abnorm. Psychol. 121 (4), 885–896. https://doi.org/10.1037/a0028270.
Weinberg, A., Kotov, R., Proudfit, G.H., 2015. Neural indicators of error processing in generalized anxiety disorder, obsessive-compulsive disorder, and major depressive disorder. J. Abnorm. Psychol. 124 (1), 172–185. https://doi.org/10.1037/abn0000019.
Wilson, R.S., 1967. Analysis of autonomic reaction patterns. Psychophysiology 4 (2), 125–142. https://doi.org/10.1111/j.1469-8986.1967.tb02750.x.
Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111 (4), 931–959.