NeuroImage 234 (2021) 117932

Contents lists available at ScienceDirect

NeuroImage

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

Commentary

A commentary on establishing norms for error-related brain activity during the arrow flanker task among young adults

Peter E. Clayson[a,*], Emily S. Kappenman[b], William J. Gehring[c], Gregory A. Miller[d,e], Michael J. Larson[f,g]

[a] Department of Psychology, University of South Florida, Tampa, FL, USA
[b] Department of Psychology, San Diego State University, San Diego, CA, USA
[c] Department of Psychology, University of Michigan, Ann Arbor, MI, USA
[d] Department of Psychology, University of California Los Angeles, Los Angeles, CA, USA
[e] Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, Los Angeles, CA, USA
[f] Department of Psychology, Brigham Young University, Provo, UT, USA
[g] Neuroscience Center, Brigham Young University, Provo, UT, USA

ARTICLE INFO

Keywords:
Standardization
Normative database
Error-related negativity (ERN)
Error positivity (Pe)
Event-related potentials (ERPs)

ABSTRACT

We suggest that a large data set for the error-related negativity (ERN) and error positivity (Pe) components of the scalp-recorded event-related brain potential (ERP) recently published as normative is not ready for such use in research and, especially, clinical application. Such efforts are challenged by an incomplete understanding of the functional significance of between-person differences in amplitudes and of nuisance factors that contribute to amplitude differences, a lack of standardization of methods, and the use of a convenience sample for the potentially normative database. To move ERPs toward standardization and useful norms, we encourage more research on the meaning of differences in ERN scores, including factors that influence between- and within-person variation, and the dissemination of protocols for data collection and processing.

We appreciate the efforts of Imburgio et al. (2020) to establish normative data for the error-related negativity (ERN) and error positivity (Pe) components of the scalp-recorded event-related brain potential (ERP). The paper will be valuable for a number of reasons, including the encouragement of standardization of procedures and publication of additional norms. However, critical issues that it did not address raise important questions regarding the establishment and use of normative ERP data. We outline these issues and associated concerns below. Although for brevity we focus here on ERN, each point applies to Pe as well.

Research indicates that ERN involves multiple neural generators and neurotransmitters and is influenced by a combination of cognitive, affective, motivational, and motor processes (Gehring et al., 2012). As a result, variation in "true" ERN signal can be due to a range of factors. The causes of individual differences in ERN scores are often unclear, and such differences have little predictive utility in isolation. For example, both larger and smaller ERNs have been observed in the context of depression, and differences in either direction have been interpreted as clinically meaningful (Clayson et al., 2020; Moran et al., 2017). Higher cardiorespiratory fitness also appears to be related to both larger (Themanson et al., 2008) and smaller ERNs (Pontifex et al., 2011), yet each study interpreted these opposing ERN findings as indicating that better fitness related to improved performance monitoring. This interpretive inconsistency about the functional significance of ERN amplitudes (e.g., larger ERNs viewed as better due to "stronger" responses, and smaller ERNs viewed as "more efficient") is common across studies and is a barrier to establishing general norms, especially when there is also inconsistency in methods across studies. In other words, without knowing the functional significance of ERN amplitude in a specific context (population, task, etc.), identifying a given individual's ERN as larger or smaller than a comparison group provides little information about brain function.

Between-person differences in ERN amplitude can also occur due to factors other than "true" ERN signal. Specifically, the amplitude and morphology of an ERP component can vary across individuals due to nuisance variables that have nothing to do with cognitive processing,[1] including skull thickness, orientation of neural generators due to cortical

[1] A number of useful texts that cover important biophysics principles necessary for rigorous EEG research are readily available. Biophysics principles apply to many of the concerns raised in this commentary. Although not an exhaustive list, we recommend these primers: Jackson and Bolger (2014) and Kappenman and Luck (2012). We also recommend these in-depth texts:

* Corresponding author.
E-mail address: clayson@usf.edu (P.E. Clayson).

https://doi.org/10.1016/j.neuroimage.2021.117932
Received 7 July 2020; Received in revised form 4 September 2020; Accepted 27 February 2021
Available online 4 March 2021

folding, non-neural bioelectric signals, and changes in unmeasured participant state variables, such as attention and fatigue (Luck et al., 2011). Although the Imburgio et al. article attempts to address these factors with the use of error-minus-correct difference waves, these between-condition difference waves do not fully mitigate this problem. For example, a difference in skull thickness that causes the ERN to be twice[2] as large in one subject as some norm would likely also cause the correct-trial ERN (CRN) to be twice as large in that individual, and this increased amplitude would therefore still be present in an error-minus-correct difference wave. To eliminate the influence of such factors with difference waves, one needs to compare the same component in two experimental conditions (e.g., the ERN from compatible versus incompatible flanker trials), but this approach was not explored. Indeed, the influence of such factors was likely underestimated in the data set by their elimination of "outlier" participants from the creation of their norms—an approach that is not standard in ERP research and seems questionable when the goal is to create a normative database representative of standard ERPs from an unselected sample.

Another nuisance factor that results in problematic variance in ERN scores is measurement error, which is reflected in the widely variable estimates of internal consistency observed in a meta-analysis of 4499 participants from 68 samples nested within 43 studies (Clayson, 2020). Estimated coefficient alphas for eight ERN trials ranged from 0.02 to 0.94, with estimates partially moderated by type of paradigm, clinical status of the sample, approach for correcting ocular artifact, measurement sensors, and approach to calculating coefficient alpha. These data demonstrate the need for standardization and for consideration of contextual factors and nuisance variables that influence ERN scores.

Flanker tasks are among the most widely used for eliciting ERN, but the numerous variants of the task and numerous approaches to data processing limit its generalizability. Tasks vary widely on a number of potentially important characteristics, including number of trials, type of stimuli, stimulus luminance, length of inter-trial intervals, use of feedback, and task instruction. The data processing pipelines and quality assurance procedures used across labs are similarly variable. Imburgio et al. acknowledged the potential for many such factors to impact ERN scores, and they themselves used different lengths of the flanker task and different recording procedures across recruitment sites in the data they pooled. However, we see this lack of standardization as fatal to a potential normative database. As acknowledged by Imburgio et al., the published normative dataset represents just one instantiation of ERN processing. This necessarily limits its generalizability. Unknown is how applicable these norms are to other labs with different variants of the flanker task, data collection systems, data quality, or analysis pipelines. Indeed, even in the case of the Imburgio study, which kept many of these factors consistent, statistically significant results in ERN difference waves were observed across sites. Taken together, consequences for other researchers, peer reviewers, or clinicians who may rely on prematurely established norms could be substantial.

The lack of standardization of methods represents a significant barrier to individual-differences research. For example, the Research Domain Criteria (RDoC) initiative emphasizes examining the feasibility of neurophysiological measures of dimensional constructs with an eye toward clinical prediction (Kozak and Cuthbert, 2016). The ERN was initially investigated in healthy participants and was later used to study-group differences in clinical populations (Gehring et al., 2018). However, neurophysiological measures of group/condition differences do not easily translate to individual-differences research (Hajcak et al., 2017; Infantolino et al., 2018), and ERN research still has such obstacles to overcome.

As an example of a challenge in establishing norms, the mean ± standard deviation for ERN scores from 326 males in Imburgio et al. (Table 7) was +3.18 ± 6.50 μV, and the mean ERN score from 429 males (ERP Analysis section, Fig. S3) in Fischer et al. (2016) was −5.37 μV. These two studies had large samples with different demographic characteristics, used different variations of the flanker task, and varied in recording and data-reduction parameters. Each study employed high-quality methods and made reasonable decisions with regard to each characteristic. If the Imburgio et al. database were used to characterize the "average" male participant from the Fischer et al. sample, an ERN score of −5.37 would correspond to a *z* score of −1.32 (percentile rank = 9.34% or 90.66%). This could be interpreted as indicating that the average male in the Fischer et al. sample is abnormal, which is rather unlikely.

Numerous other issues arise when selecting a normative database, such as how representative the database is of the population(s) of interest (Mitrushina et al., 2005). To this end, sampling procedures for normative databases often stratify on age, sex, race/ethnicity, education level, and socioeconomic status. Imburgio et al. did not report using a standardized sampling procedure[3] and excluded participants with ERP scores greater than three standard deviations away from the mean, which truncates the distribution, leading to overestimation of deficits. Excluding outliers mischaracterizes the population and compromises the normative data. Unsystematic sampling procedures can yield unrepresentative cell sizes for each demographic characteristic, limiting generalizability.

More "ERPology" (Luck, 2014) is required to understand the functional significance of differences in ERN scores, including the diverse factors that influence between- and within-person variation. The Imburgio et al. data set is a valuable basis for that. The publication of protocols for ERN data processing is a necessary first step. Missing information about data processing appears to be a significant problem for ERP research broadly (Clayson et al., 2019; Keil et al., 2014), not just ERN research. Some labs have moved toward publishing supporting documentation that outlines all data recording and processing procedures (e.g., see Farrens et al., 2019). This practice serves to improve the replicability of processing pipelines, and such communication is crucial for standardization.

Opening up our lab notebooks by depositing ERN paradigms, scripts, etc. that are routinely used in-house via repositories will help to disseminate paradigms for optimization and standardization. The development of the ERP CORE (Compendium of Open Resources and Experiments) represents such an effort (https://erpinfo.org/erp-core; Kappenman et al., 2020). ERP CORE is a resource of open EEG paradigms, data, and processing scripts aimed at optimization and standardization of task and analysis procedures. After sufficient optimization and standardization, stratified samples can then be collected to build normative databases. In short, we appreciate the work of Imburgio et al. but believe that the characterization of values obtained for ERN and Pe in a single paradigm and analysis pipeline from a convenience sample

Nunez and Srinivasan (2006), Buzsaki, Anastassiou, and Koch (2012), and Zahn, Carpenter, and McGlashan (1981).

[2] Skull thickness has a multiplicative rather than additive impact on voltages measured at the scalp—illustrated by Ohm's law (*voltage* = *current* x *resistance*). Variance in skull thickness alters resistance (impedance), which will have a multiplicative impact on voltage measured at the scalp. This is especially relevant for difference scores in light of variability in skull thickness (and resistance) across people and across the lifespan (e.g., Frodl et al., 2001; Lillie, Urban, Lynch, Weaver, & Stitzel, 2016). Multiplicative differences in ERPs can also lead to mistaken statistical inferences in the analysis of interaction effects (McCarthy & Wood, 1985).

[3] Standardization samples comprise data that adhere to rigorous standards, including a standard procedure for recruiting participants. The recruited sample of participants should be appropriately stratified to reflect important demographic characteristics of the population of interest (see Mitrushina et al., 2005; Strauss, Sherman, & Spreen, 2006). In addition, tests should be administered and scored in a systematic and standardized fashion. Without proper standardized procedures, scores that are deviant from the normative sample could be due to any number of factors in the administration or scoring of the measures, and spurious interpretations can be made (Bigler & Dodrill, 1997).

as norms is premature for use in research and, especially, clinical application.

References

Bigler, E.D., Dodrill, C.B., 1997. Assessment of neuropsychological testing. Neurology 49, 1180–1182. doi:10.1212/WNL.49.4.1180-a.

Buzsaki, G., Anastassiou, C.A., Koch, C., 2012. The origin of extracellular fields and currents – EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. 13, 407–420. doi:10.1038/nrn3241.

Clayson, P.E., 2020. Moderators of the internal consistency of error-related negativity scores: a meta-analysis of internal consistency estimates. Psychophysiology 57, e13583. doi:10.1111/psyp.13583.

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 111, 5–17. doi:10.1111/psyp.13437.

Clayson, P.E., Carbine, K.A., Larson, M.J., 2020. Error-related negativity and reward positivity as biomarkers of depression: P-curving the evidence. Int. J. Psychophysiol. 150, 50–72. doi:10.1016/j.ijpsycho.2020.01.005.

Farrens, J. L., Simmons, A., Luck, S. J., & Kappenman, E. S. (2019). Electroencephalogram (EEG) Recording Protocol for Cognitive and Affective Human Neuroscience Research (Publication no. 10.21203/rs.2.18328/v2+). Retrieved Feb. 4, 2020

Fischer, A.G., Danielmeier, C., Villringer, A., Klein, T.A., Ullsperger, M., 2016. Gender influences on brain responses to errors and post-error adjustments. Sci. Rep. 6, 24435. doi:10.1038/srep24435.

Frodl, T., Meisenzahl, E.M., Müller, D., Leinsinger, G., Juckel, G., Hahn, K., Hegerl, U., 2001. The effect of the skull on event-related P300. Clin. Neurophysiol. 112. doi:10.1016/S1388-2457(01)00587-9.

Gehring, W.J., Goss, B., Coles, M.G.H., Meyer, D.E., Donchin, E, 2018. The error-related negativity. Perspect. Psychol. Sci. 13, 200–204. doi:10.1177/1745691617715310.

Gehring, W.J., Liu, Y., Orr, J.M., Carp, J., 2012. The error-related negativity (ERN/Ne). In: Luck, S.J., Kappenman, E.S. (Eds.), Oxford Handbook of Event-Related Potential Components. Oxford University Press, pp. 231–291.

Hajcak, G., Meyer, A., Kotov, R., 2017. Psychometrics and the neuroscience of individual differences: internal consistency limits between-subjects effects. J. Abnorm. Psychol. 126, 823–834. doi:10.1037/abn0000274.

Imburgio, M.J., Banica, I., Hill, K.E., Weinberg, A., Foti, D., Macnamara, A., 2020. Establishing norms for error-related brain activity during the arrow Flanker task among young adults. NeuroImage 213, 116694. doi:10.1016/j.neuroimage.2020.116694.

Infantolino, Z.P., Luking, K.R., Sauder, C.L., Curtin, J.J., Hajcak, G., 2018. Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. NeuroImage 173, 146–152. doi:10.1016/j.neuroimage.2018.02.024.

Jackson, A.F., Bolger, D.J., 2014. The neurophysiological bases of EEG and EEG measurement: a review for the rest of us. Psychophysiology 51, 1061–1071. doi:10.1111/psyp.12283.

Kappenman, E. S., Farrens, J., Zhang, W., Stewart, A. X., & Luck, S. J. (2020). ERP CORE: An Open Resource for Human Event-Related Potential Research. PsyArXiv. doi:10.31234/osf.io/4azqm

Kappenman, E.S., Luck, S.J, 2012. ERP components: the ups and downs of brainwave recordings. In: Luck, S.J., Kappenman, E.S. (Eds.), The Oxford Handbook of Event-Related Potential Components. Oxford University Press, Inc, New York, NY, pp. 3–30.

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, 1–21.

Kozak, M.J., Cuthbert, B.N., 2016. The NIMH research domain criteria initiative: background, issues, and pragmatics. Psychophysiology 53, 286–297. doi:10.1111/psyp.12518.

Lillie, E.M., Urban, J.E., Lynch, S.K., Weaver, A.A., Stitzel, J.D., 2016. Evaluation of skull cortical thickness changes with age and sex from computed tomography scans. J. Bone Miner. Res. 31, 299–307. doi:10.1002/jbmr.2613.

Luck, S.J., 2014. An Introduction to the Event-Related Potential Technique, 2nd ed. The MIT Press, Cambridge, MA.

Luck, S.J., Mathalon, D.H., O'Donnell, B.F., Hämäläinen, M.S, Spencer, K.M., Javitt, D.C., Uhlhaas, P.J, 2011. A roadmap for the development and validation of event-related potential biomarkers in schizophrenia research. Biol. Psychiatry 70, 28–34. doi:10.1016/j.biopsych.2010.09.021.

McCarthy, G., Wood, C.C., 1985. Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. Electroencephalogr. Clin. Neurophysiol. 62, 203–208. doi:10.1016/0168-5597(85)90015-2.

Mitrushina, M., Boone, K.B., Razani, J., D'Elia, L.F, 2005. Handbook of Normative Data for Neuropsychological Assessment. Oxford University Press.

Moran, T.P., Schroder, H.S., Kneip, C., Moser, J.S., 2017. Meta-analysis and psychophysiology: a tutorial using depression and action-monitoring event-related potentials. Int. J. Psychophysiol. 111, 17–32. doi:10.1016/j.ijpsycho.2016.07.001.

Nunez, P.L., Srinivasan, R., 2006. Electric Fields of the Brain: The Neurophysics of EEG, 2nd ed. Oxford University Press, Oxford; New York.

Pontifex, M.B., Raine, L.B., Johnson, C.R., Chaddock, L., Voss, M.W., Cohen, N.J., Hillman, C.H., 2011. Cardiorespiratory fitness and the flexible modulation of cognitive control in preadolescent children. J. Cognit. Neurosci. 23, 1332–1345. doi:10.1162/jocn.2010.21528.

Strauss, E., Sherman, E.M.S., Spreen, O, 2006. A Compendium of Neuropsychological Tests, 3rd ed. Oxford University Press, New York.

Themanson, J.R., Pontifex, M.B., Hillman, C.H., 2008. Fitness and action monitoring: evidence for improved cognitive flexibility in young adults. Neuroscience 157, 319–328. doi:10.1016/j.neuroscience.2008.09.014.

Zahn, T.P., Carpenter, W.T., McGlashan, T.H., 1981. Autonomic nervous system activity in acute schizophrenia: I. Method and comparison with normal controls. Arch. Gen. Psychiatry 38, 251–258.