A Bayesian Approach to Evaluating the Psychometric Properties of ERP Measurements

Peter E. Clayson

University of California, Los Angeles

INTRODUCTION

- To draw valid inferences from observed scores, data must first be demonstrated as reliable in the specific context where the data were collected (Smith & McCarthy, 1995).
  - Examples of context-relevant factors include the population being sampled or the setting being generalized to.
  - Failing to consider score reliability can lead to the mismeasurement of psychological constructs and, consequently, mistaken conclusions.
- Studies employing event-related brain potentials (ERPs) have additional contextual factors to consider that impact score reliability (Clayson & Miller, in press).
  - The processing pipeline needed to obtain ERP scores from continuous EEG requires a researcher to make various decisions (filtering, ocular artifact adjustment, statistical extraction approach, etc.) that may differ across ERP components, paradigms, hardware, and labs.
  - To draw solid conclusions about the psychology-biology relationships assessed by ERPs, the measurement approach used to quantify ERP components must first be demonstrated as reliable.
- Generalizability (G) theory provides some advantages over classical test theory for analyzing ERP score reliability (Baldwin et al., 2015; Clayson & Miller, in press, under review).
  - G theory provides a multifaceted approach for simultaneously estimating sources of measurement error, such as diagnostic category or numbers of trials needed for stable ERP measurements.
  - G theory can handle unbalanced designs. It is common for the number of trials retained for averaging to vary between participants.
  - G theory can handle unequal variances and covariances between parallel forms of measurement (e.g., split-half, test-retest, multiple tasks).
- The present study demonstrates the use of an open-source Matlab program, ERP Reliability Analysis (ERA) Toolbox, to evaluate ERP score dependability (a G-theory analog to internal consistency reliability) using generalizability theory.
- The purpose of the ERA toolbox is to characterize the reliability of ERP measurements to facilitate the calculation and reporting of these estimates.

METHOD

- The error-related negativity (ERN) data presented here represent a re-analysis of some of the data reported in Baldwin et al. (2015).
- EEG was recorded from 29 participants with an anxiety disorder (ANX), 319 healthy comparison subjects (Controls), and 34 participants with major depressive disorder (MDD) while completing a modified Eriksen flanker task.
- ERN amplitude was quantified as the average EEG activity from 0 to 100ms following the participant's erroneous responses across four leads: Fcz, Cz, and two leads just posterior and lateral to FCz
- Present reliability analyses examined the impact of the number of trials retained for averaging and diagnostic status on the dependability of ERN measurements.
- A level of .70 was considered the threshold for acceptable dependability coefficients.

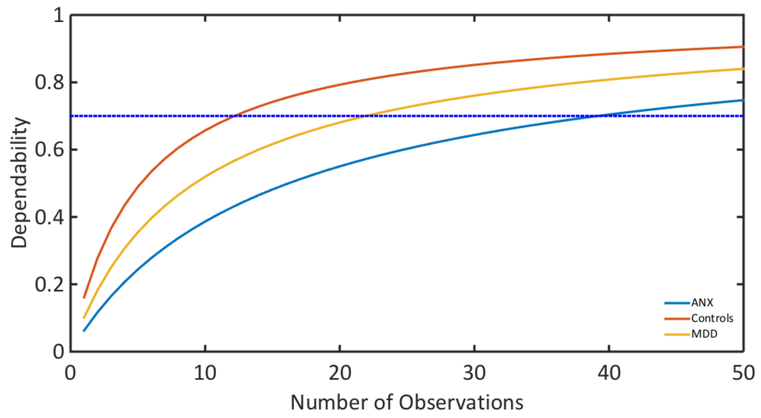
RESULTS

Relative Sizes of Sources of Variance

	Between-Person Standard Deviation	Within-Person Standard Deviation	Intraclass Correlation Coefficient
ANX	1.18 CI (0.75, 1.75)	4.65 CI (4.47, 4.84)	.06 CI (.03, .12)
Controls	2.18 CI (1.98, 2.39)	4.97 CI (4.91, 5.03)	.16 CI (.14, .19)
MDD	1.67 CI (1.23, 2.24)	5.01 CI (4.85, 5.17)	.10 CI (.06, .17)

- The point estimates and 95% credible intervals (CIs) for between- and within-person standard deviations and intraclass correlation coefficients (ICCs).
- Between-person standard deviations represent the expected distance between an idealized or average person's ERN score and the diagnostic group mean.
- Within-person standard deviations represent the variability in single-trial ERN scores and estimated measurement error.
- ICCs represent the proportion of between-person variance to total variance. When the ICC is high, between-person variance is large compared to error variance.
  - The ICC point estimate for the control group is higher than the ICC point estimates for the ANX and MDD groups, indicating that fewer trials will be needed in the control group for stable ERN scores.

Dependability Point Estimates

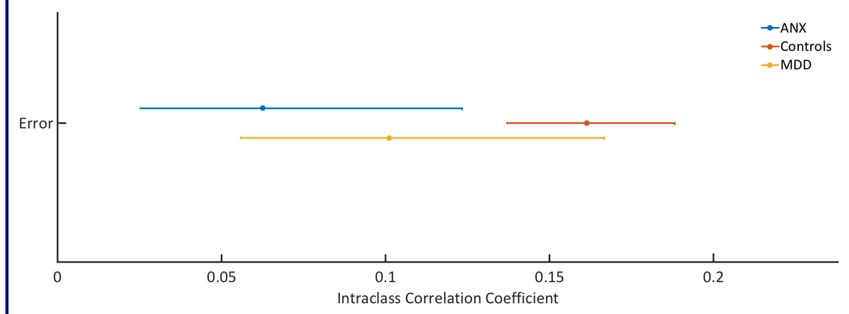


- The numbers of trials needed to obtain dependable ERN measurements (.70 criterion) was 40 for the ANX group, 13 for controls, and 22 for the MDD group.
- At each given number of trials included in an average, dependability estimates were highest for the healthy control group, followed by the MDD group and the ANX group.

	n Included	*n* Excluded	Dependability
ANX	15	14	.78 CI (.61, .90)
Controls	284	35	.90 CI (.88, .91)
MDD	31	3	.87 CI (.79, .93)

- Participants with too few trials to obtain an acceptable dependability threshold should be excluded: 14 participants in ANX group were rejected, 35 in controls, and 3 in MDD group.
- Data for each remaining participant had acceptable dependability point estimates, and those data can then be used for subsequent statistical analysis.
- The overall dependability point estimates for each group were also acceptable.

Intraclass Correlation Coefficients



- The ICCs and their 95% credible intervals for the ANX group, control group, and MDD group.

SUMMARY AND CONCLUSIONS

- In the present example, the ERA Toolbox was used to evaluate the reliability of ERN measurements three groups: an ANX group, control group, and MDD group.
- The toolbox estimated the contribution of the number of trials retained for averaging and diagnostic category to observed score variance.
  - ICCs were highest for the control group, which indicated that the control group would need the fewest trials for dependable ERN scores, followed by the MDD group, then the ANX group.
- The toolbox also provided information regarding the number of trials needed for dependable ERN measurements.
  - Each group required a different number of trials to obtain an acceptable reliability threshold. Ignoring diagnostic category in the estimation of score reliability would have resulted in unreliable estimates for the ANX and MDD groups.
- Participants with too few trials for each diagnostic category were identified and removed. An overall dependability estimate was then calculated for each diagnostic category.
- This demonstration of the toolbox also highlights how one contextual factor, diagnostic category, can impact of the dependability of ERN scores (see also Baldwin et al., 2015).
- The ERA Toolbox is a tool for examining the reliability of ERP scores on a study-by-study basis. It also facilitates the use of reliability thresholds for the exclusion of participant data with too few trials for stable measurements, which is an improvement on using the number of trials retained for averaging as a proxy for reliability.

References

Baldwin, S. A., Larson, M. J., & Clayson, P. E. (2015). The dependability of electrophysiological measurements of performance monitoring in a clinical sample: A generalizability and decision analysis of the ERN and Pe. *Psychophysiology, 52*, 790-800. doi: 10.1111/psyp.12401

Clayson, P. E., & Miller, G. A. (under review). ERP Reliability Analysis (ERA) Toolbox: An open-source toolbox for analyzing the reliability of event-related potentials.

Clayson, P. E., & Miller, G. A. (in press). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Pyschophysiology*. doi: 10.1016/j.ijpsycho.2016.09.005

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments, *Psychological Assessment,* 7, 300-308. doi: 10.1037/1014-3590.7.3.300

ERA Toolbox: http://peclayson.github.io/ERA_Toolbox/

Correspondence address: peter.clayson@gmail.com