ERP Reliability Analysis (ERA) Toolbox

created by

Peter E. Clayson

User Manual for Version 0.3.1

Last Modified 20 April 2016

https://github.com/peclayson/ERA_Toolbox/wiki

Contents

Introduction

Overview

The ERP Reliability Analysis (ERA) toolbox is an open-source Matlab program that uses generalizability (G) theory to evaluate the reliability of ERP data. The purpose of the toolbox is to characterize the dependability (G-theory analog of reliability) of ERP scores to facilitate their calculation on a study-by-study basis and increase the reporting of these estimates.

The ERA Toolbox provides information about the minimum number of trials needed for dependable ERP scores and describes the overall dependability of ERP measurements. All information provided by the ERA Toolbox is stratified by group and condition to allow the user to directly compare dependability (e.g., a particular group may require more trials to achieve an acceptable level of dependability than another group).

This document is a user manual for using the ERA Toolbox. For links to the most up-to-date version, bug reports, feature requests, and relevant publications, see the GitHub wiki.

Why another toolbox?

Reliability is a property of scores (the data in hand), not a property of measures. This means that P3, error-related negativity (ERN), late positive potential (LPP), (insert your favorite ERP component here) is not reliable in some "universal" sense. Since reliability is context dependent, demonstrating the reliability of LPP scores in undergraduates at UCLA does not mean LPP scores recorded from children in New York can be assumed to be reliable. Measurement reliability needs to be demonstrated on a population-by-population, study-by-study, component-by-component basis.

The purpose of the ERA Toolbox is to facilitate the calculation of dependability estimates to characterize observed ERP scores. ERP psychometric studies have been useful in suggesting cutoffs and characterizing the overall reliability of ERP components in those studies. When designing a study, that information can help guide decisions about, for example, the number of trials to present to a participant for a given population. However, just because the observed data meet the

previously recommended trial cutoff does not mean that the ERP scores are necessarily reliable. ERP score reliability cannot be inferred from trial counts, despite that trial counts and reliability are certainly related.

My hope is that the ERA Toolbox will make it easier to demonstrate the reliability of ERP scores on a study-by-study basis.

Mismeasurement of ERPs leads to misunderstood phenomena and mistaken conclusions. Poor ERP score reliability from mismeasurement compromises validity. Improving ERP measurement, by ensuring score reliability, can improve our trust of the inferences drawn from observed scores and the likelihood of our findings replicating.

For a detailed discussion of reliability as it relates to ERP scores, see Clayson and Miller (under review).

Getting Started

The ERA Toolbox can be downloaded here. The toolbox has been tested using Matlab versions 2014b and 2015b on both Mac OS X and Windows 7. Once you have downloaded the toolbox, add the directories to your Matlab path (use 'Add with subfolders…').

The toolbox has three dependents: MatlabStan, CmdStan, and MatlabProcessManager. The wiki containing instructions for downloading these can be found here. The current version of the toolbox has been tested using MatlabStan (downloaded 18 April 2016), CmdStan (v 2.0.1), and MatlabProcessManager (v 0.4.0).

Getting Help

If you run into any problems, check the wiki FAQ section and GitHub issues page for help. If there is not an answer to your question on the issues page, feel free to post a new issue. When posting an issue, please describe how you encountered the error and then copy and paste the Matlab error into the issue. Please also indicate the Matlab version that you are using and the version of the ERA Toolbox (which can be found at the startup of the toolbox).

I will do my best to respond as quickly as I am able. I hope that by using the GitHub issues page, that people will be able to find answers more quickly (if it's a

previously encountered problem) than emailing me directly. But, if you would prefer to email me, my email is peter.clayson@gmail.com.

The GitHub issues page is also a great place to recommend other features for the toolbox. I would love to hear any suggestions that you have!

Acknowledgements

Many people have helped me, whether it be directly or indirectly, with making this toolbox.

First, I have to thank Scott A. Baldwin for developing the dependability formulas for analyzing ERP scores (Baldwin, Larson, & Clayson, 2015). In addition to developing these formulas, he has been integral in the development of my own understanding of generalizability theory. He has also provided his own recommendations of what I should include in the toolbox.

I also need to thank Gregory A. Miller for encouraging me to develop the toolbox. We wrote a paper on the importance of assessing reliability on a study-by-study basis and provided guidelines for how reliability should be reported in ERP studies (Clayson & Miller, under review). As part of this paper, Dr. Miller recommended that we include code for assessing dependability estimates.

Lastly, I must thank Michael J. Larson for encouraging me to develop the toolbox, but especially for encouraging me to develop a GUI for the toolbox. I remember being at a conference and discussing the toolbox with Dr. Larson. He assumed it had so many features that I had not even thought about implementing at that point. I'm still working on implementing some of them! Dr. Larson also helped beta test the software for me. Actually, he is the one that introduced me to the importance of measurement.

Now that I have thanked Drs. Baldwin, Miller, and Larson, I need to say that although I have accepted a lot of feedback from them, I coded the actual toolbox. Thus, any bugs or errors are my responsibility, and I take full credit for them.

Citations

When using the ERA toolbox, please cite (Baldwin et al., 2015) for the formulas using generalizability theory and CmdStan (Stan Development Team, 2016) and Stan for the actual crunching (Carpenter et al., in press). I think that it is important

to cite both CmdStan and Stan, as 1) the toolbox would not do anything without them and 2) citing the paper and software help the Stan Development Team get more funding, which means more features for us to use! I have also submitted a poster for the Society for Psychophysiological Research conference later this year. The poster can be cited when mentioning the actual ERA toolbox (Clayson, under review).

Example sentence to include in paper:

Dependability estimates were calculated following the formulas provided by Baldwin et al. (2015) using the ERP Reliability Analysis (ERA) Toolbox v 0.3.1 (Clayson, under review). The ERA Toolbox used CmdStan v 2.0.1 (Stan Development Team, 2016) to implement the statistical models in Stan (Carpenter et al., in press).

Again, I hope that by citing the Stan Development team, they can get more funding to give us cool features!

License

Analyzing Data

Preparing Data

The ERA Toolbox works on single-subject, single-trial ERP measurements. It does not matter what kind of measurement is used (e.g., latency v amplitude). Before using the toolbox, create a spreadsheet with all of the single-subject, single-trial measurements.

Use a long format for all inputted data. Each line should correspond to one trial, and there can be multiple measurements per line.

At the very least, each line should have an ID variable and a measurement variable. When processing the data in the toolbox, you will be asked to specify which variable is which. Additionally, you can have columns that indicate group membership and event type, although they are not required columns.

	A	B	C	D
1	ID	Error_Label	ERN_ROI	Group
2	AD_6002	Correct	2.534287	MDD
3	AD_6002	Correct	-3.93958	MDD
4	AD_6002	Correct	1.378525	MDD
5	AD_6002	Correct	-0.55206	MDD
6	AD_6002	Correct	2.23271	MDD
7	AD_6002	Correct	1.482636	MDD
8	AD_6002	Correct	-0.27898	MDD
9	AD_6002	Correct	1.105055	MDD
10	AD_6002	Correct	-0.34711	MDD
11	AD_6002	Correct	-2.51458	MDD

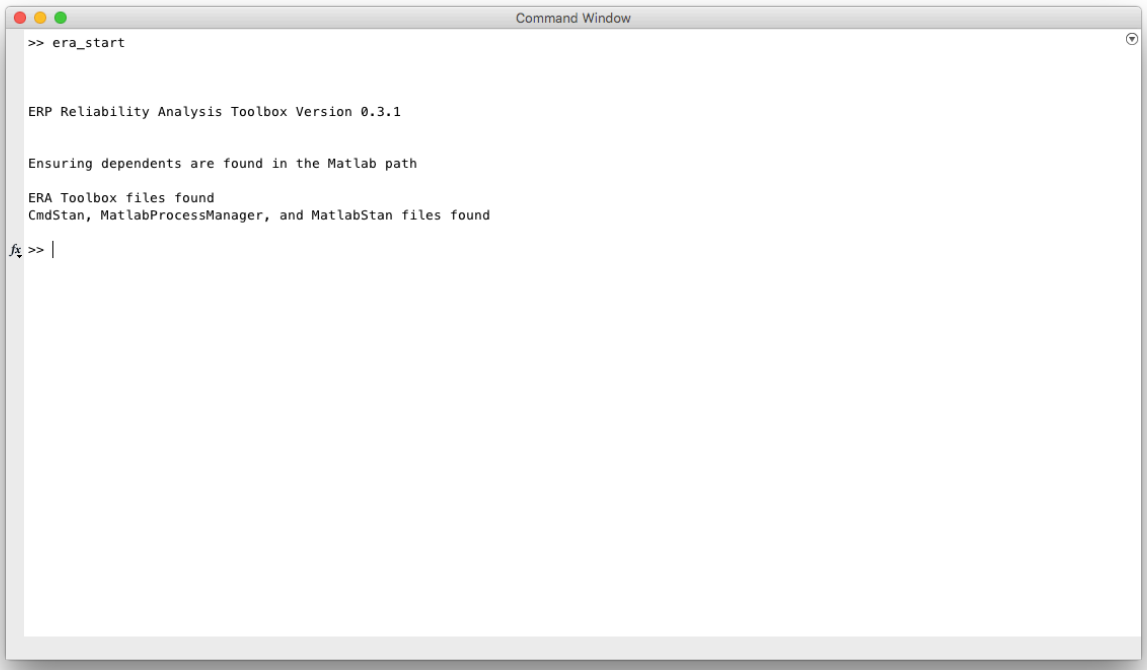
This is an example of how your data could be set up. There are columns for the subject ID (ID), group membership (Group), event type (Error_Label), and ERN amplitude measurement (ERN_ROI).

The following file formats are supported by the toolbox.

- .xlsx – Excel File
- .xls – Excel File 97-2003
- .csv – comma-separated value file
- .dat – tab-delimited text
- .ods – OpenDocument spreadsheet

Start the toolbox

In order to start the ERA Toolbox, simply type era_start in the Matlab command window and click Enter.



After clicking Enter, you will see some information printed in the Command Window.

The first thing printed is the name of the toolbox and version number. (This is where you can get the version number for reporting bugs/issues.)

The toolbox will then check to ensure that the dependents are located in your Matlab path. It will notify you whether they are found. If the ERA Toolbox is not found, make sure that you added the ERA_Toolbox directory using 'Add with subfolders…'. If the toolbox could not locate CmdStan, MatlabProcessManager, or MatlabStan, then you need to follow the instructions here for downloading and installing them.

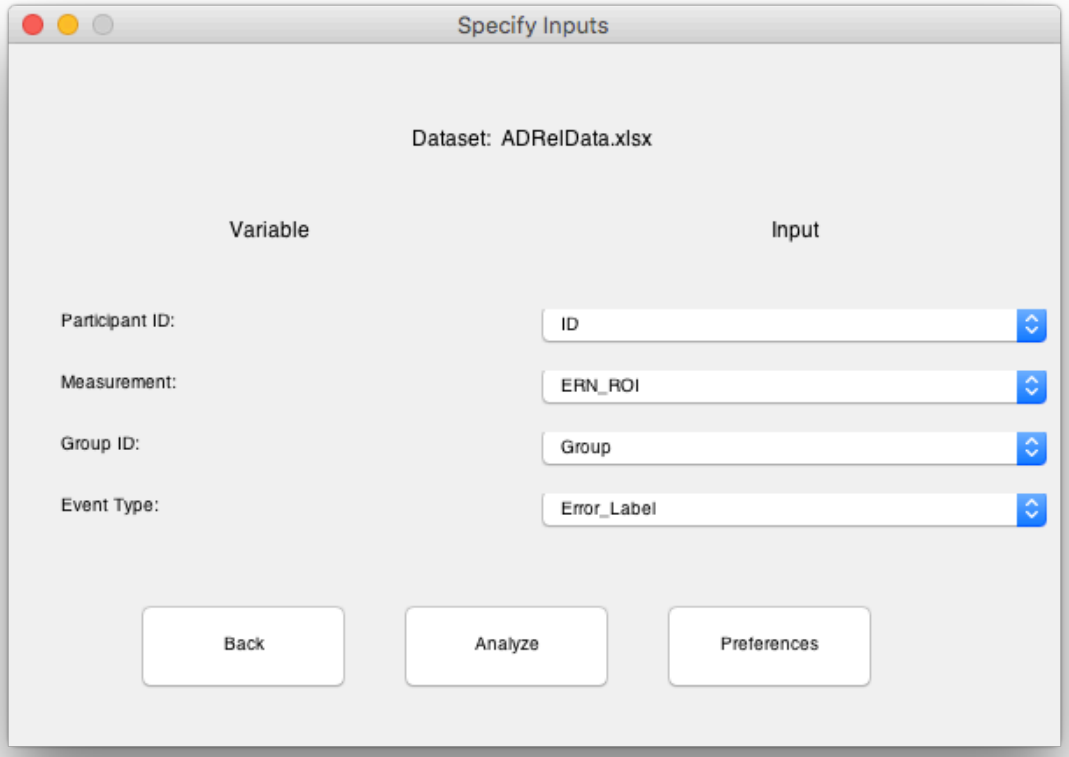
If everything goes smoothly you will be presented with a gui that requires you to indicated whether you would like to Process New Data or View Results.

Process New Data: If data have not yet been analyzed by the toolbox, select this option to begin analyzing the data in CmdStan.

View Results: If the data have already been processed in the toolbox, select this option to summarize the data.

Process New Data

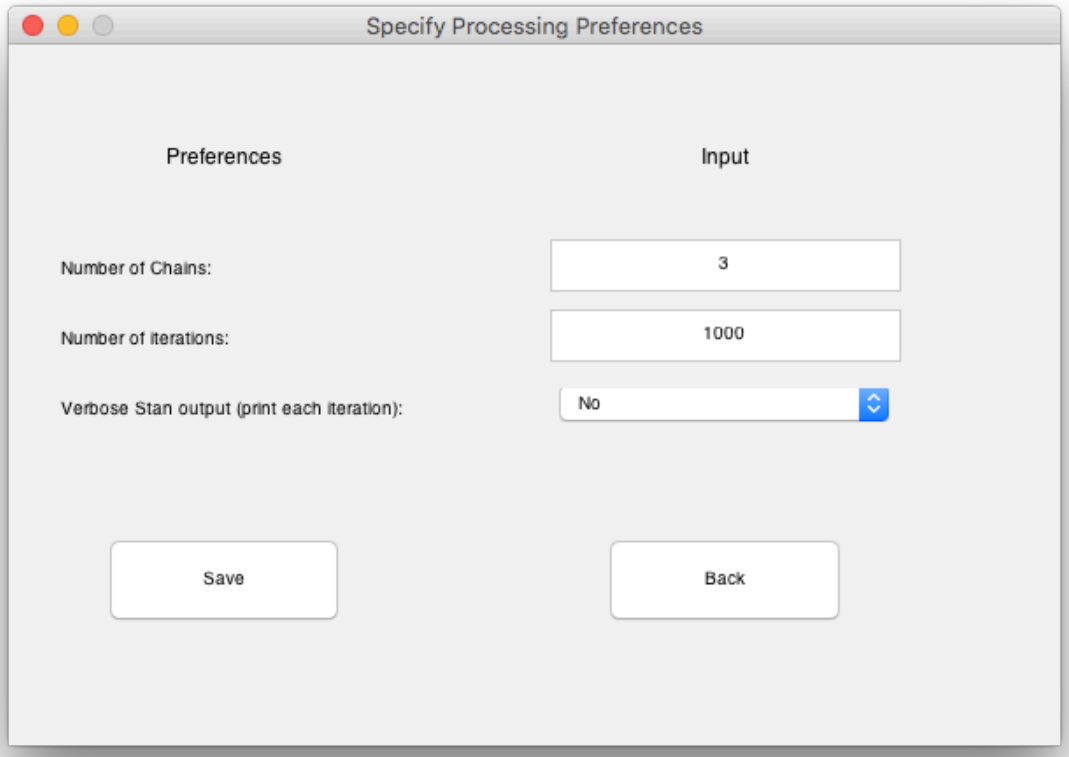
Selecting the Process New Data button will bring up a window to choose the file that is to be processed. After selecting a file, a gui will pop up allowing you to specify which columns correspond to which variables.



The gui will display the filename of the loaded dataset. This screen will allow you to select which columns from the spreadsheet belong to which variable: Participant ID, Measurement, Group ID, and Event Type. The columns in your data will be populated in the pulldown lists on the right.

At the very least, the Participant ID and Measurement rows must be populated with column names (notice ‘none’ is not an option). If there are not columns indicating group membership or event type in the data, then select ‘none’ from the pulldown list.

Once you have selected which columns contain the data to be analyzed, select the Preferences button.



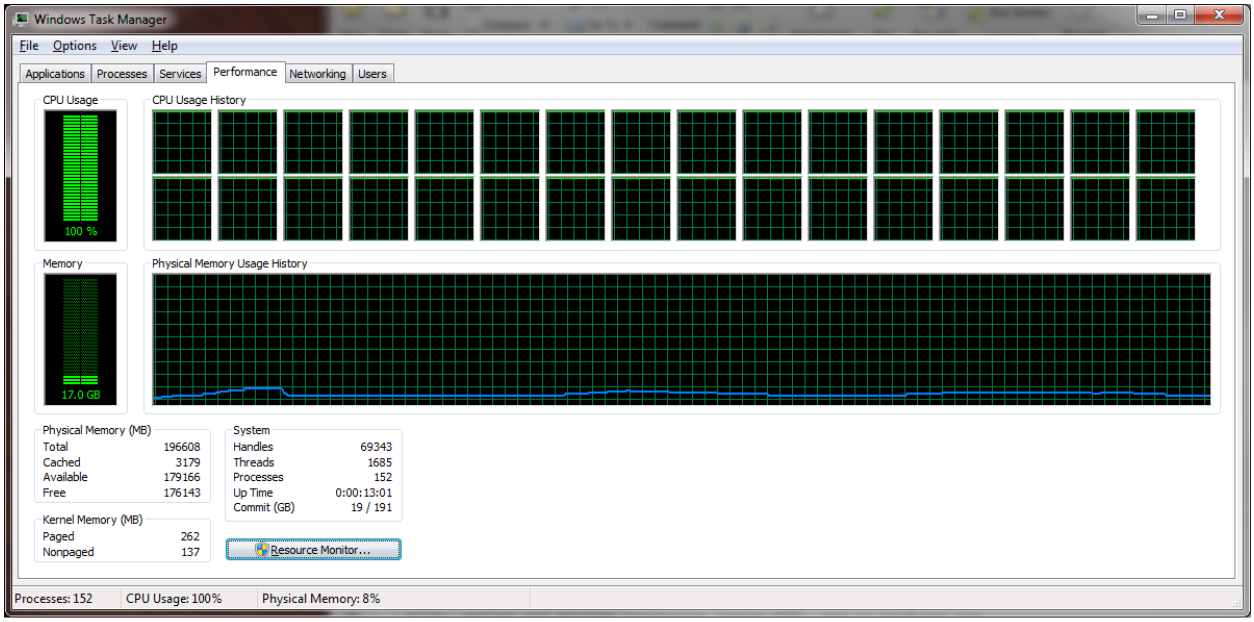
There are three preferences to choose from for processing.

- Number of Chains: Specifies the number of chains to run for the Markov Chain Monte Carlo estimation procedure in CmdStan. You need at least 3 chains in order to properly assess convergence (the toolbox will not allow you to choose fewer than 3). You can do more if you would like. CmdStan will run the chains in parallel.

  Holding the number of iterations constant, increasing the number of chains will increase the number of draws that go into your estimates. Thus, increasing the number of chains can increase the precision of your estimates.

  CmdStan will run each chain in parallel on your computer. If you have a computer with 8 cores (and the RAM to support the computation), you can use more than 3 chains (up to 8 in parallel). CmdStan will not let you run more chains than you have cores in parallel. If you specify more chains than you have cores, CmdStan will run as many as it can in parallel, then execute the other

chains when a core frees up. This is a quick way to push your computer to its limit. I rather enjoy it. As you can see from this beautiful screenshot, not exactly RAM heavy (~17GB for 32 models), but I use all of my processors… The RAM used depends heavily on the number of participants and trials in the dataset.



- Number of Iterations: Specifies the number of iterations for each chain. There is not a good way to estimate a priori the number of iterations that you need to get estimates that converge. If your model is having trouble converging, then you need to increase the number of iterations. The toolkit's default is 1,000. That is a pretty arbitrary number. I frequently use 10,000, which gives me good convergence for the ERP data I've run through the toolbox (but is usually overkill). Use as many iterations you need until your model converges.

- Verbose Stan Output: If no, nothing will be printed to the command window while CmdStan is crunching. If yes, the progress of CmdStan crunching will be printed in the command window. The model being run will be printed to the command window and the progress of the crunching will be shown.

To save the preferences click Save. If you do not want to save the changes, just click Back.

Once you have set the preferences you want, click Analyze to run the model.

Analyze

After clicking Analyze, the model will be executed. CmdStan will translate the Stan program to C++ and compile the resulting C++ to an executable. The model may take a while to set up.

CmdStan will take a while to crunch depending on the size of the dataset and the number of iterations (and your computing power).

A few things to note.

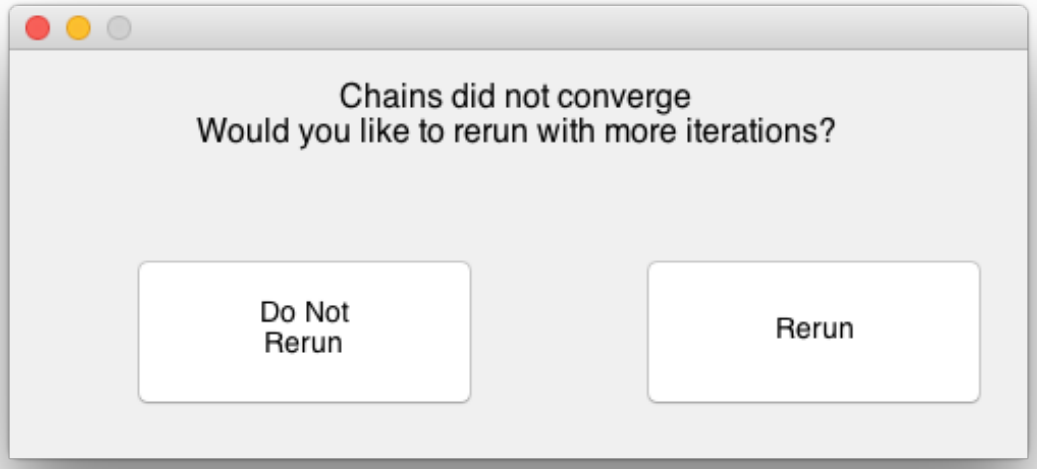
The models run in C++, not Matlab. Thus, terminating Matlab's processing (using ctrl+c) does not terminate the C++ models. If for some reason you need to terminate the processing, you need to use the task manager to terminate the cmdstan processes (there will be as many separate processes running as chains that you specified). You'll also need to use ctrl+c to terminate Matlab's processing. The Matlab command window will not take inputs while CmdStan is running the chains.

Convergence of the chains will automatically be checked after the model has finished processing. Convergence is assessed in two ways (see Gelman et al., 2013, pp. 281-286).

1. The potential scale reduction for the scalar estimands ($\hat{R}$) is checked to see whether it is lower than a threshold set at 1.1.

2. The effective sample size for each scalar estimand is checked to ensure it is greater than 10 times the number of chains.

If the convergence criteria are not satisfied, then a gui will pop up asking whether the model should be rerun with more iterations. If convergence criteria are satisfied, the user will be taken to the screen to setup which tables and plots to show for processed data.



If Rerun is selected, then the toolbox (somewhat arbitrarily) will double the number of iterations are rerun the model. If the number of iterations after being doubled is not greater than 1,000, the number of iterations will be set to 1,000 for the next round of processing.

If Do Not Rerun is selected, the user will be taken back to the gui for specifying the inputs of the model.

It is difficult to know how many iterations are necessary to satisfy the convergence criteria. If you have the time and computing power, it may be easier to select Do Not Rerun and greatly increase the number of iterations (increase to 10,000?). It can be troublesome to keep rerunning the model over and over again with only doubling the number of iterations each time.

Viewing Data

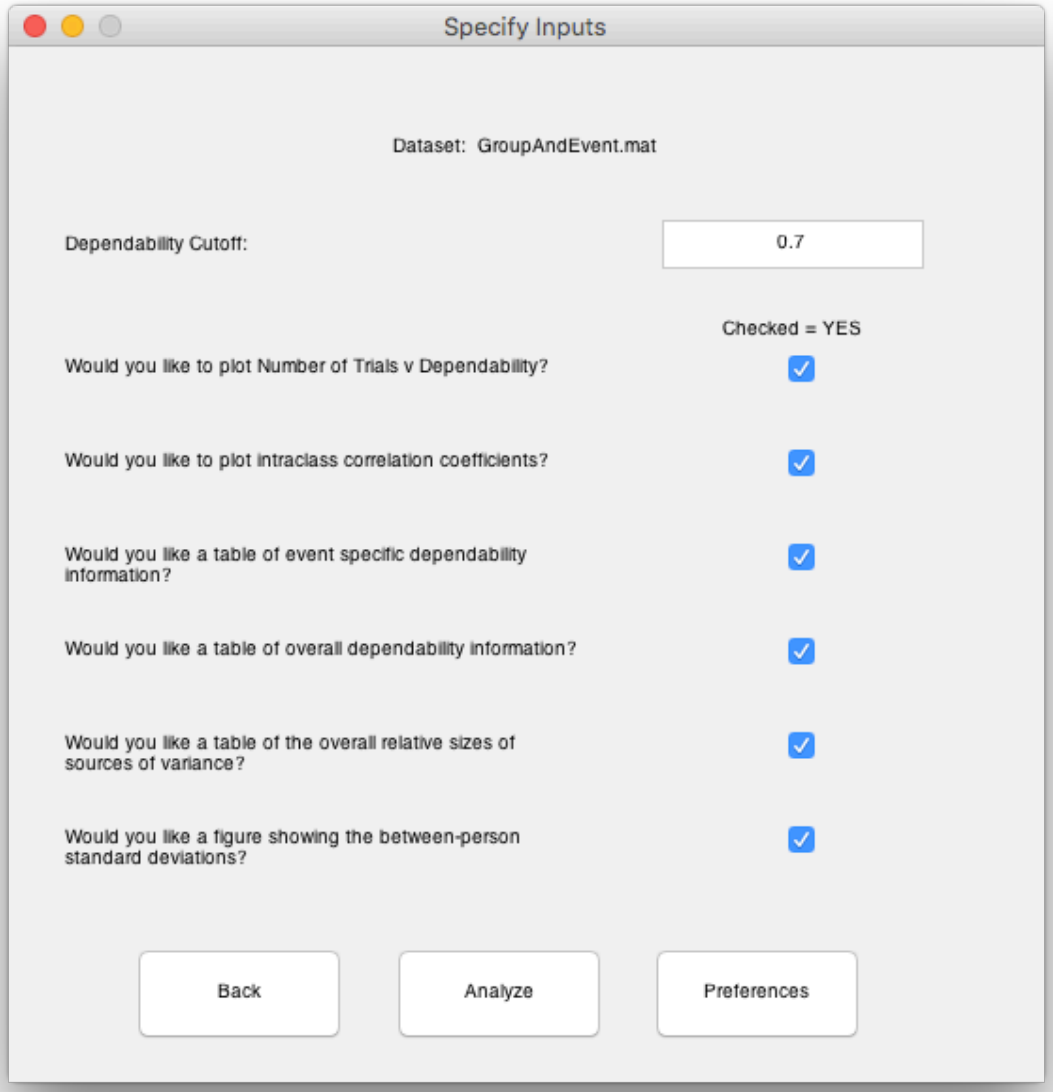
There are two ways to get to the gui for specifying the inputs to view the data.

If the data have just been processed, the gui for specifying the inputs will automatically pop up once the convergence criteria have been met.

Alternatively, you can execute era_start in the Matlab command window and select View Results.

If View Results is selected, a window will pop up for selecting a data file.

Specify Inputs



There are many options for viewing the processed data. If the box is checked, that indicates that you would like to view that table or plot. If applicable, data will be stratified by group, event, or group and event. More information about each plot is provided in its own section below.

- Dependability Cutoff: Specify the threshold for acceptable dependability. The default threshold is .70, although a threshold of .80 is recommended. This is the cutoff that the toolbox will use to search for the number of trials that are needed for dependable estimates. For guidelines and

recommendations, as well as the deleterious effects that reliability can have on statistics, see Clayson and Miller (under review).

- Would you like to plot Number of Trials v Dependability?: This produces a plot that shows the relationship between the number of trials included in the estimate and the dependability estimate.

- Would you like to plot intraclass correlation coefficients?: This will display a plot that visualizes the intraclass correlation coefficients (ICCs).

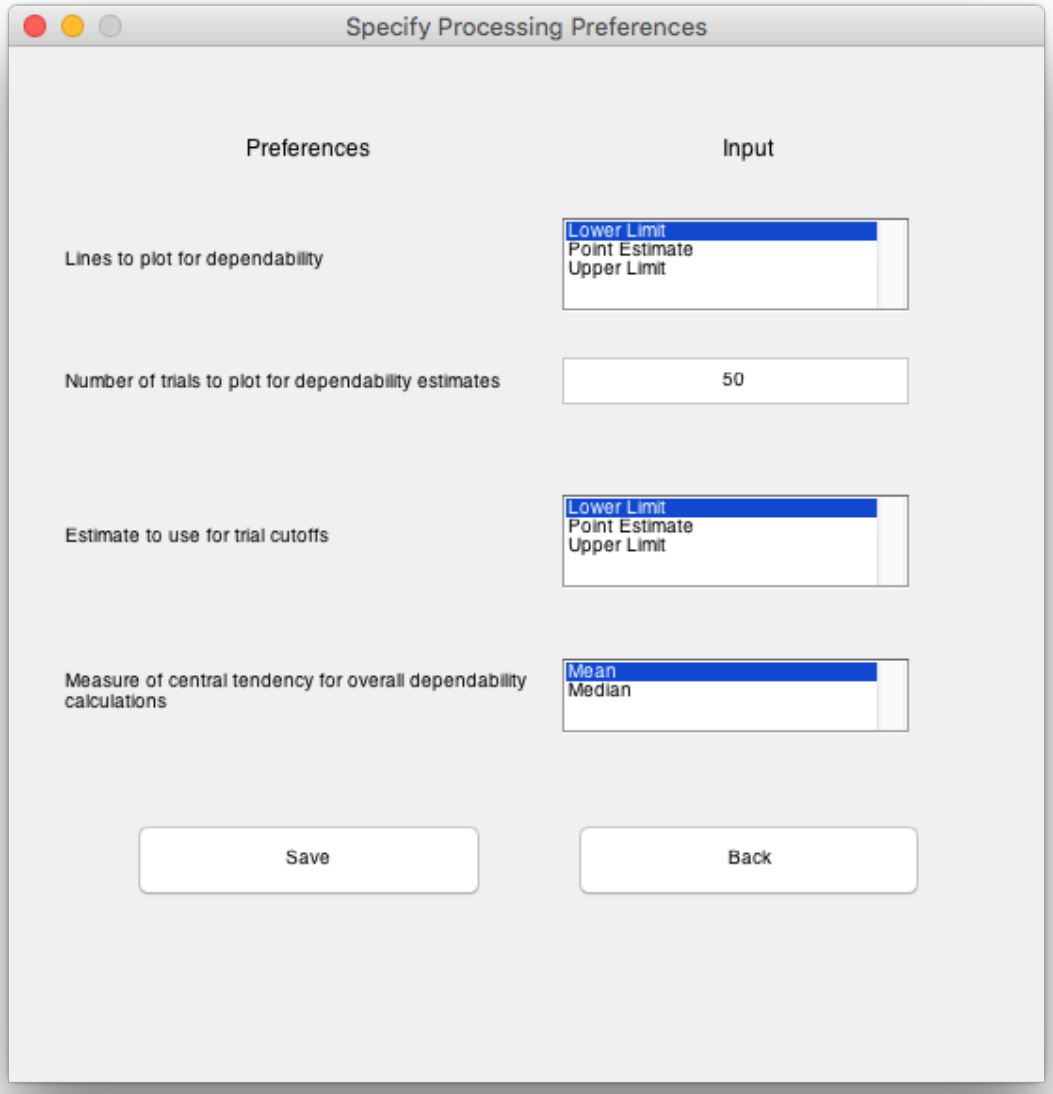
- Would you like a table of event specific dependability information?: This will show a table with the number of trials needed to achieve the specified dependability threshold and the associated dependability point estimate with a 95% credible interval (aka, fiducial trust limits; these are kind of similar to confidence intervals).

- Would you like a table of overall dependability information?: This table contains most of the summary information for the data. It includes the number of participants who met the threshold for each given event. It also indicates the number of participants who did not meet the threshold. It includes the overall dependability of the data and the ICCs. Summary information (mean, minimum, and maximum) for the number of trials for participants with "good" data (those participants with data that met the dependability threshold) is also shown.

- Would you like a table of overall relative sizes of sources of variance?: This will produce a plot of the ICCs.

- Would you like a figure showing the between-person standard deviations?: Provides a plot show the between-person standard deviations.

Selecting Preferences will bring up various options for viewing the data.



- Lines to plot for dependability: Specifies the line to be plotted for the Number of Trials v Dependability plot. Options include the dependability point estimate and the lower and upper limits of the 95% credible interval for the dependability estimate.

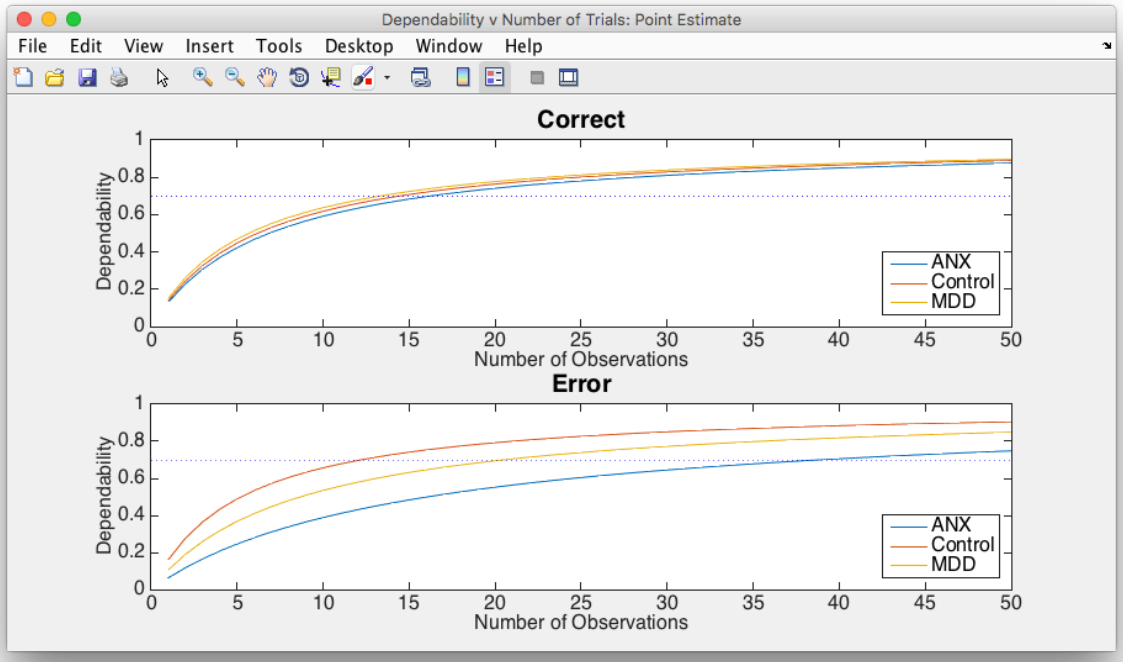
- Number of trials to plot for dependability estimates: Indicate the number of trials (x-axis) that should be plotted for the Number of Trials v Dependability plot.

- Estimate to use for trial cutoffs: Specifies which estimates to use for finding the number of trials to achieve acceptable dependability. Options include the dependability point estimate and the lower and upper limits of the 95% credible interval for the dependability estimate.

- Measure of central tendency to use for overall dependability calculations: Specify whether to use the mean or median number of trials in the calculation of the overall dependability estimates.

Viewing Outputs

Dependability Plot

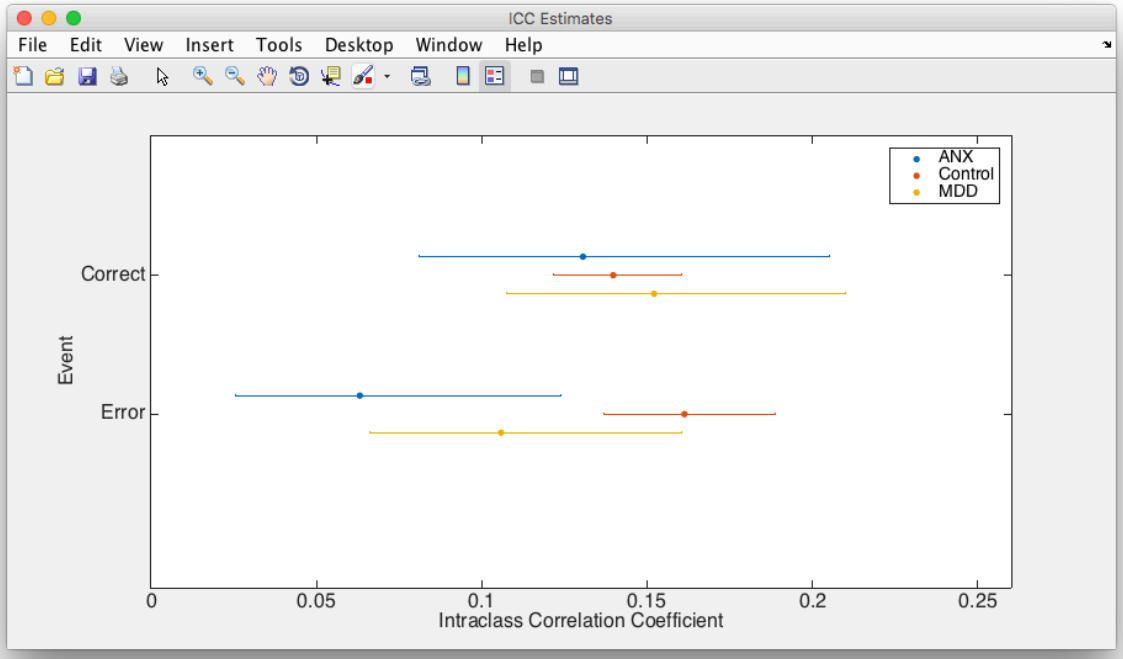


This plot shows the impact of increasing the number of trials on dependability estimates. The y-axis is the dependability estimate and the x-axis is the number of observations (trials) included in the estimate (upper limit of x-axis is defined in preferences). Each event is plotted separately, and separate lines show information for separate groups (legend provided). The dotted horizontal line is the specified dependability cutoff. This particular plot shows the point estimate, which is

indicated in the menu bar of the window. The plot can be saved by pressing File, then Save.

In the example given, groups show a similar relationship between numbers of trials and dependability for correct trials. The relationship is very different for error trials. Controls obtain an acceptable level of dependability more quickly than the other groups.

ICC Plot

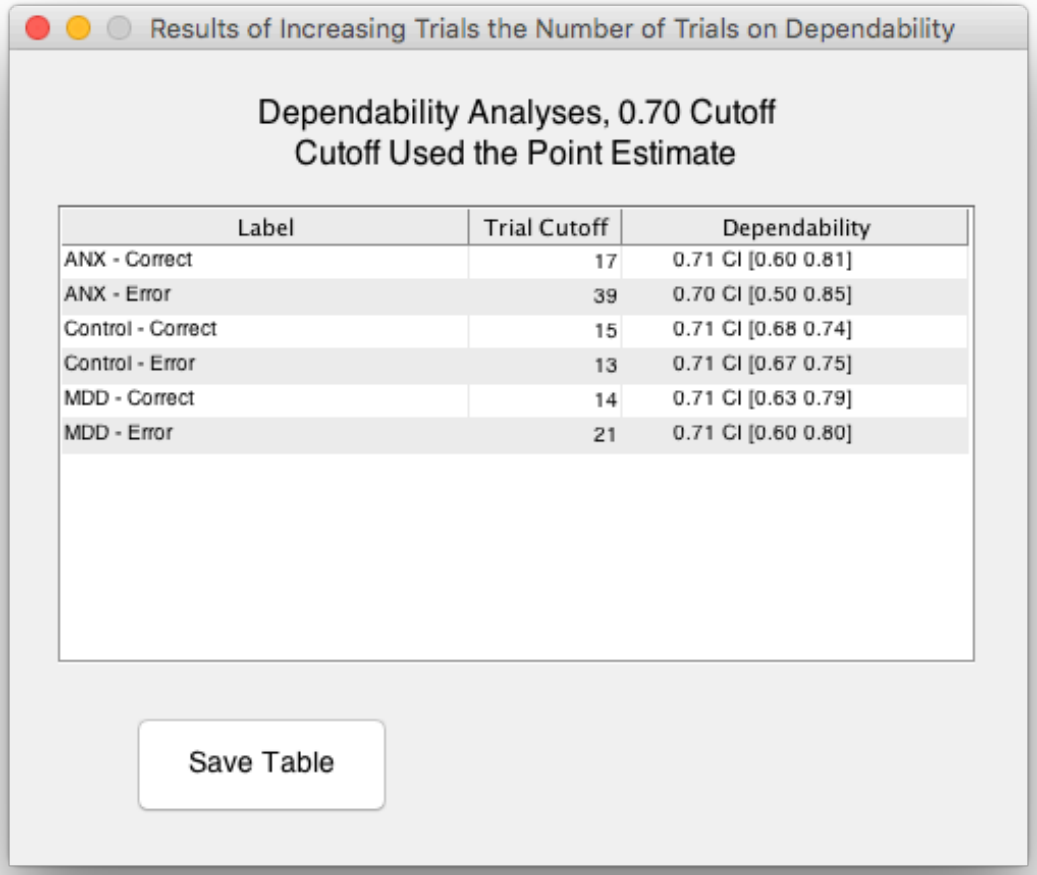


This plot shows the point estimate of the intraclass correlation coefficients (ICCs) with their associated 95% credible intervals. The y-axis is separate events, and the x-axis is ICC. Separate lines are provided for each group. The plot can be saved by pressing File, then Save.

The ICC is calculated using the formula: between-person variance/(between-person variance + error variance). This can be interpreted as the proportion of the total variance that is between persons. When ICCs are below .3, it indicates that the majority of the observed variance is within persons and suggests that many trials will likely be needed for stable measurements.

In the example given, groups show fairly similar ICCs for correct trials. For error trials, ICCs are different between the groups. The ANX group is much lower than the Control group and suggests that the ANX group is going to need more error trials than the Control group to achieve dependable measurements.

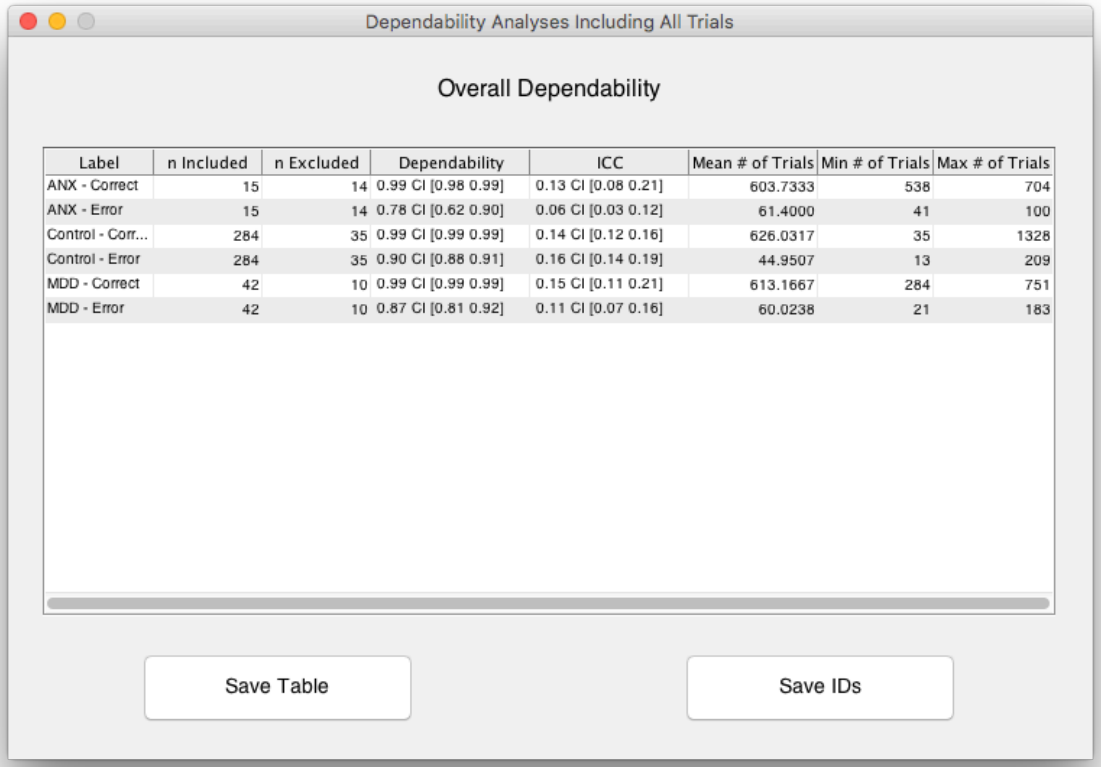
Cutoff Inclusion Table



This table shows number of trials needed to achieve an acceptable level of dependability, as specified by the user (indicated in the title of the table). Every possible group and event combination is shown with its specified cutoff and dependability point estimate and 95% credible interval. There is a button to save this table to a file.

Based on this table the user can get a sense of how many trials are needed for the data in hand. These cutoffs could then be used to guide decisions for excluding participants with unreliable measurements (Clayson & Miller, under review).

Overall Dependability Table



This table summarizes the data set. This gui has two buttons. The Save Table button will save all of this information to a file. The Save IDs button will save a file that includes a column for the participant IDs with good data (those in n Included column) and a column for the participant IDs with bad data (those in n Excluded column). The information in this file could then be used to remove those participants with unreliable measurements from statistical analyses.

- n Included: number of participants with enough trials to survive the dependability threshold. Participants will only be included if they have enough trials for all possible events.

- n Excluded: number of participants that did not have enough trials to survive the dependability threshold for at least one event.

- Dependability: dependability point estimate with 95% credible interval. This is the estimate for the dependability of the overall data. Either the mean or median (specified in the viewing preferences) can be used in this dependability equation. If the data are extremely skewed, you may consider using the median over the mean.

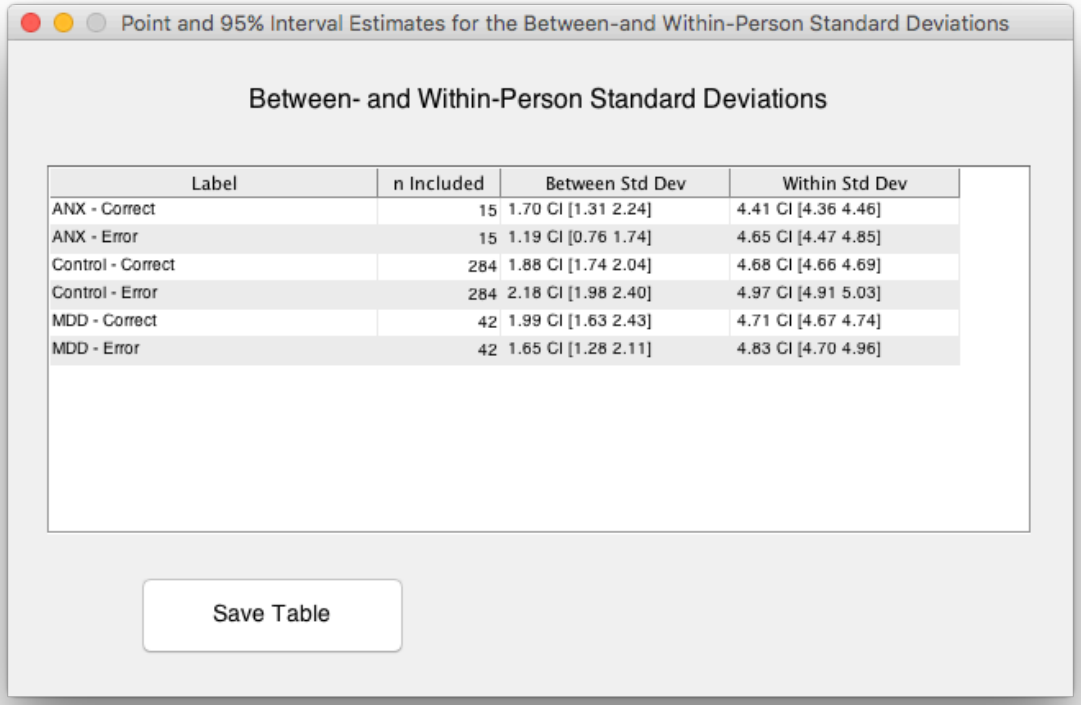
- ICC: ICC point estimates with 95% credible interval.

- Mean # of Trials: Mean number of trials for those participants with enough trials to survive the dependability threshold for each event (those participants in the n Included column).

- Min # of Trials: Minimum number of trials for those participants with enough trials to survive the dependability threshold for each event (those participants in the n Included column).

- Max # of Trials: Maximum number of trials for those participants with enough trials to survive the dependability threshold for each event (those participants in the n Included column).

Relative Sizes of Variance Table



This table shows the between- and within-person standard deviations. This plot is another way to look at the relative contributions of variance (compared to ICCs). The table can be saved to a file by pressing the Save Table button.

- n Included: number of participants with enough trials to survive the dependability threshold. Participants will only be included if they have enough trials for all possible events. This is information is just for reference and is not factored into the standard deviation analyses.

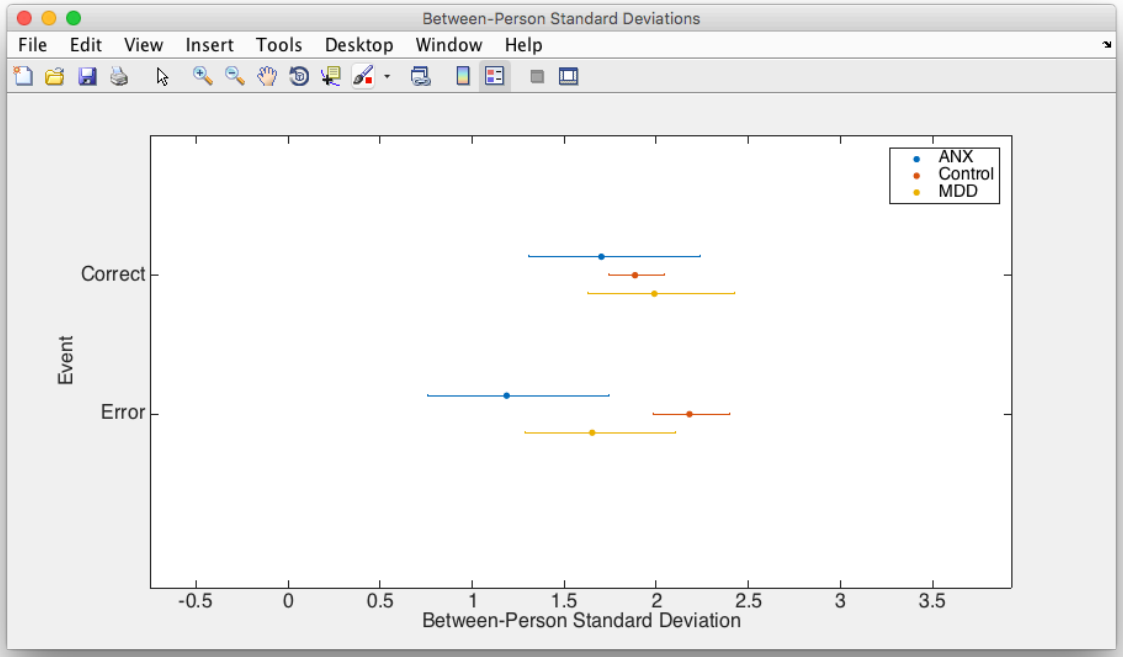
- Between Std Dev: Point estimate for between-person standard deviations and the 95% credible interval.

- Within Std Dev: Point estimate for within-person standard deviations and the 95% credible interval.

If within-person standard deviations are large compared to between-person standard deviations, many trials are likely to be needed for stable measurements.

Comparing the standard deviations between groups and events also gives a sense of the underlying data contributing to these measurements. For example, the point estimate for Control-Error for between-person standard deviations is higher than the point estimates for the same condition (Error) for ANX and MDD groups. Given this, it makes sense as to why the Control group needs fewer trials than the ANX and MDD groups to obtain stable measurements.

Between-Person Standard Deviation Plot



This plot shows the point estimate of the between-person standard deviations with their associated 95% credible intervals. The y-axis is separate events, and the x-axis is between-person standard deviation. Separate lines are provided for separate groups. The plot can be saved by pressing File, then Save.

See section above for how to use between-person standard deviations.

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

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (in press). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Clayson, P. E. (under review). *A bayesian approach to evaluating the psychometric properties of ERP measurements: The ERP Reliability Analysis Toolbox*. Paper presented at the 56th Annual Meeting for the Society for Psychophysiological Research.

Clayson, P. E., & Miller, G. A. (under review). Psychometric considerations in the measurement of event-related brain potentials: Guideliens for measurement and reporting.

Gelman, A., Carlin, J. B., Stern, H. E., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Pres.

Stan Development Team. (2016). CmdStan: The command-line interface to Stan, version 2.9.0. from http://mc-stan.org/