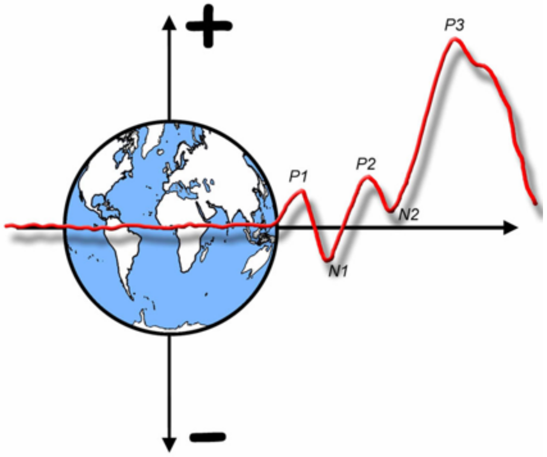


CENTER FOR MIND AND BRAIN
UNIVERSITY OF CALIFORNIA, DAVIS

This presentation was made possible by NIH grant R25MH080794

Standardized Measurement Error

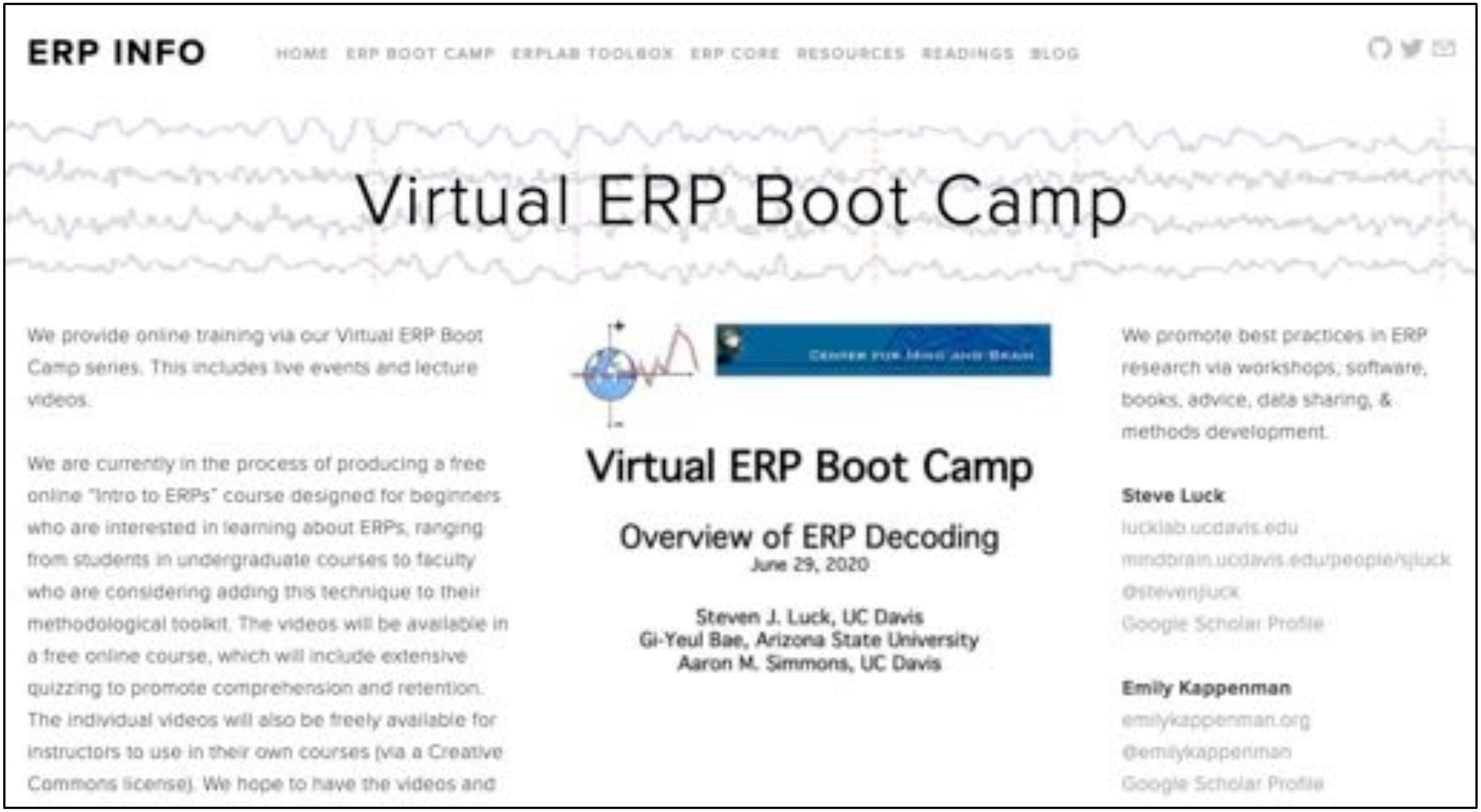
A Universal Measure of Data Quality for Averaged ERPs

Steve Luck, Andrew Stewart, & Aaron Simmons
University of California, Davis

If you use the Standardized Measurement Error in a publication, please cite: Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2020). Standardized Measurement Error: A Universal Measure of Data Quality for Averaged Event-Related Potentials. PsyArXiv. https://doi.org/10.31234/osf.io/dwm64 (or the peer-reviewed version once it is published).

Recording

- A recording will be available for 1 week
- Check https://erpinfo.org/virtual-boot-camp
  - The link should be available by the end of the day tomorrow



Virtual ERP Boot Camp

We provide online training via our Virtual ERP Boot Camp series. This includes live events and lecture videos.

We are currently in the process of producing a free online "Intro to ERPs" course designed for beginners who are interested in learning about ERPs, ranging from students in undergraduate courses to faculty who are considering adding this technique to their methodological toolkit. The videos will be available in a free online course, which will include extensive quizzing to promote comprehension and retention. The individual videos will also be freely available for instructors to use in their own courses (via a Creative Commons license). We hope to have the videos and

Virtual ERP Boot Camp

Overview of ERP Decoding
June 29, 2020

Steven J. Luck, UC Davis
Gi-Yeul Bae, Arizona State University
Aaron M. Simmons, UC Davis

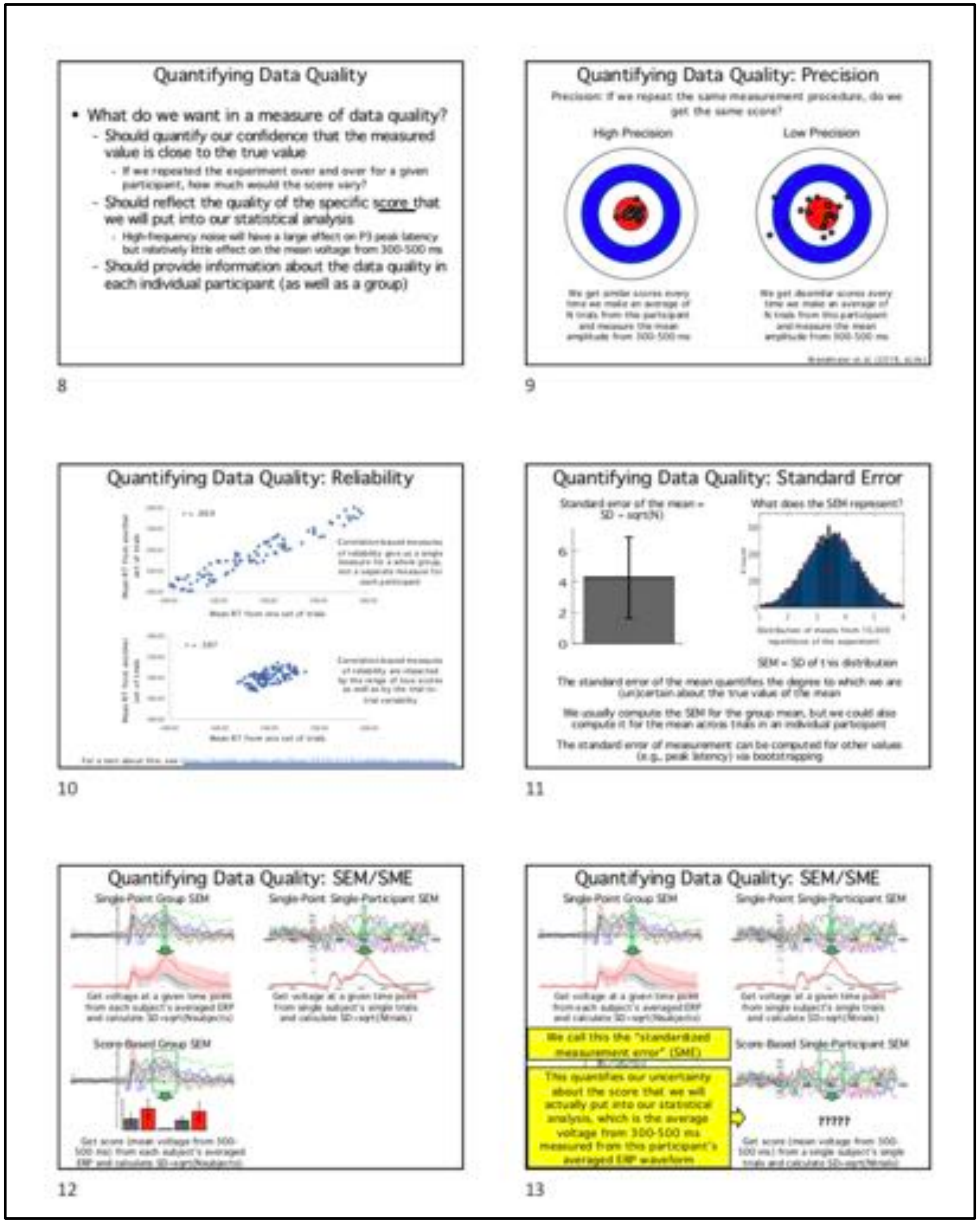
We promote best practices in ERP research via workshops, software, books, advice, data sharing, & methods development.

Steve Luck
lucklab.ucdavis.edu
mindbrain.ucdavis.edu/people/sjluck
@stevenjluck
Google Scholar Profile

Emily Kappenman
emilykappenman.org
@emilykappenman
Google Scholar Profile

Slides

- A PDF of the slides is available right now at https://bit.ly/3i3QecO
    - Also available at erpinfo.org/virtual-boot-camp
- Please do not attempt to download or share the webinar video
- But the PDF of the slides can be shared under the terms of a Creative Commons license



Quantifying Data Quality

- What do we want in a measure of data quality?
    - Should quantify our confidence that the measured value is close to the true value
        - If we repeated the experiment over and over for a given participant, how much would the score vary?
    - Should reflect the quality of the specific score that we will put into our statistical analysis
        - High-frequency noise will have a large effect on P3 peak latency but relatively little effect on the mean voltage from 300-500 ms
    - Should provide information about the data quality in each individual participant (as well as a group)

8

Quantifying Data Quality: Precision

Precision: If we repeat the same measurement procedure, do we get the same score?

We get similar scores every time we make an average of N trials from this participant and measure the mean amplitude from 300-500 ms

We get dissimilar scores every time we make an average of N trials from this participant and measure the mean amplitude from 300-500 ms

9

Quantifying Data Quality: Reliability

Mean RT from another set of trials
Mean RT from one set of trials

Correlation-based measures of reliability give us a single measure for a whole group, not a separate measure for each participant

Correlation-based measures of reliability are impacted by the range of true scores as well as by the trial-to-trial variability

For a rant about this, see [illegible]

10

Quantifying Data Quality: Standard Error

Standard error of the mean = SD ÷ sqrt(N)

What does the SEM represent?

Distribution of means from 10,000 repetitions of the experiment

SEM = SD of this distribution

The standard error of the mean quantifies the degree to which we are (un)certain about the true value of the mean

We usually compute the SEM for the group mean, but we could also compute it for the mean across trials in an individual participant

The standard error of measurement can be computed for other values (e.g., peak latency) via bootstrapping

11

Quantifying Data Quality: SEM/SME

Single-Point Group SEM

Get voltage at a given time point from each subject's averaged ERP and calculate SD÷sqrt(Nsubjects)

Single-Point Single-Participant SEM

Get voltage at a given time point from single subject's single trials and calculate SD÷sqrt(Ntrials)

Score-Based Group SEM

Get score (mean voltage from 300-500 ms) from each subject's averaged ERP and calculate SD÷sqrt(Nsubjects)

12

Quantifying Data Quality: SEM/SME

Single-Point Group SEM

Get voltage at a given time point from each subject's averaged ERP and calculate SD÷sqrt(Nsubjects)

Single-Point Single-Participant SEM

Get voltage at a given time point from single subject's single trials and calculate SD÷sqrt(Ntrials)

We call this the "standardized measurement error" (SME)

This quantifies our uncertainty about the score that we will actually put into our statistical analysis, which is the average voltage from 300-500 ms measured from this participant's averaged ERP waveform

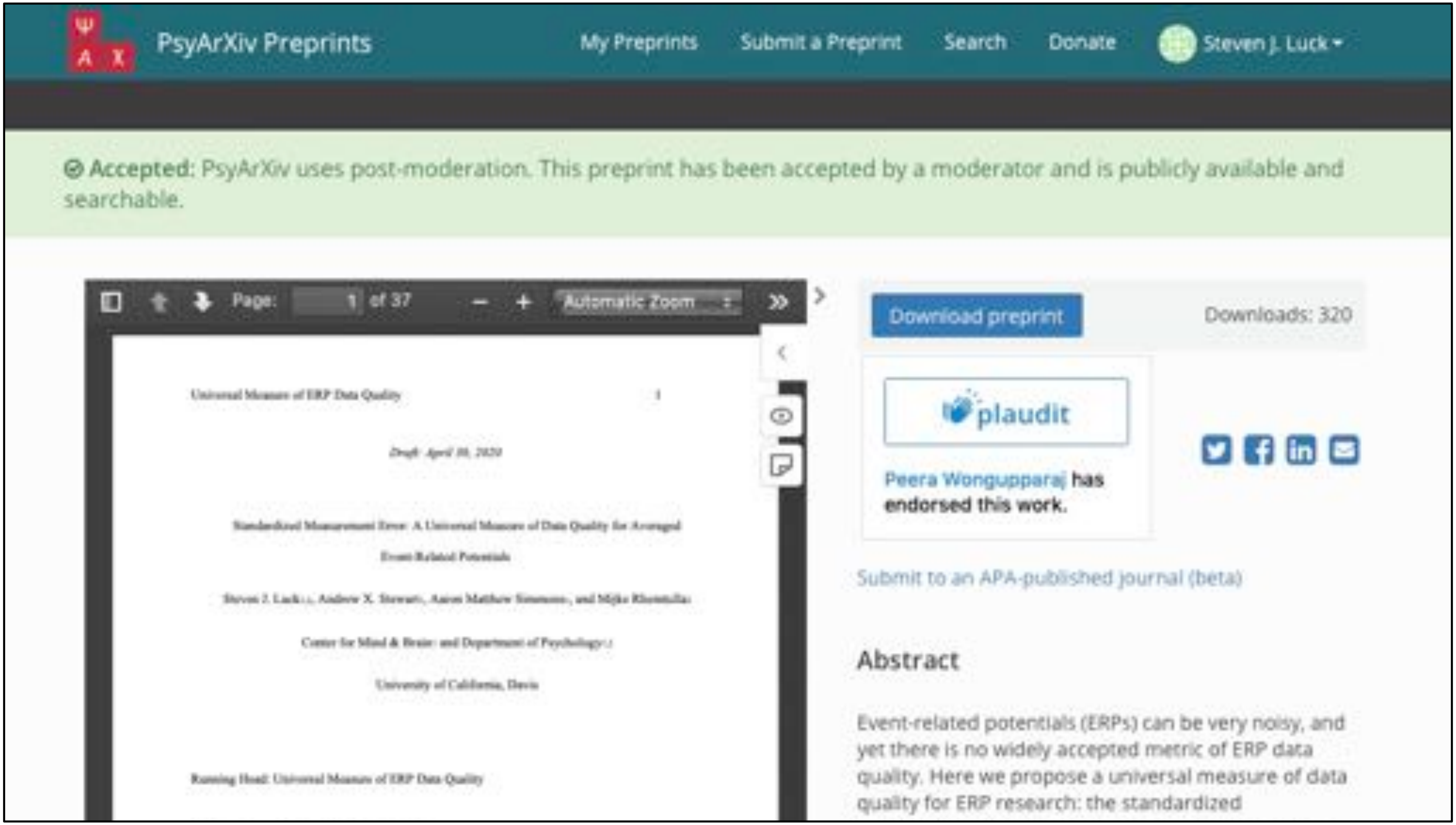
Score-Based Single-Participant SEM

?????

Get score (mean voltage from 300-500 ms) from a single subject's single trials and calculate SD÷sqrt(Ntrials)

13

Preprint



Accepted: PsyArXiv uses post-moderation. This preprint has been accepted by a moderator and is publicly available and searchable.

Universal Measure of ERP Data Quality 1

Draft: April 30, 2020

Standardized Measurement Error: A Universal Measure of Data Quality for Averaged Event-Related Potentials

Steven J. Luck[1,2], Andrew X. Stewart[1], Aaron Matthew Simmons[1], and Mijke Rhemtulla[2]

Center for Mind & Brain[1] and Department of Psychology[1,2]

University of California, Davis

Running Head: Universal Measure of ERP Data Quality

Downloads: 320

Peera Wongupparaj has endorsed this work.

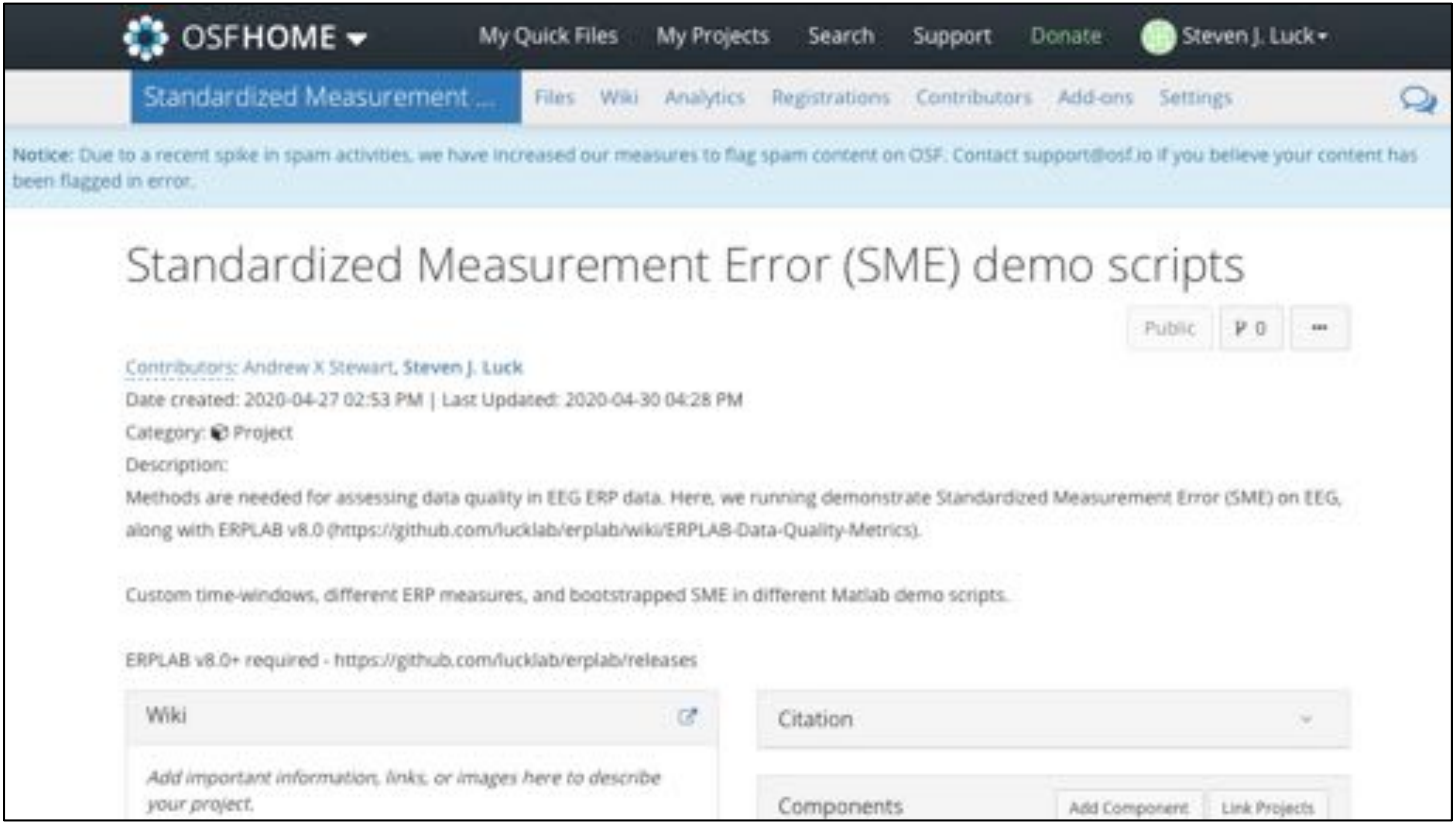
Submit to an APA-published journal (beta)

Abstract

Event-related potentials (ERPs) can be very noisy, and yet there is no widely accepted metric of ERP data quality. Here we propose a universal measure of data quality for ERP research: the standardized

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2020). Standardized Measurement Error: A Universal Measure of Data Quality for Averaged Event-Related Potentials. *PsyArXiv*. https://doi.org/10.31234/osf.io/dwm64

Demo Data and Scripts



Notice: Due to a recent spike in spam activities, we have increased our measures to flag spam content on OSF. Contact support@osf.io if you believe your content has been flagged in error.

Standardized Measurement Error (SME) demo scripts

Contributors: Andrew X Stewart, **Steven J. Luck**

Date created: 2020-04-27 02:53 PM | Last Updated: 2020-04-30 04:28 PM

Category: Project

Description:

Methods are needed for assessing data quality in EEG ERP data. Here, we running demonstrate Standardized Measurement Error (SME) on EEG, along with ERPLAB v8.0 (https://github.com/lucklab/erplab/wiki/ERPLAB-Data-Quality-Metrics).

Custom time-windows, different ERP measures, and bootstrapped SME in different Matlab demo scripts.

ERPLAB v8.0+ required - https://github.com/lucklab/erplab/releases

Wiki

Add important information, links, or images here to describe your project.

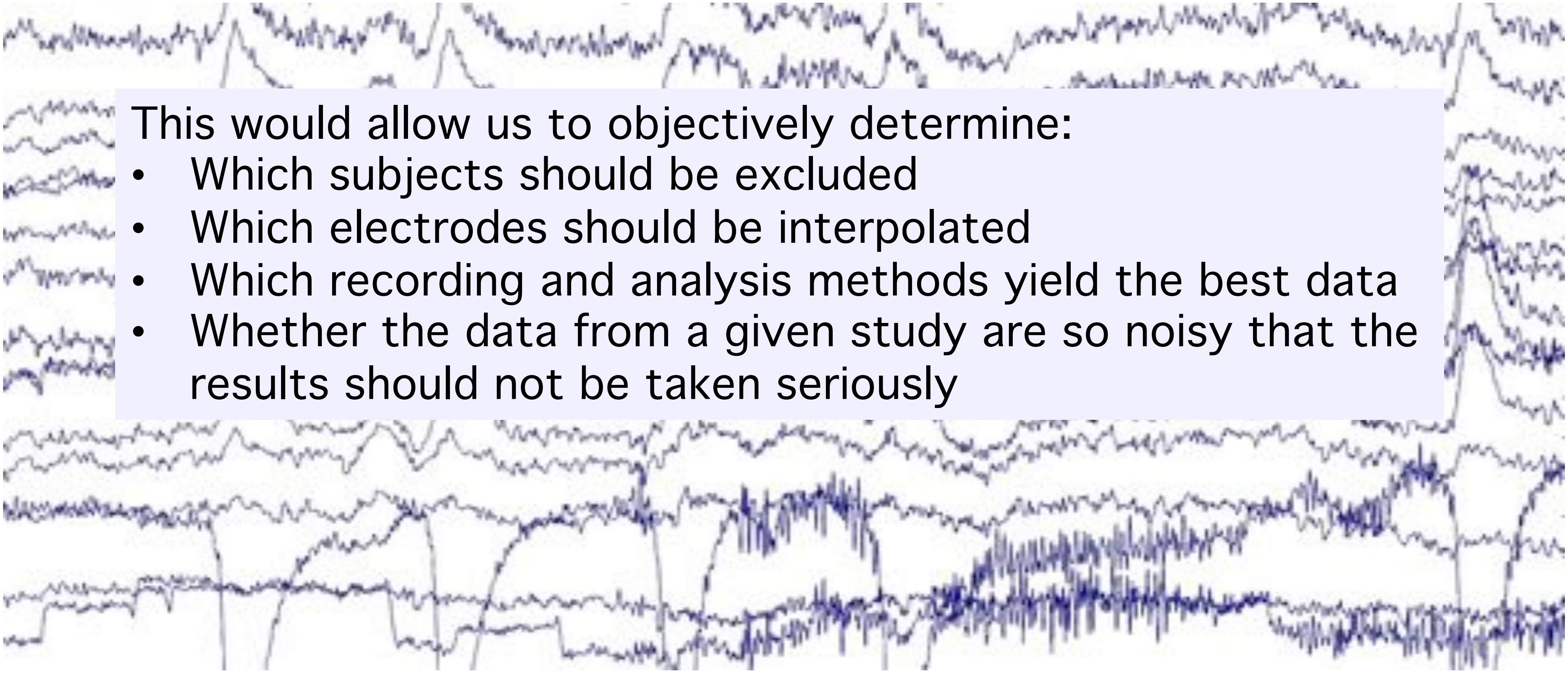
Citation

Components

https://osf.io/a4huc/

We must be insane to think we can average together a few dozen trials of EEG data and get a stable ERP waveform

At a minimum, we should have an objective metric of the quality of our averaged ERP waveforms



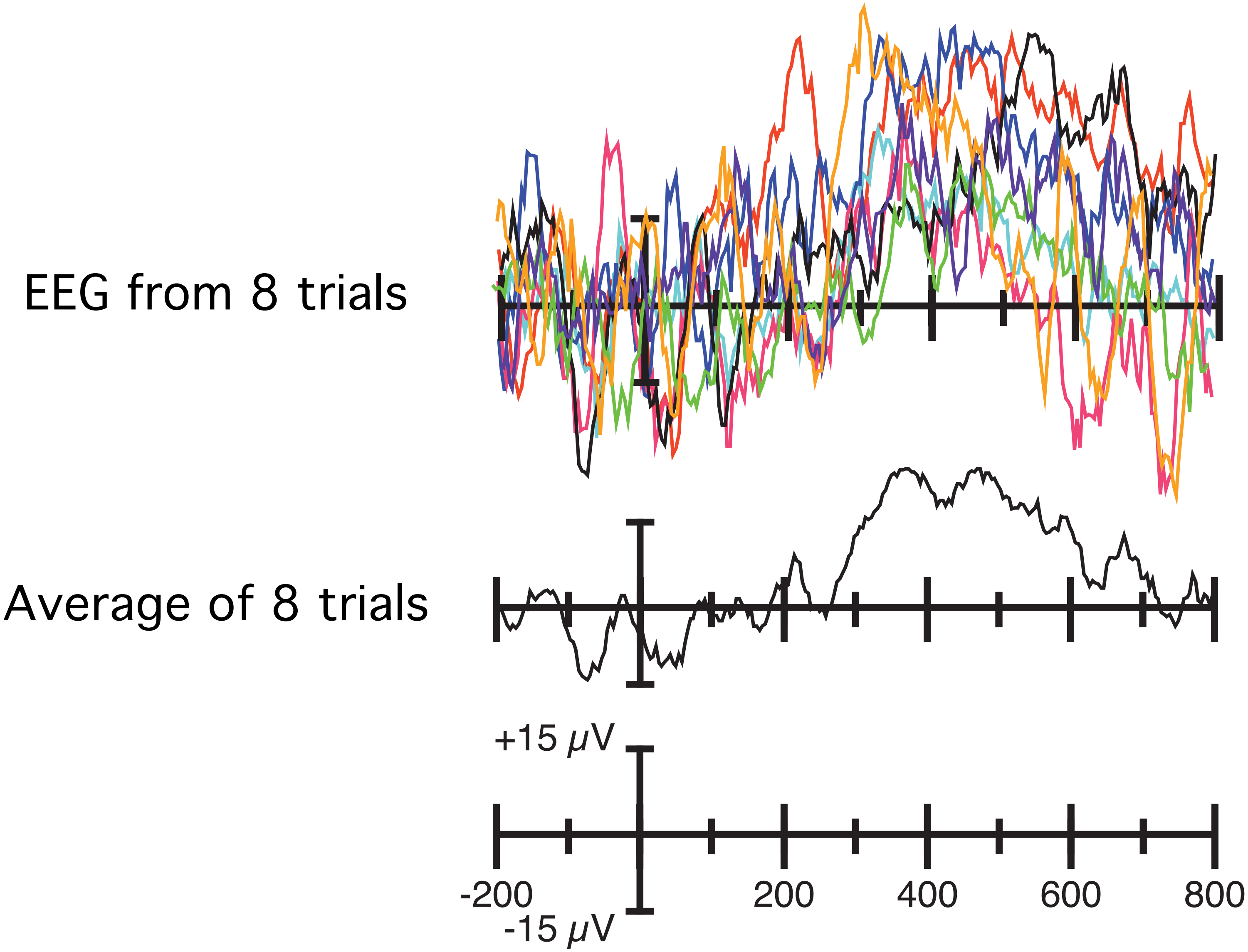
This would allow us to objectively determine:

- Which subjects should be excluded
- Which electrodes should be interpolated
- Which recording and analysis methods yield the best data
- Whether the data from a given study are so noisy that the results should not be taken seriously

Today’s Plan

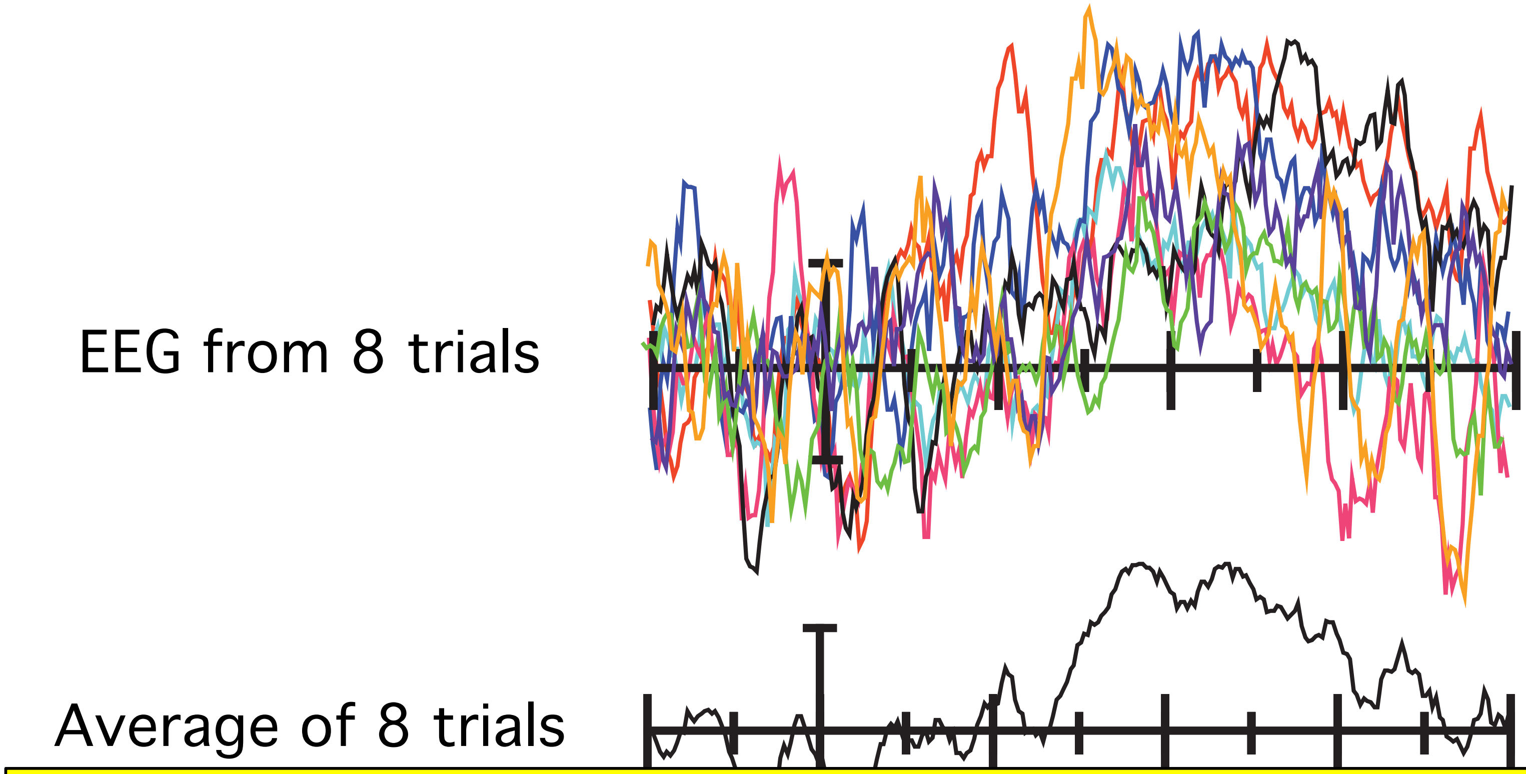
- Desirable properties of a metric of ERP data quality
  - We need a measure of “precision”
- Using the standard error to quantify precision
  - The traditional approach (standard error of group mean)
  - Standardized Measurement Error (SME): the standard error of a single subject’s amplitude or latency “score”
- Computing the “analytic” SME using ERPLAB Toolbox
  - Appropriate if your score is the mean amplitude over some time window (e.g., 300-500 ms)
- The bootstrapped SME for other scores (e.g., peaks)
  - Requires some simple Matlab scripting
- Using SME to understand how measurement error impacts effect size and statistical power
  - Can predict exactly how the effect size or statistical power will change if you increase or decrease the number of trials

Quantifying Data Quality



You've averaged N trials together. Do you now have a reasonable estimate of this participant's P3 amplitude? What about the onset latency of the P2?

Quantifying Data Quality

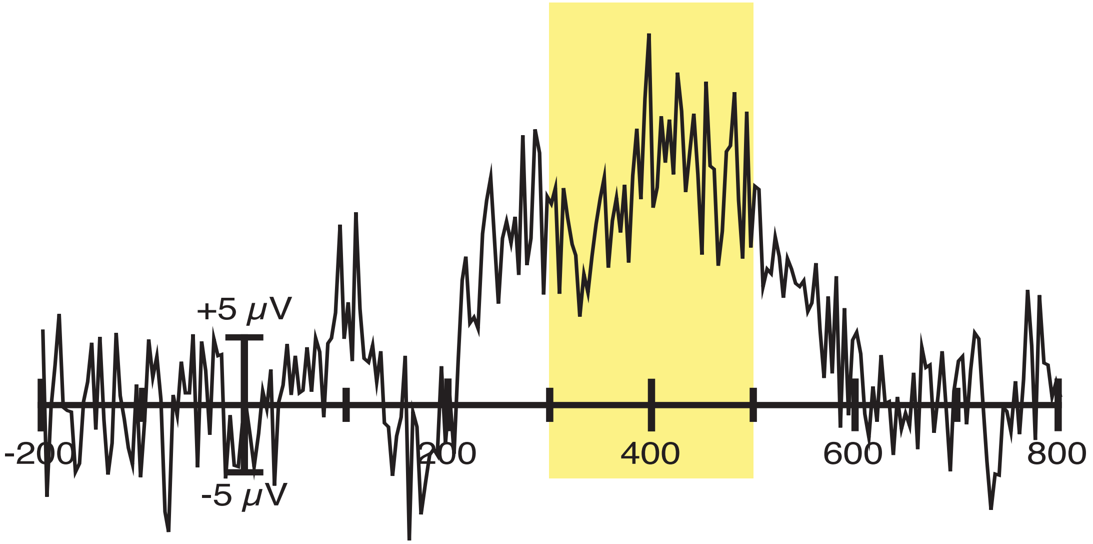


- Most ERP studies obtain amplitude or latency “scores” from averaged ERP waveforms, ignoring trial-to-trial variation
- We will be focusing on this situation
- Different methods would be needed to quantify data quality for single-trial analysis methods

You’ve averaged N trials together. Do you now have a reasonable estimate of this participant’s P3 amplitude? What about the onset latency of the P2?

Quantifying Data Quality

- What do we want in a measure of data quality?
  - Should quantify our confidence that the measured value is close to the true value for that participant
    - If we repeated the experiment over and over for a given participant, how much would the score vary?
  - Should reflect the quality of the specific <u>score</u> that we will put into our statistical analysis
    - High-frequency noise will have a large effect on peak amplitude from 300-500 ms but relatively little effect on the mean voltage

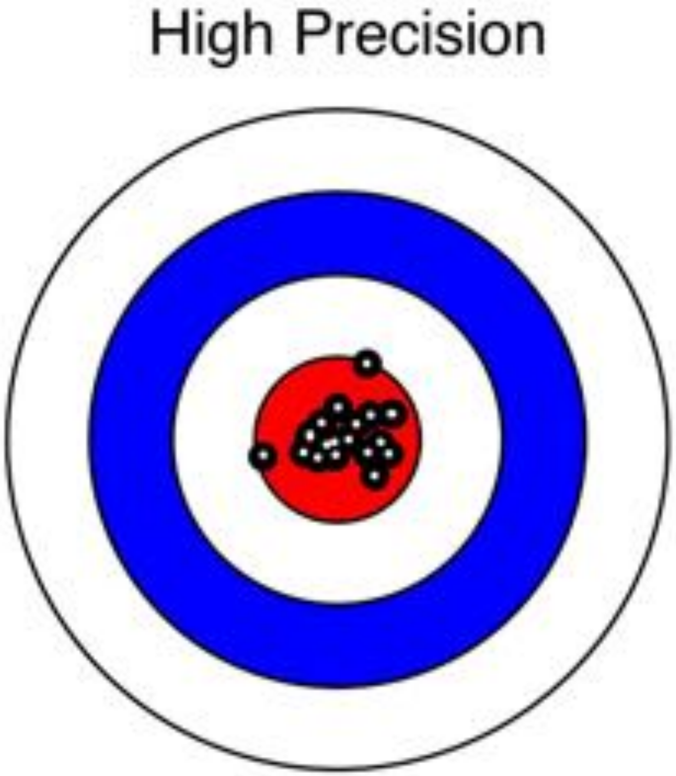


Quantifying Data Quality

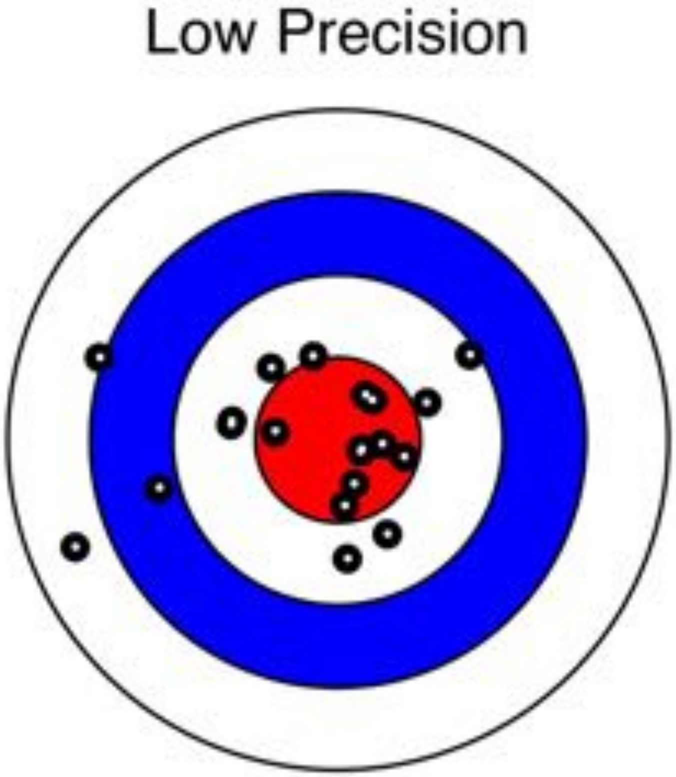
- What do we want in a measure of data quality?
  - Should quantify our confidence that the measured value is close to the true value for that participant
    - If we repeated the experiment over and over for a given participant, how much would the score vary?
  - Should reflect the quality of the specific <u>score</u> that we will put into our statistical analysis
    - High-frequency noise will have a large effect on peak amplitude from 300-500 ms but relatively little effect on the mean voltage
  - Should provide information about data quality for each individual participant (as well as a group)

Quantifying Data Quality: Precision

Precision: If we repeat the same measurement procedure, do we get the same score?



We get similar scores every time we make an average of N trials from this participant and measure the mean amplitude from 300-500 ms



We get dissimilar scores every time we make an average of N trials from this participant and measure the mean amplitude from 300-500 ms

Brandmaier et al. (2018, eLife)

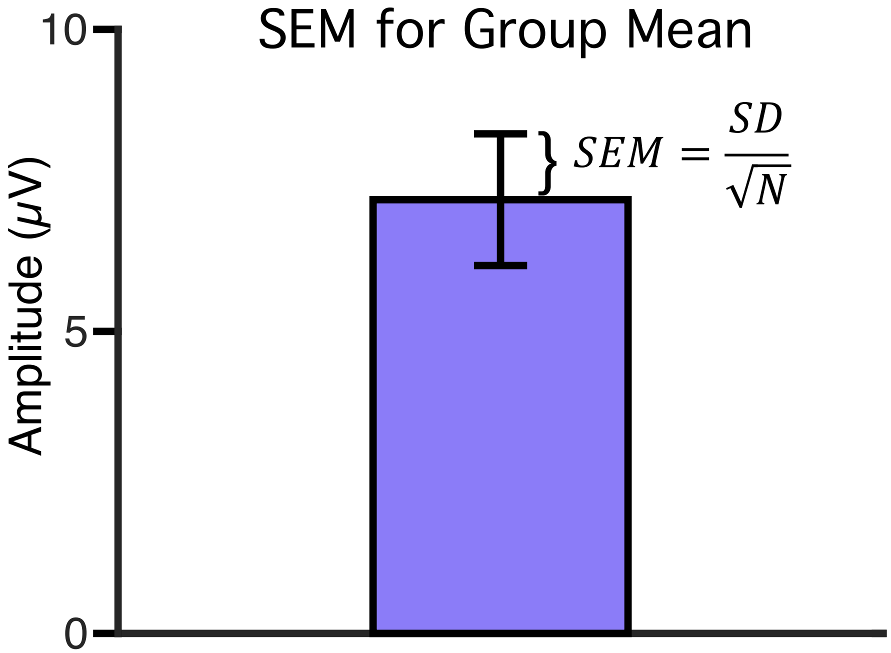
Quantifying Data Quality: Reliability

- Traditional psychometric measures of reliability:
  - Provide a group value but not single-subject values
  - Are impacted by the range of true scores, not just by the quality of the data

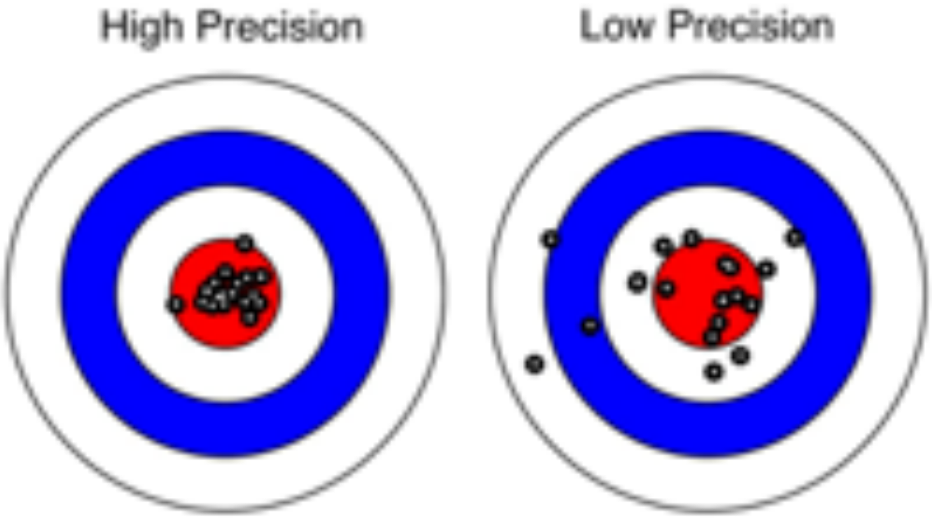
For a rant about this, see
https://lucklab.ucdavis.edu/blog/2019/2/19/reliability-and-precision

For a thoughtful paper, see
Hedge et al. (2018), https://doi.org/10.3758/s13428-017-0935-1

Quantifying Data Quality: Standard Error



Mean P3 amplitude (±1 SEM) across a group of 12 participants



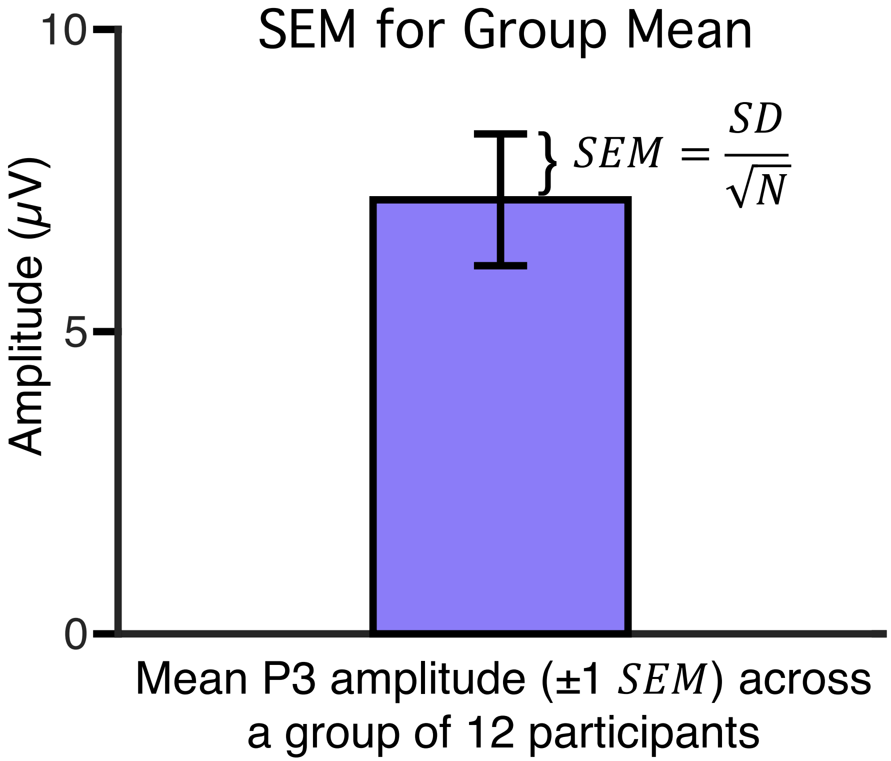
- Make an averaged ERP waveform for each of 12 subjects
- Measure P3 amplitude in each of the 12 averaged ERP waveforms
- Take the mean of these 12 values
- Take the SD of these 12 values
- $\text{SEM} = \; SD \,/\, \sqrt{Nsubjects}$

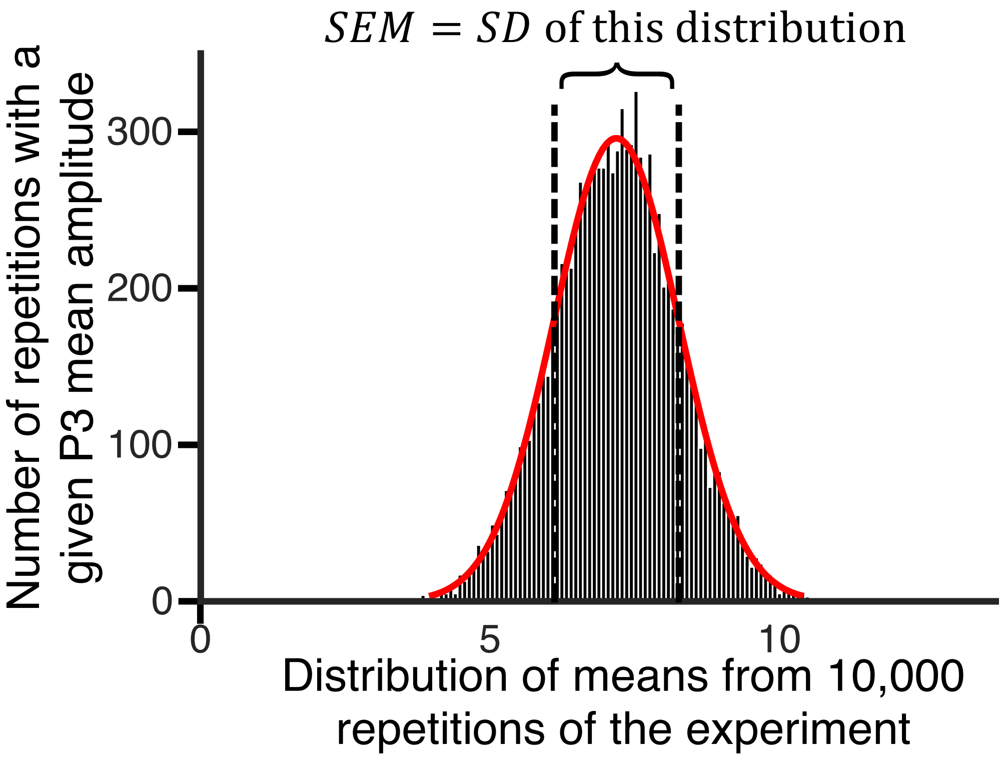
But what does the SEM actually represent?

The SEM tells us the precision of the group mean

If we repeated the experiment 10,000 times, calculating the group mean for each experiment, how much variability would there be in the group mean?

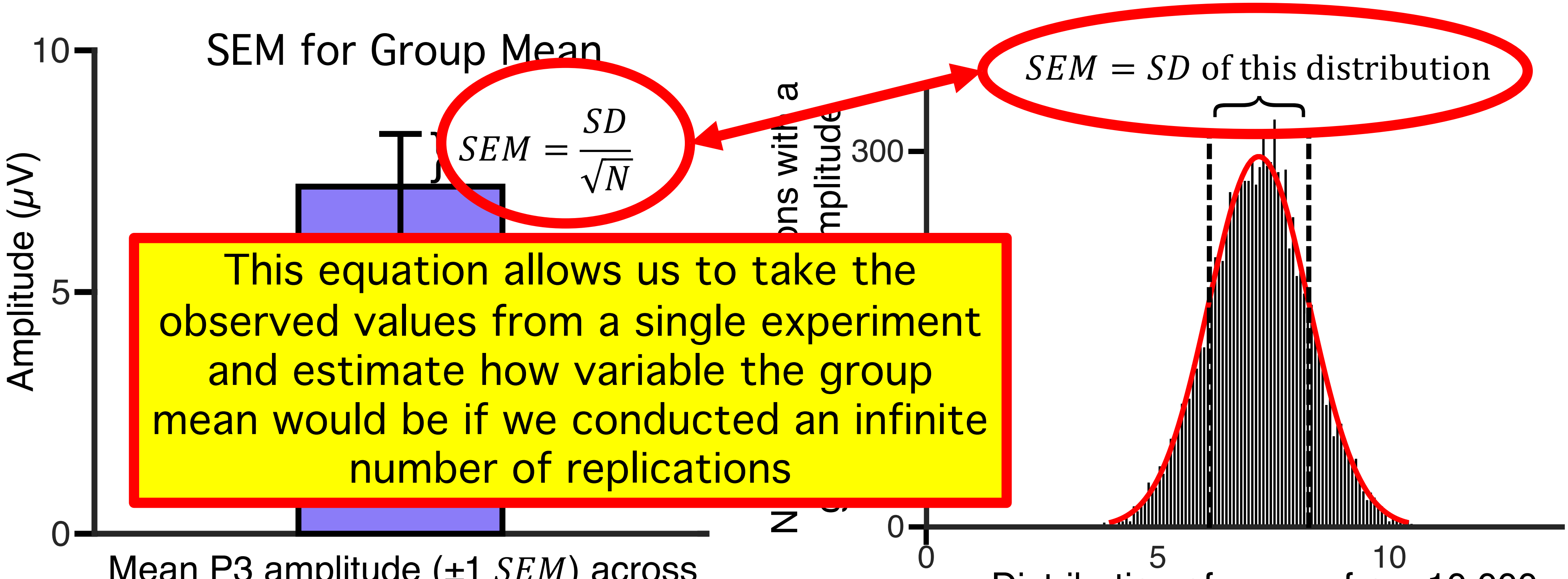
Quantifying Data Quality: Standard Error





If we repeated the experiment 10,000 times, calculating the group mean for each experiment, how much variability would there be in the group mean?

Quantifying Data Quality: Standard Error



$$SEM = \frac{SD}{\sqrt{N}}$$

$SEM = SD$ of this distribution

This equation allows us to take the observed values from a single experiment and estimate how variable the group mean would be if we conducted an infinite number of replications

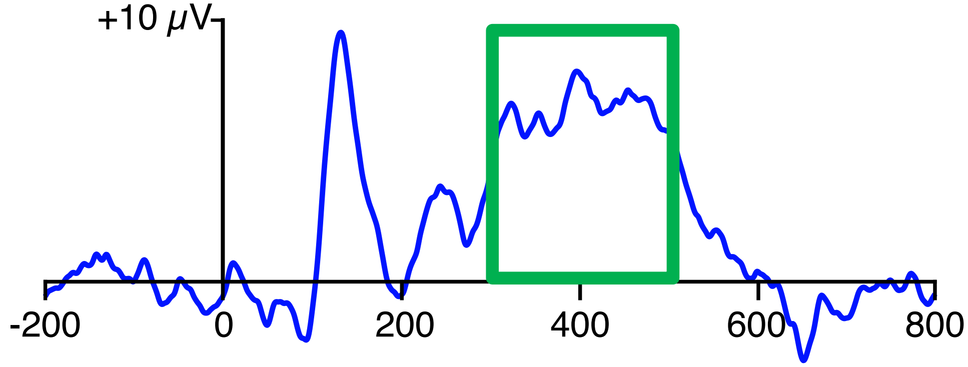
Mean P3 amplitude (±1 SEM) across a group of 12 participants

Distribution of means from 10,000 repetitions of the experiment

- This equation does not assume a normal distribution
- This equation only works for the mean (not for the median, etc.)
  - We can use *bootstrapping* for other kinds of standard errors
- Our metric of ERP data quality involves computing the standard error of the *score* from a single subject’s averaged ERP waveform
  - Example: Standard error of peak amplitude for the P3 wave
  - Peak is complicated, so we will start with “time-window mean amplitude”

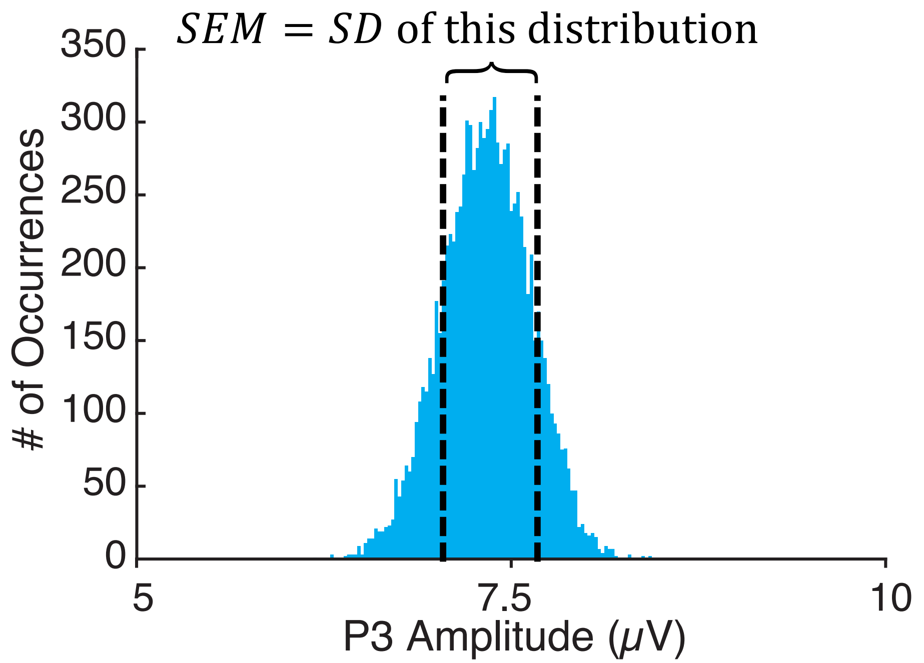
Standardized Measurement Error (SME)

Averaged ERP from a single subject

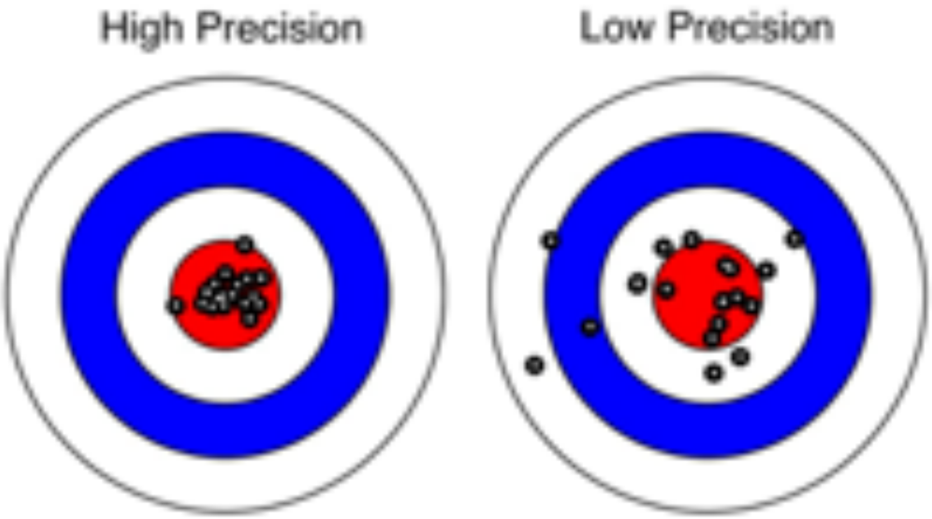


Time-window mean amplitude: mean voltage during a time period (e.g., 300-500 ms), measured from an averaged ERP waveform

We want to quantify the precision of this measure



Distribution of means from 10,000 repetitions of the experiment

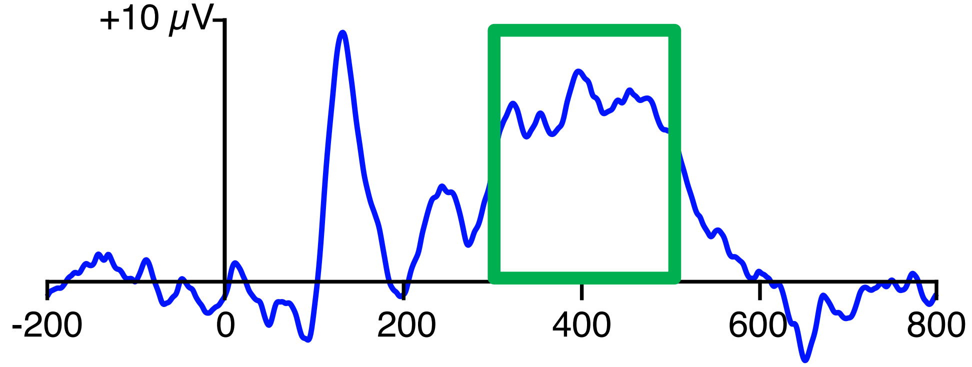


Empirical approach

- Repeat the session 10,000 times for this subject
- For each session, make an averaged ERP waveform and calculate the time-window mean amplitude (300-500 ms)
- Take the SD of these 10,000 values
- This SD is the standard error of measurement for the time-window mean amplitude

Standardized Measurement Error (SME)

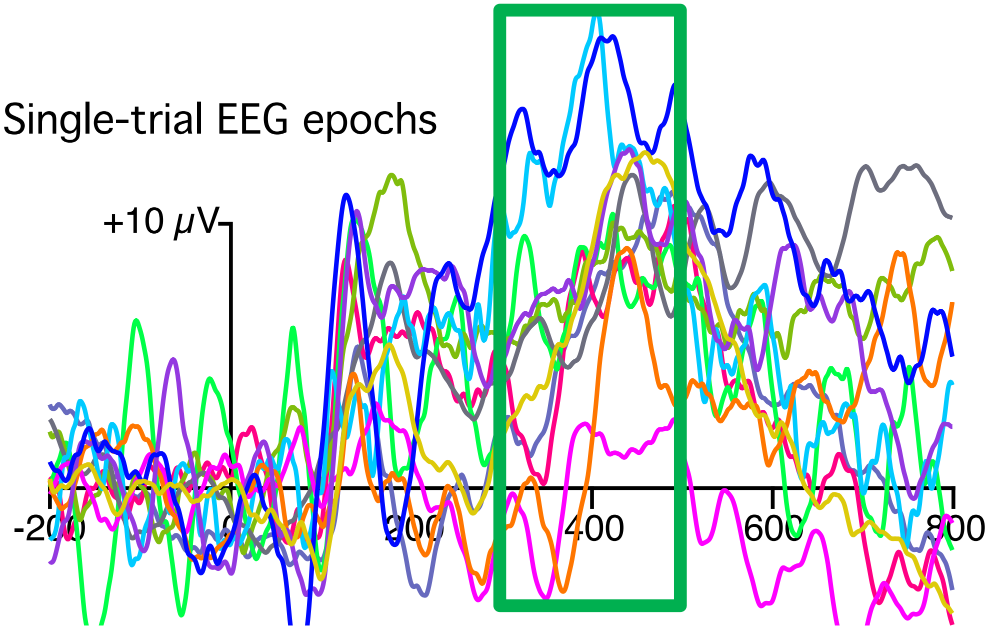
Averaged ERP from a single subject

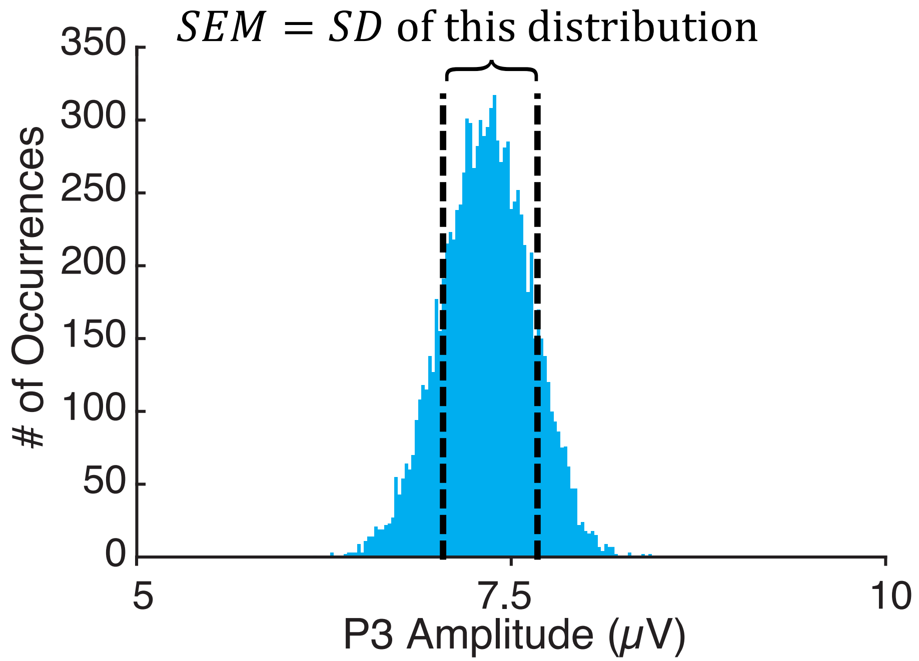


Time-window mean amplitude: mean voltage during a time period (e.g., 300-500 ms), measured from an averaged ERP waveform

We want to quantify the precision of this measure

Single-trial EEG epochs





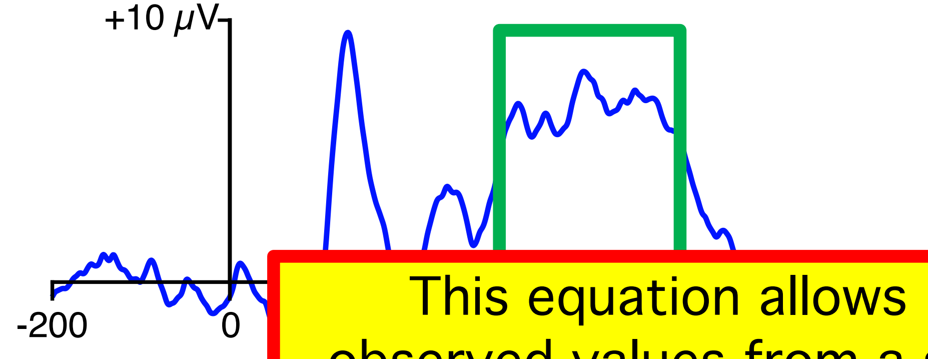
Distribution of means from 10,000 repetitions of the experiment

Analytic approach

- Measure the time-window mean amplitude (300-500 ms) on each trial
- Take the SD of these values
- $\mathrm{SEM} = SD \,/\, \sqrt{Ntrials}$
- When the SEM is used in this way, we call it the Standardized Measurement Error (SME)

Standardized Measurement Error (SME)

Averaged ERP from a single subject



Time-windo

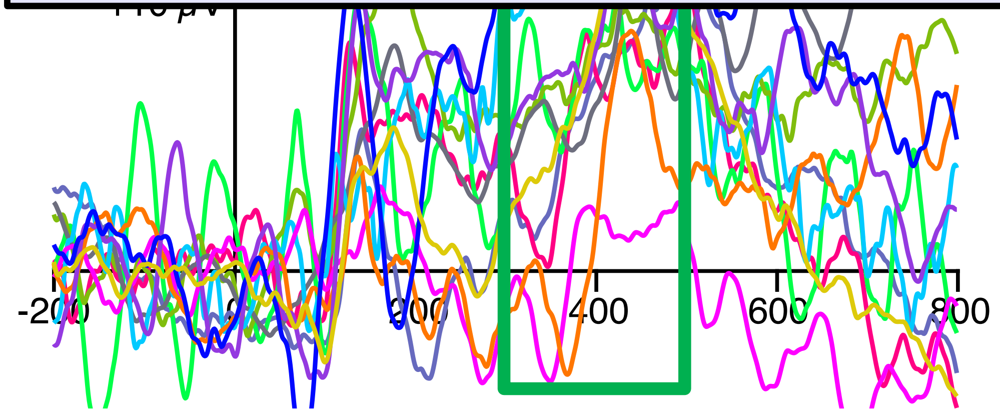
during a ti

measured f

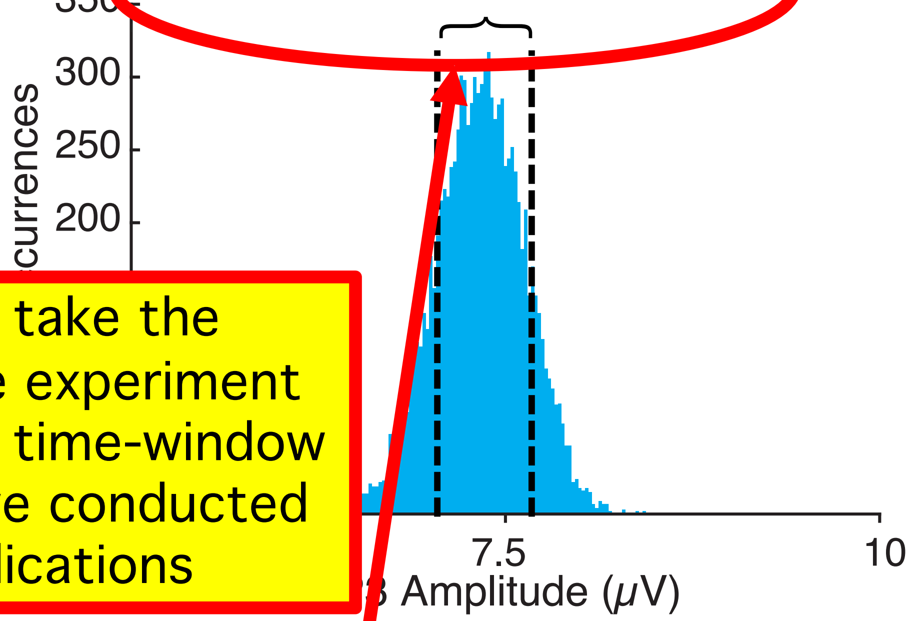
We want to quantify the precision

This equation allows us to take the observed values from a single experiment and estimate how variable the time-window mean amplitude would be if we conducted an infinite number of replications

ERPLAB (v8 or later) automatically calculates SME with default time windows whenever you create an averaged ERP waveform



$SEM = SD$ of this distribution



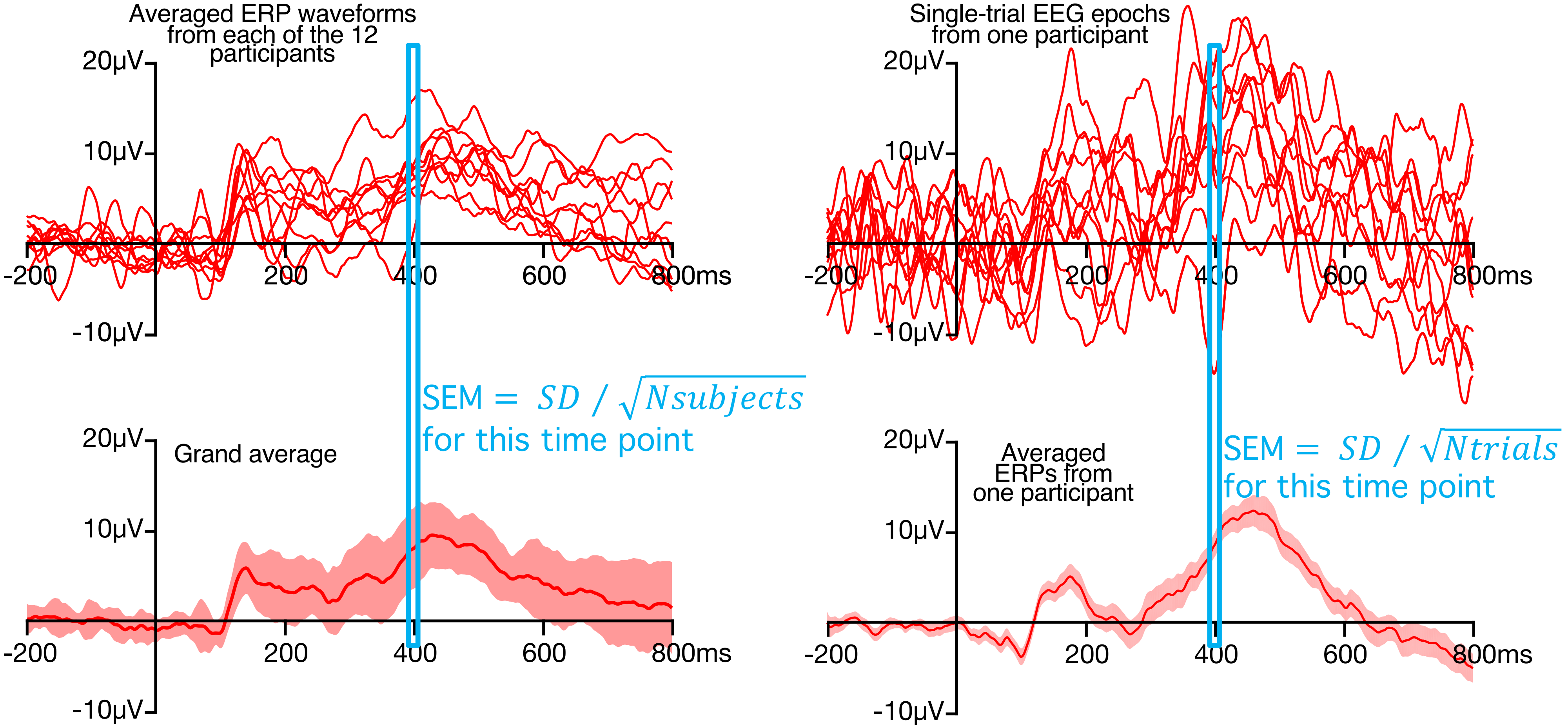
Distribution of means from 10,000 repetitions of the experiment

lytic approach

easure the time-window mean amplitude (300-500 ms) on each trial

- Take the SD of these values
- $\mathrm{SEM} = SD \,/\, \sqrt{Ntrials}$
- When the SEM is used in this way, we call it the Standardized Measurement Error (SME)

Separate SEM at Each Time Point?



Averaged ERP waveforms from each of the 12 participants

Single-trial EEG epochs from one participant

Grand average

Averaged ERPs from one participant

SEM = $SD \ / \ \sqrt{Nsubjects}$ for this time point

SEM = $SD \ / \ \sqrt{Ntrials}$ for this time point

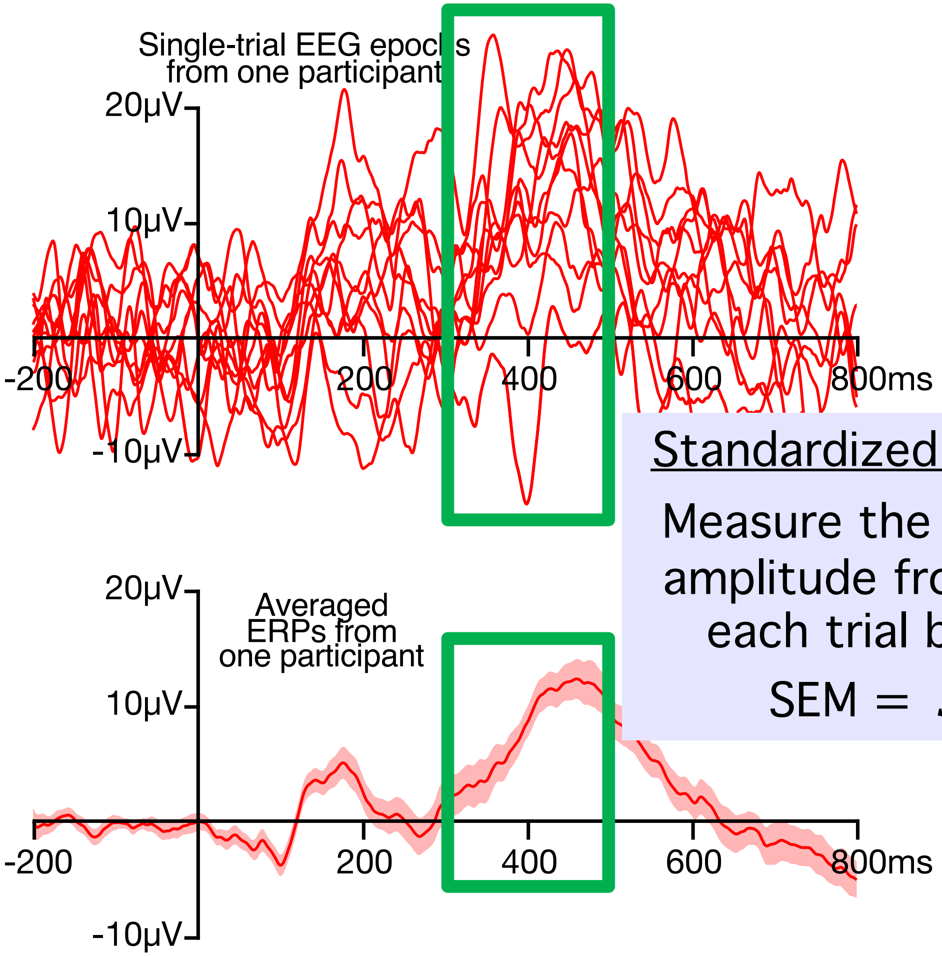
Reflects both measurement error and true differences between subjects

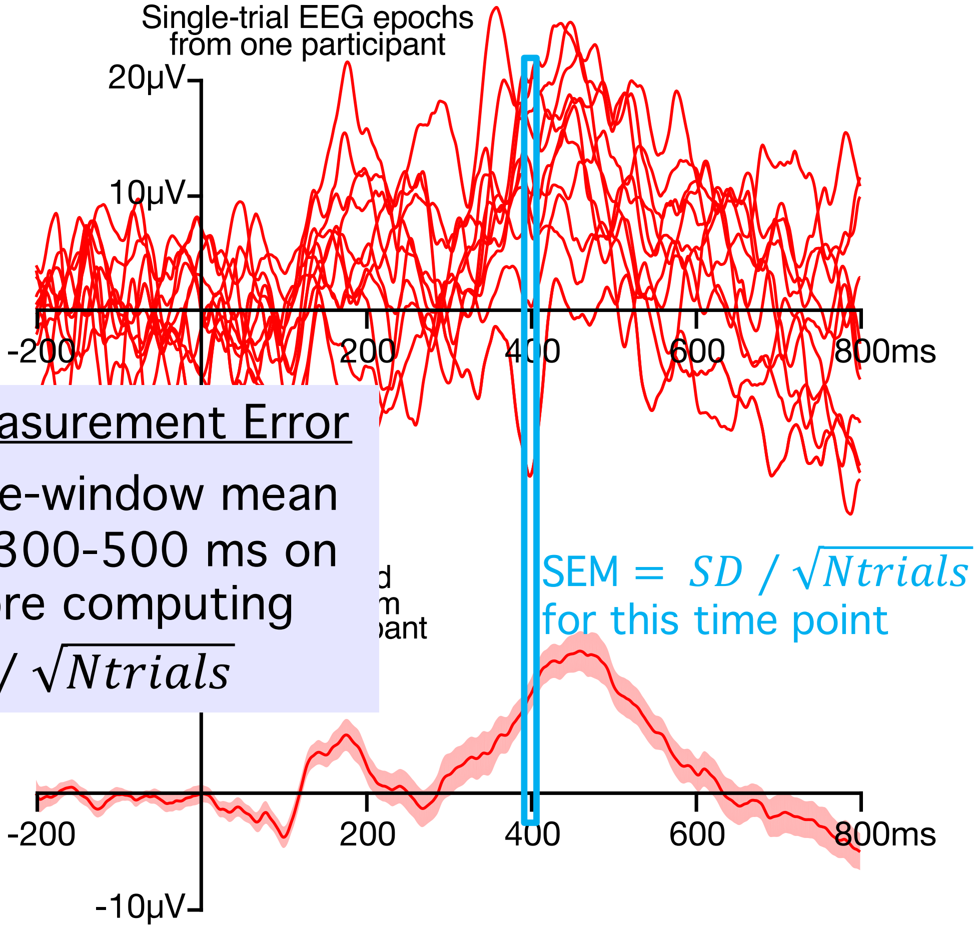
Solely reflects measurement error

Doesn’t tell us about the precision of our score (time-window mean amplitude from 300-500 ms)

Doesn’t tell us about the precision of our score (time-window mean amplitude from 300-500 ms)

Separate SEM at Each Time Point?





SEM $= SD / \sqrt{Ntrials}$
for this time point

Standardized Measurement Error

Measure the time-window mean amplitude from 300-500 ms on each trial before computing

SEM $= SD / \sqrt{Ntrials}$

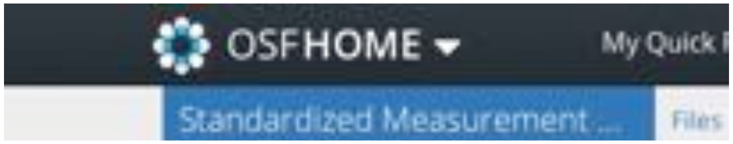
Solely reflects measurement error

Tells us about the precision of our score (time-window mean amplitude from 300-500 ms)

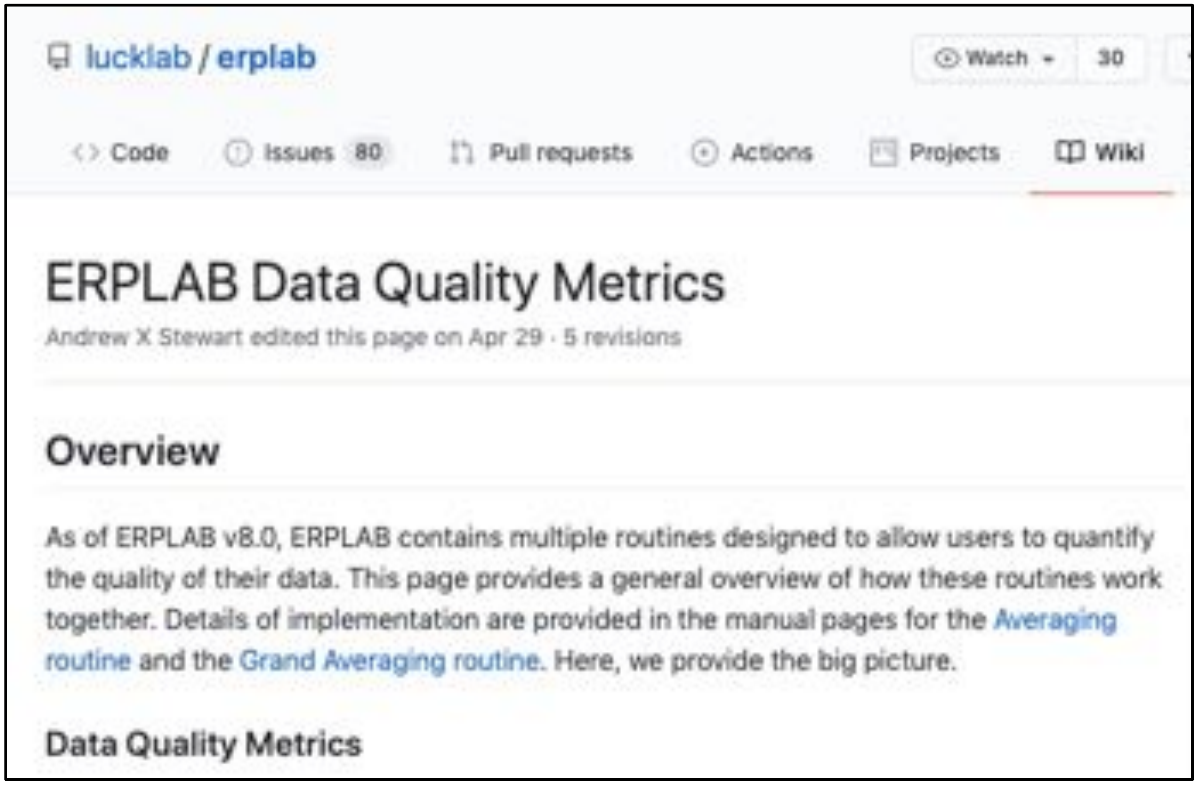
Solely reflects measurement error

Doesn't tell us about the precision of our score (time-window mean amplitude from 300-500 ms)

Example



https://osf.io/a4huc/



ERPLAB Data Quality Metrics

Andrew X Stewart edited this page on Apr 29 · 5 revisions

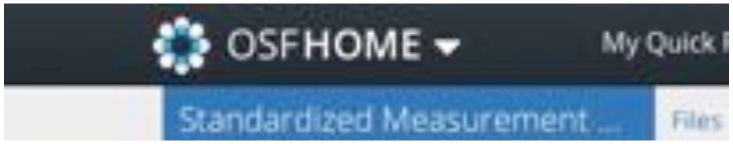
Overview

As of ERPLAB v8.0, ERPLAB contains multiple routines designed to allow users to quantify the quality of their data. This page provides a general overview of how these routines work together. Details of implementation are provided in the manual pages for the Averaging routine and the Grand Averaging routine. Here, we provide the big picture.

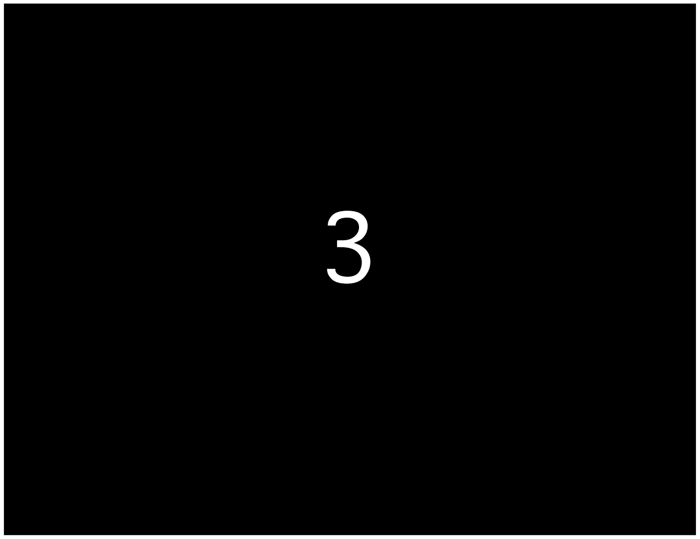
Data Quality Metrics

https://github.com/lucklab/erplab/wiki/ERPLAB-Data-Quality-Metrics

Example



https://osf.io/a4huc/

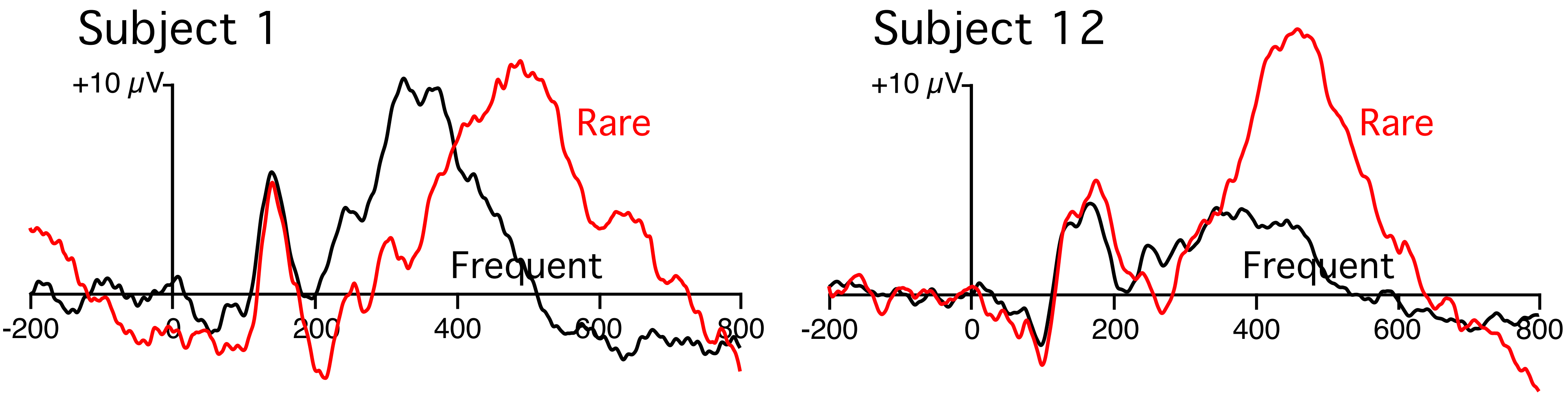


Letters, 80%, Press Left

Digits, 20%, Press Right

Counter-balanced

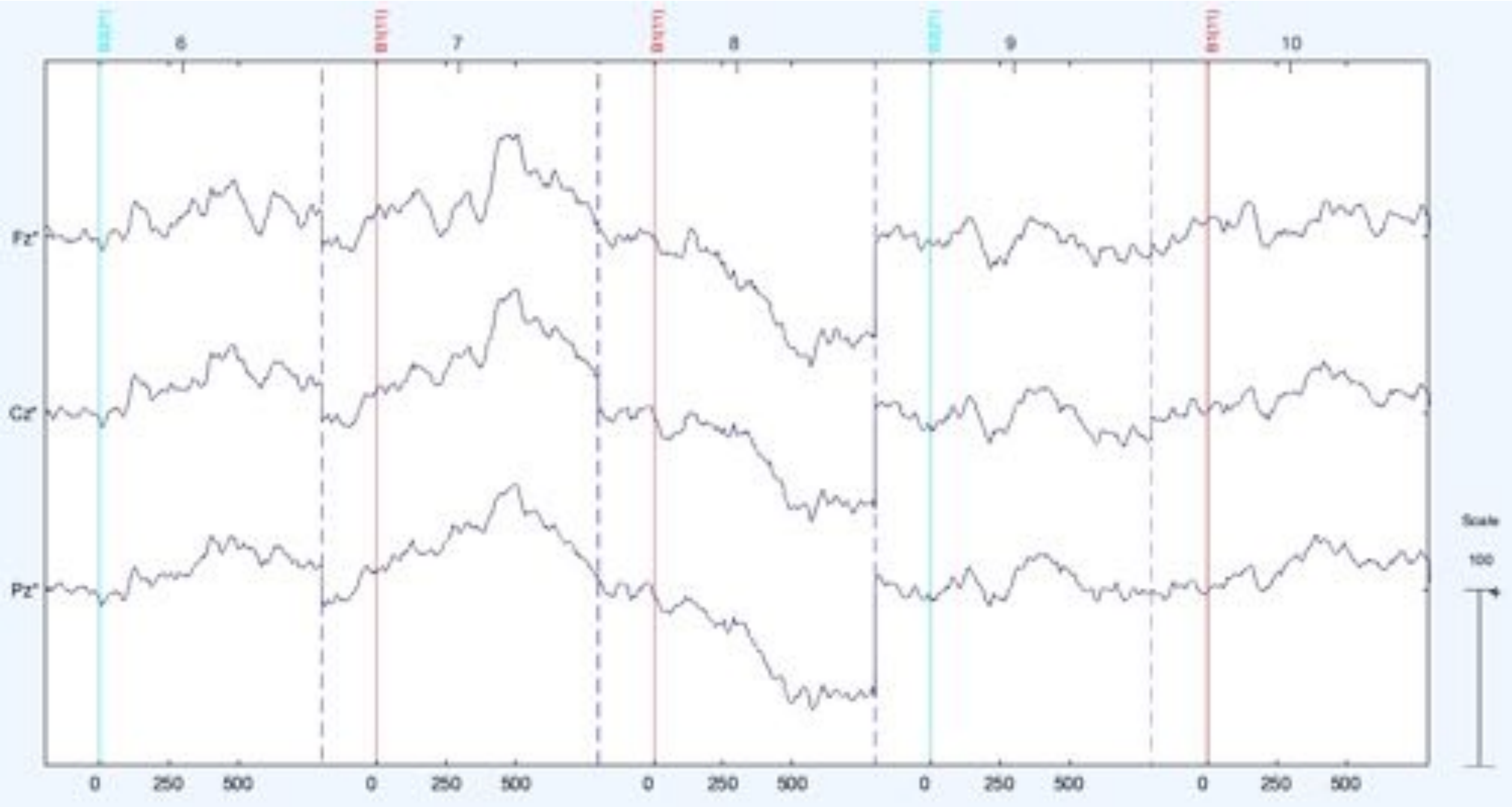
The data have been preprocessed so that every subject has 20 artifact-free rare stimuli and 80 artifact-free frequent stimuli (Fz, Cz, and Pz only)


Subject 1

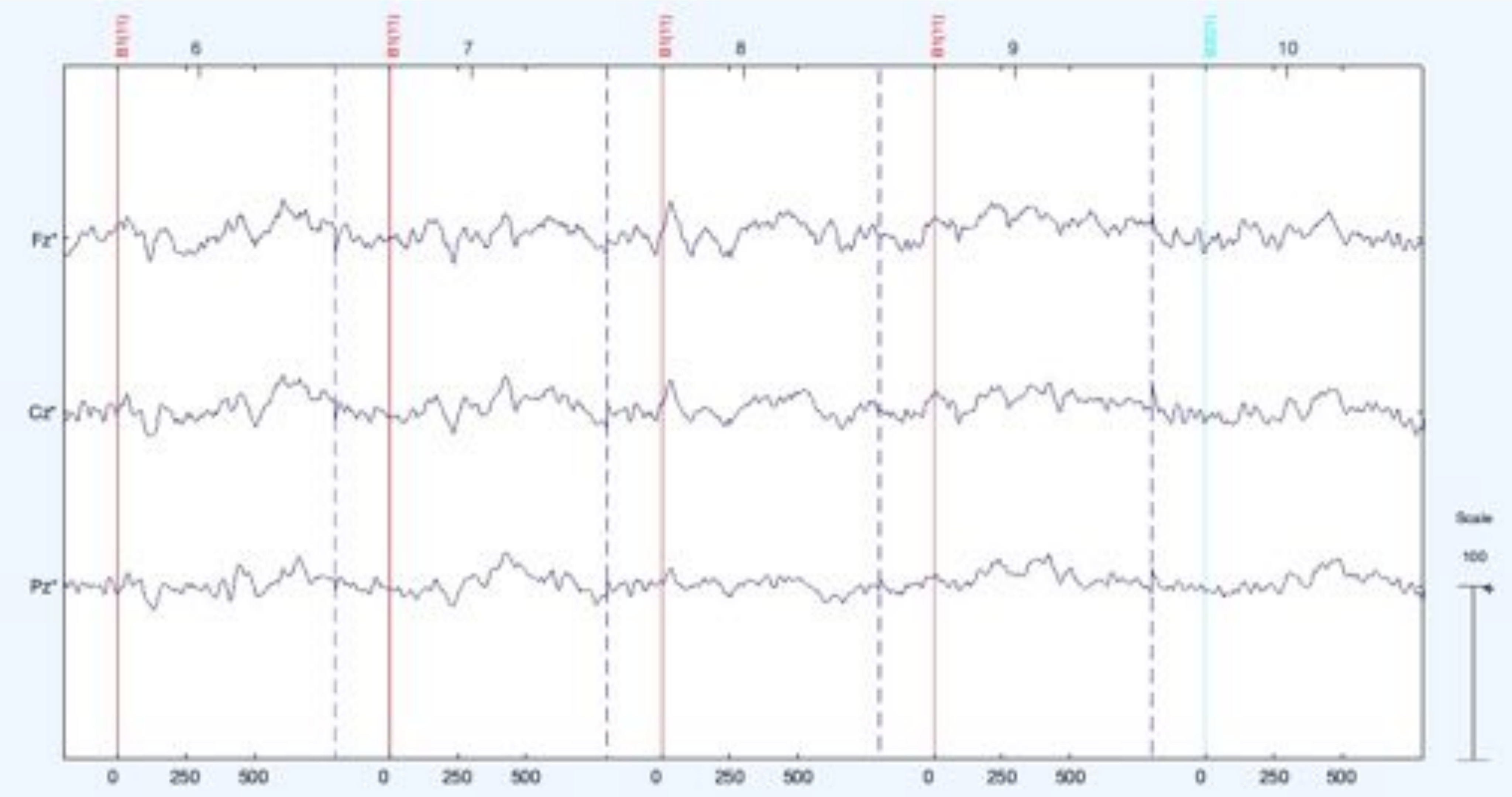
Subject 12

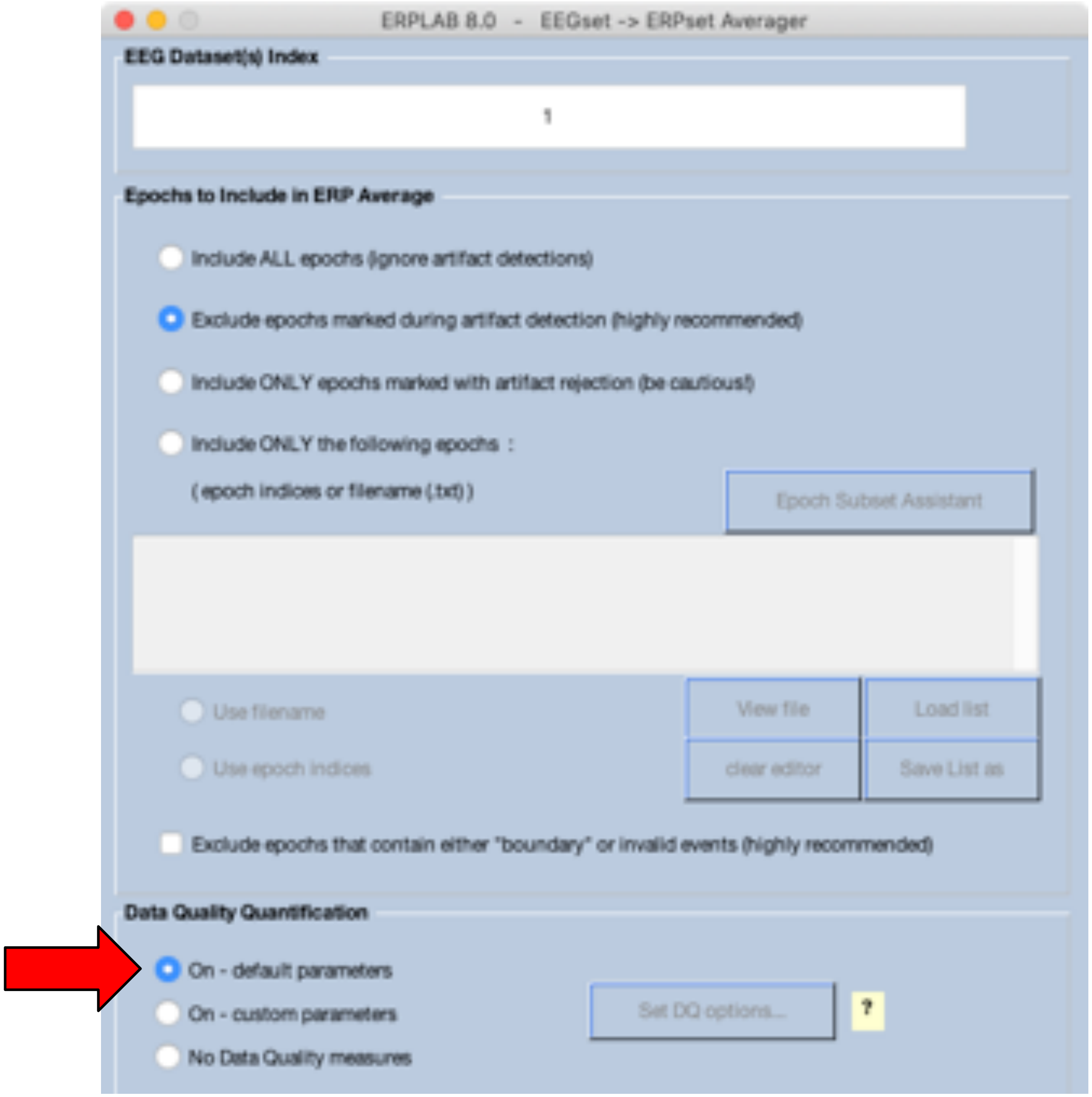
Which subject has noisier data?

Subject 1: Epoched EEG



Subject 12: Epoched EEG





ERPLAB 8.0 - EEGset -> ERPset Averager

EEG Dataset(s) Index

1

Epochs to Include in ERP Average

○ Include ALL epochs (ignore artifact detections)

● Exclude epochs marked during artifact detection (highly recommended)

○ Include ONLY epochs marked with artifact rejection (be cautious!)

○ Include ONLY the following epochs :

(epoch indices or filename (.txt))

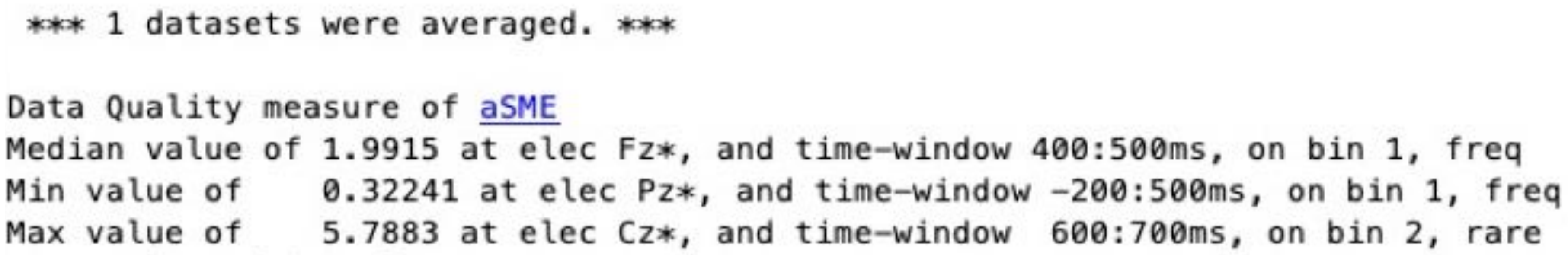
□ Exclude epochs that contain either "boundary" or invalid events (highly recommended)

Data Quality Quantification

● On - default parameters

○ On - custom parameters

○ No Data Quality measures

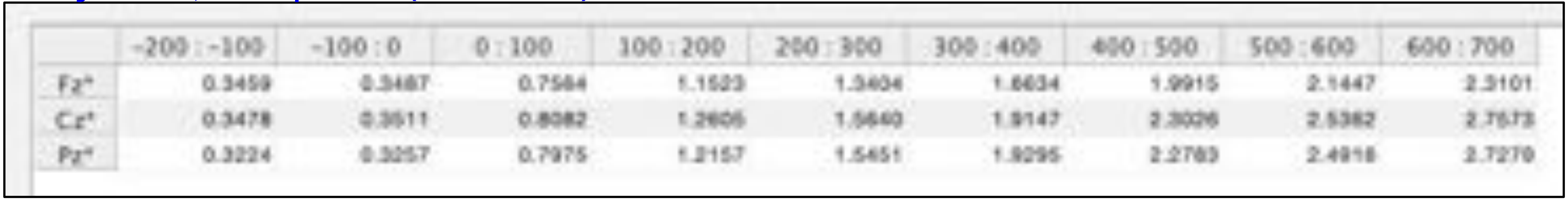


```
 *** 1 datasets were averaged. ***

Data Quality measure of aSME
Median value of 1.9915 at elec Fz*, and time-window 400:500ms, on bin 1, freq
Min value of    0.32241 at elec Pz*, and time-window -200:500ms, on bin 1, freq
Max value of    5.7883 at elec Cz*, and time-window  600:700ms, on bin 2, rare
```

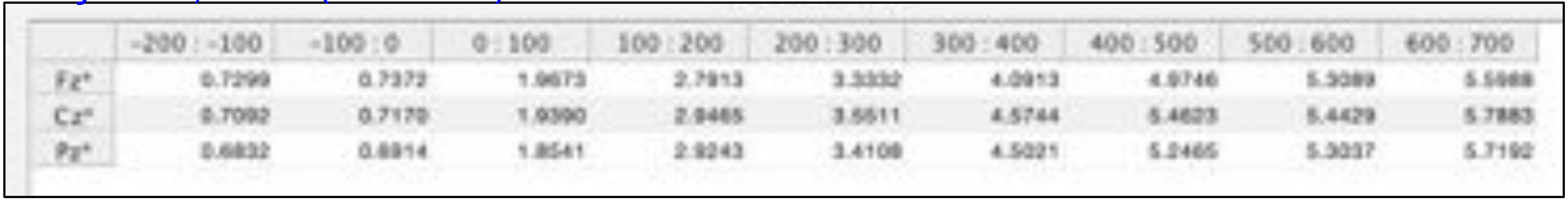
Analytic SME (aSME) Values

Subject 1, Frequent (80 trials)



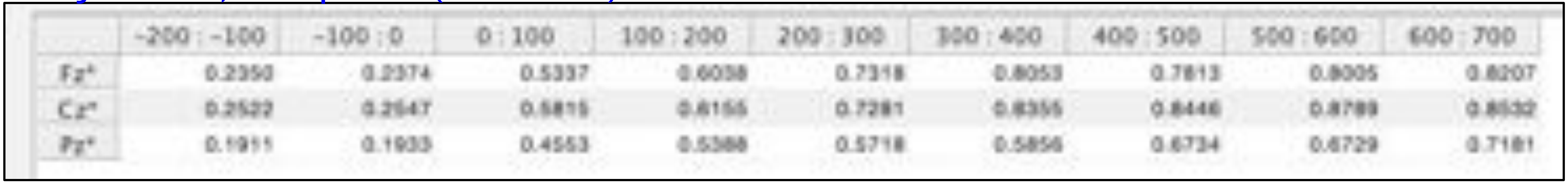
	-200 : -100	-100 : 0	0 : 100	100 : 200	200 : 300	300 : 400	400 : 500	500 : 600	600 : 700
Fz*	0.3459	0.3487	0.7564	1.1523	1.3404	1.6634	1.9915	2.1447	2.3101
Cz*	0.3478	0.3511	0.8082	1.2605	1.5640	1.9147	2.3026	2.5382	2.7573
Pz*	0.3224	0.3257	0.7975	1.2157	1.5451	1.9295	2.2783	2.4918	2.7270

Subject 1, Rare (20 trials)



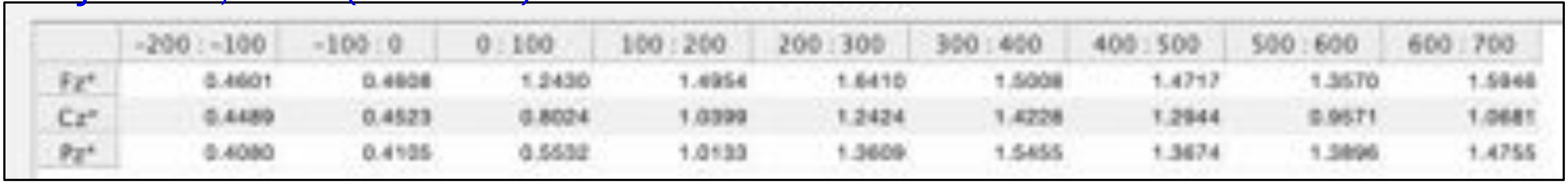
	-200 : -100	-100 : 0	0 : 100	100 : 200	200 : 300	300 : 400	400 : 500	500 : 600	600 : 700
Fz*	0.7299	0.7372	1.9673	2.7913	3.3332	4.0913	4.9746	5.3089	5.5988
Cz*	0.7092	0.7170	1.9390	2.9465	3.5511	4.5744	5.4623	5.4429	5.7883
Pz*	0.6832	0.6914	1.8541	2.9243	3.4108	4.5021	5.2465	5.3037	5.7192

Subject 12, Frequent (80 trials)



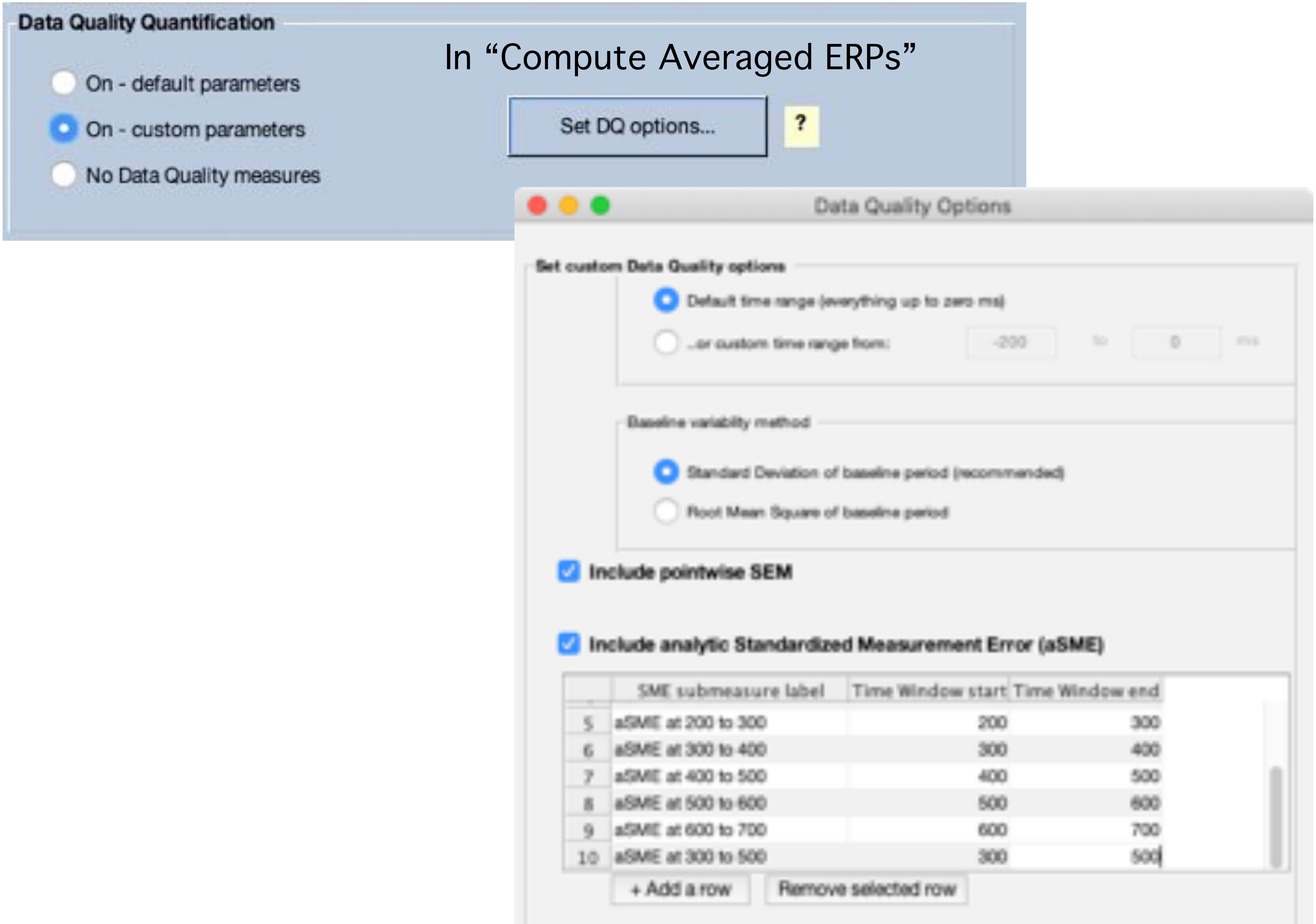
	-200 : -100	-100 : 0	0 : 100	100 : 200	200 : 300	300 : 400	400 : 500	500 : 600	600 : 700
Fz*	0.2350	0.2374	0.5337	0.6038	0.7318	0.8053	0.7813	0.8005	0.8207
Cz*	0.2522	0.2547	0.5815	0.6155	0.7281	0.8355	0.8446	0.8789	0.8532
Pz*	0.1911	0.1933	0.4553	0.5388	0.5718	0.5856	0.6734	0.6729	0.7181

Subject 12, Rare (20 trials)



	-200 : -100	-100 : 0	0 : 100	100 : 200	200 : 300	300 : 400	400 : 500	500 : 600	600 : 700
Fz*	0.4601	0.4608	1.2430	1.4954	1.6410	1.5008	1.4717	1.3570	1.5948
Cz*	0.4489	0.4523	0.8024	1.0399	1.2424	1.4228	1.2944	0.9571	1.0681
Pz*	0.4080	0.4105	0.5532	1.0133	1.3609	1.5455	1.3674	1.3896	1.4755

Custom Time Periods



In “Compute Averaged ERPs”

Data Quality Quantification

○ On - default parameters

● On - custom parameters

○ No Data Quality measures

Set custom Data Quality options

● Default time range (everything up to zero ms)

○ ...or custom time range from: -200 to 0 ms

Baseline variability method

● Standard Deviation of baseline period (recommended)

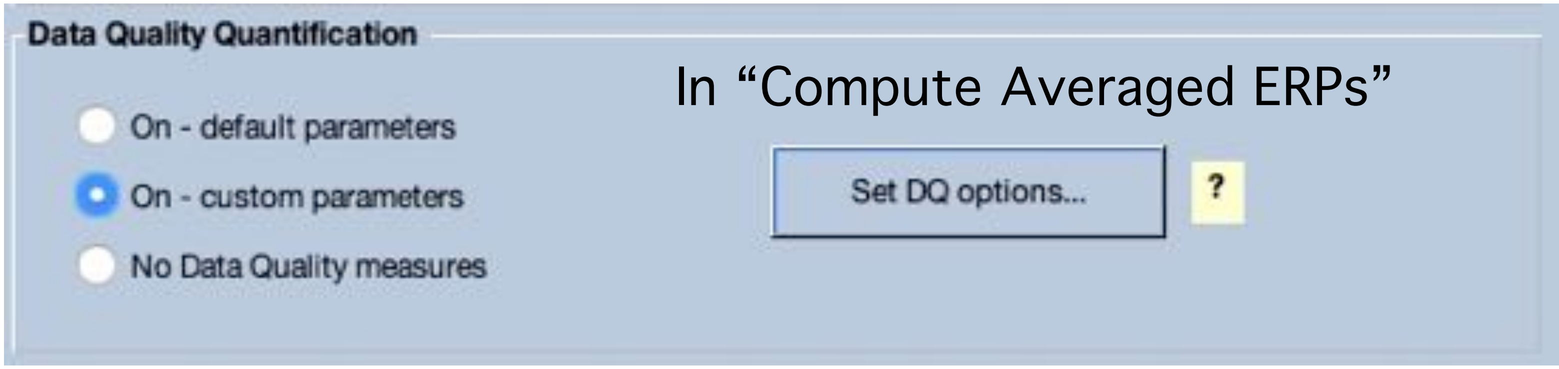
○ Root Mean Square of baseline period

☑ Include pointwise SEM

☑ Include analytic Standardized Measurement Error (aSME)

	SME submeasure label	Time Window start	Time Window end
5	aSME at 200 to 300	200	300
6	aSME at 300 to 400	300	400
7	aSME at 400 to 500	400	500
8	aSME at 500 to 600	500	600
9	aSME at 600 to 700	600	700
10	aSME at 300 to 500	300	500

Custom Time Periods



In “Compute Averaged ERPs”

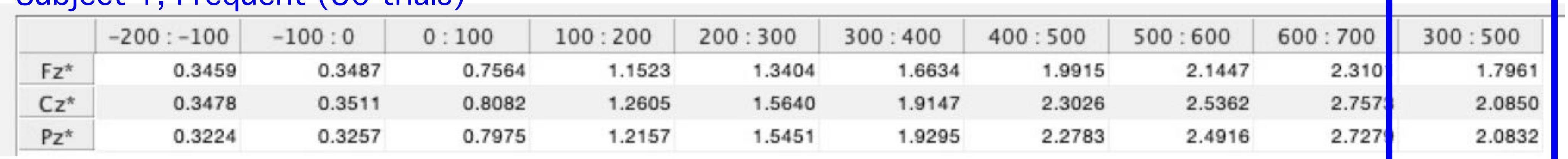
Data Quality Quantification

○ On - default parameters

● On - custom parameters

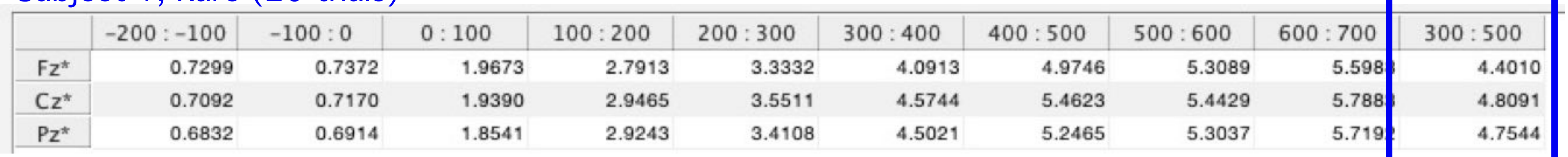
○ No Data Quality measures

Subject 1, Frequent (80 trials)



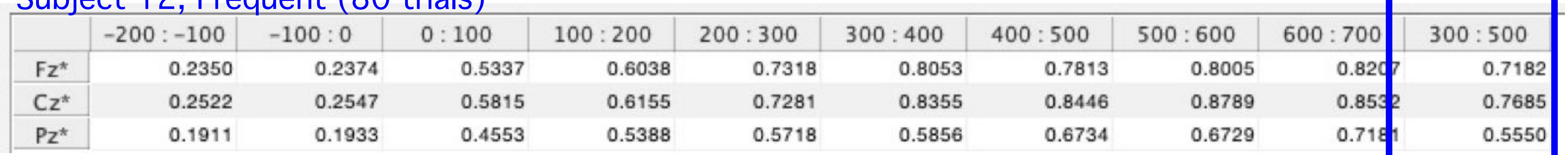
	-200 : -100	-100 : 0	0 : 100	100 : 200	200 : 300	300 : 400	400 : 500	500 : 600	600 : 700	300 : 500
Fz*	0.3459	0.3487	0.7564	1.1523	1.3404	1.6634	1.9915	2.1447	2.3101	1.7961
Cz*	0.3478	0.3511	0.8082	1.2605	1.5640	1.9147	2.3026	2.5362	2.7573	2.0850
Pz*	0.3224	0.3257	0.7975	1.2157	1.5451	1.9295	2.2783	2.4916	2.7279	2.0832

Subject 1, Rare (20 trials)



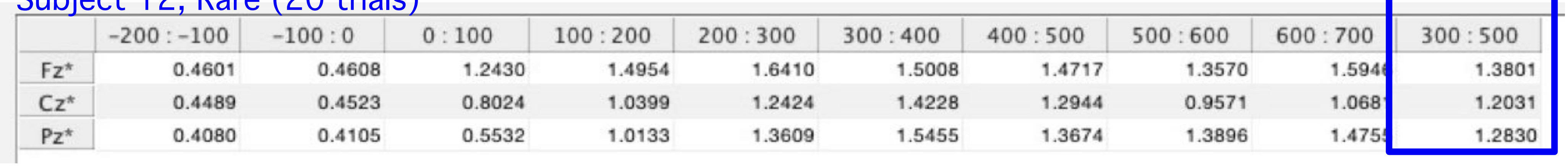
	-200 : -100	-100 : 0	0 : 100	100 : 200	200 : 300	300 : 400	400 : 500	500 : 600	600 : 700	300 : 500
Fz*	0.7299	0.7372	1.9673	2.7913	3.3332	4.0913	4.9746	5.3089	5.5983	4.4010
Cz*	0.7092	0.7170	1.9390	2.9465	3.5511	4.5744	5.4623	5.4429	5.7883	4.8091
Pz*	0.6832	0.6914	1.8541	2.9243	3.4108	4.5021	5.2465	5.3037	5.7192	4.7544

Subject 12, Frequent (80 trials)



	-200 : -100	-100 : 0	0 : 100	100 : 200	200 : 300	300 : 400	400 : 500	500 : 600	600 : 700	300 : 500
Fz*	0.2350	0.2374	0.5337	0.6038	0.7318	0.8053	0.7813	0.8005	0.8207	0.7182
Cz*	0.2522	0.2547	0.5815	0.6155	0.7281	0.8355	0.8446	0.8789	0.8532	0.7685
Pz*	0.1911	0.1933	0.4553	0.5388	0.5718	0.5856	0.6734	0.6729	0.7181	0.5550

Subject 12, Rare (20 trials)

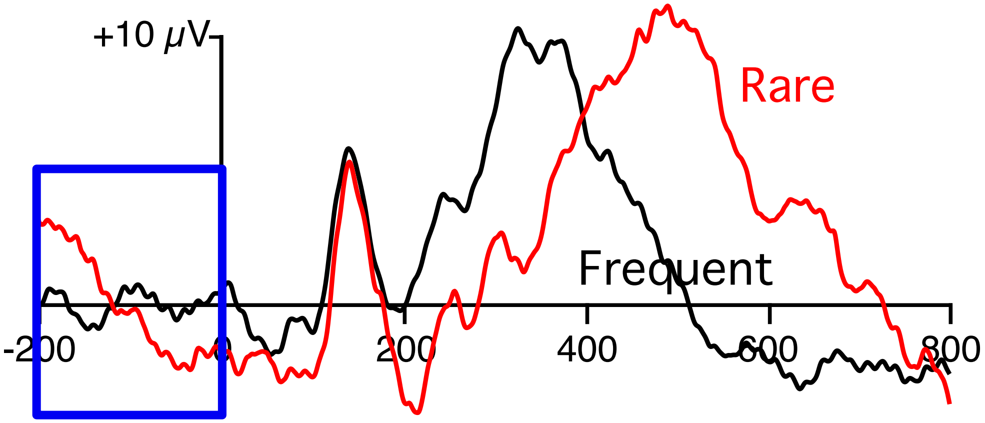


	-200 : -100	-100 : 0	0 : 100	100 : 200	200 : 300	300 : 400	400 : 500	500 : 600	600 : 700	300 : 500
Fz*	0.4601	0.4608	1.2430	1.4954	1.6410	1.5008	1.4717	1.3570	1.5946	1.3801
Cz*	0.4489	0.4523	0.8024	1.0399	1.2424	1.4228	1.2944	0.9571	1.0681	1.2031
Pz*	0.4080	0.4105	0.5532	1.0133	1.3609	1.5455	1.3674	1.3896	1.4755	1.2830

Baseline Noise in Averaged ERP

Algorithm: Take the sequence of voltages during the baseline period of the average and calculate the standard deviation

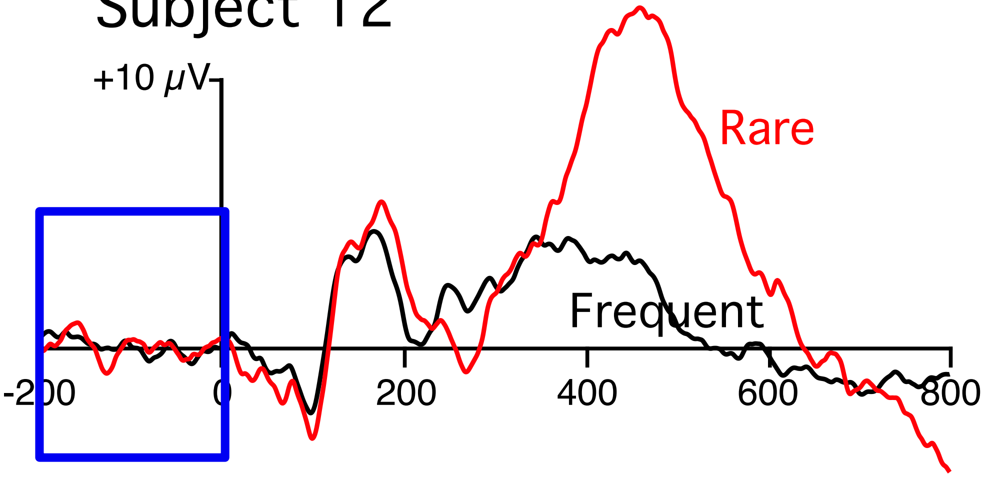
Subject 1



Frequent: SD = 0.4571 μV

Rare: SD = 1.9123 μV

Subject 12

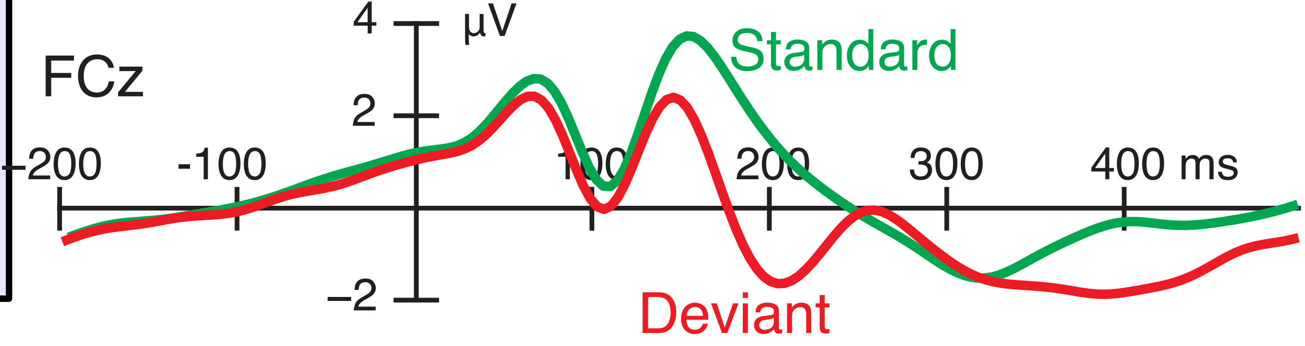


Frequent: SD = 0.3167 μV

Rare: SD = 0.4145 μV

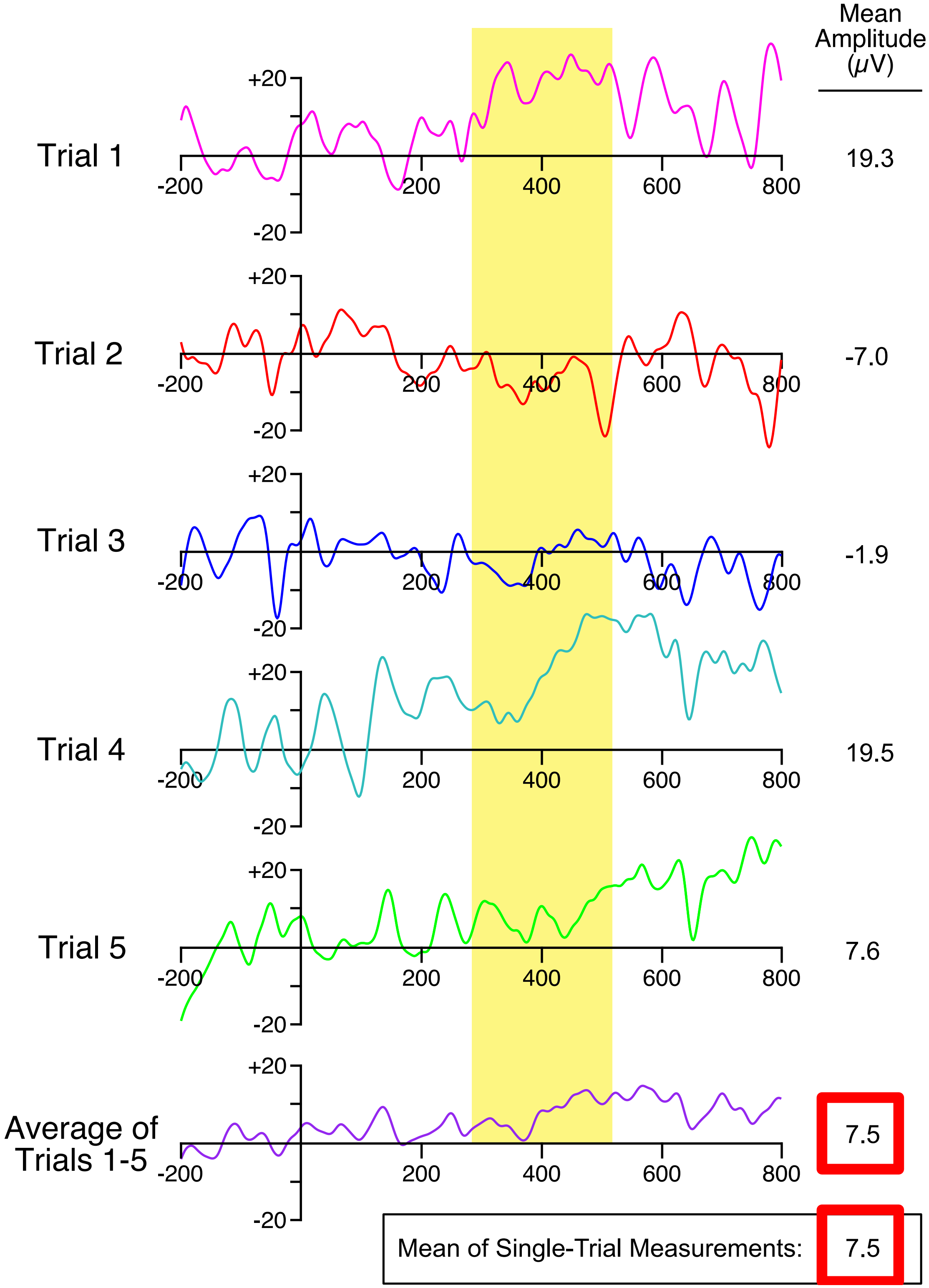
This assumes that all variation across the baseline period arises from noise, which is often untrue

Mismatch Negativity (MMN)

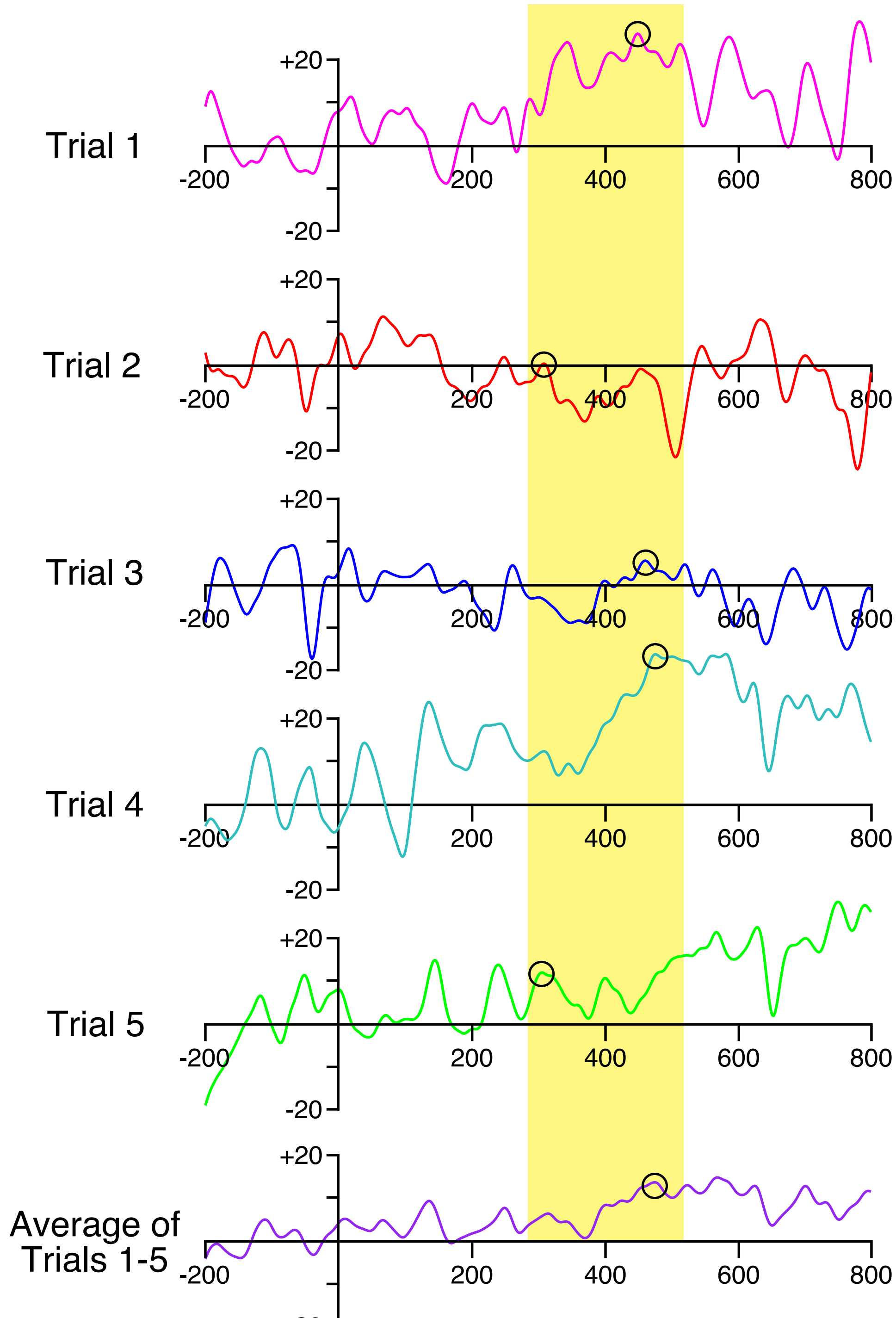


SME for Other Measures

- When we use the standard SEM equation (SD / $\sqrt{N}$) to calculate the SME, we call this the “analytic SME” (aSME)
- The analytic SME is appropriate when our score is the mean voltage within a time window (e.g., 300-500 ms)
- However, aSME is not appropriate for other measures (e.g., peak amplitude, peak latency, onset latency)
- In these cases, we need to use bootstrapping (“bootstrapped SME” or bSME)



Measuring the mean amplitudes on the single trials and then taking the average yields the same value as measuring the mean amplitude from the averaged ERP waveform.



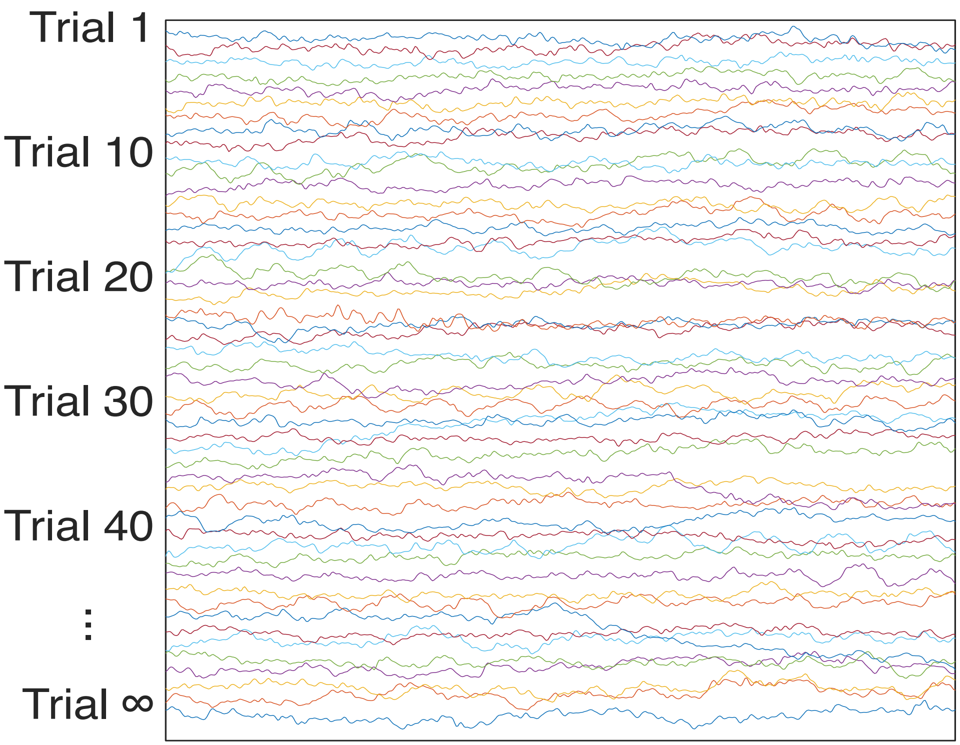
	Mean Amplitude (μV)	Peak Amplitude (μV)
Trial 1	19.3	26.1
Trial 2	-7.0	0.3
Trial 3	-1.9	5.6
Trial 4	19.5	35.1
Trial 5	7.6	11.9
Average of Trials 1-5	7.5	13.6
Mean of Single-Trial Measurements:	7.5	15.8

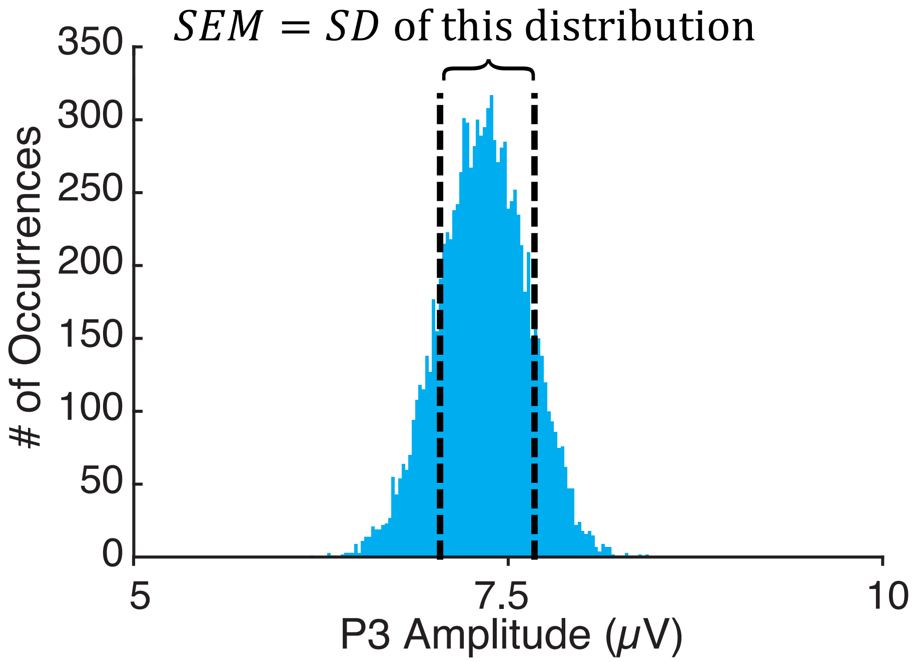
Measuring the peak amplitudes on the single trials and then taking the average does not yield the same value as measuring the peak amplitude from the averaged ERP waveform.

The SEM calculated from the single-trial peak amplitudes is the standard error of the mean of the single-trial peak amplitudes, not the standard error of the peak amplitude of the averaged waveform.

We can use bootstrapping to estimate the standard error of the peak amplitude.

Essence of Bootstrapping

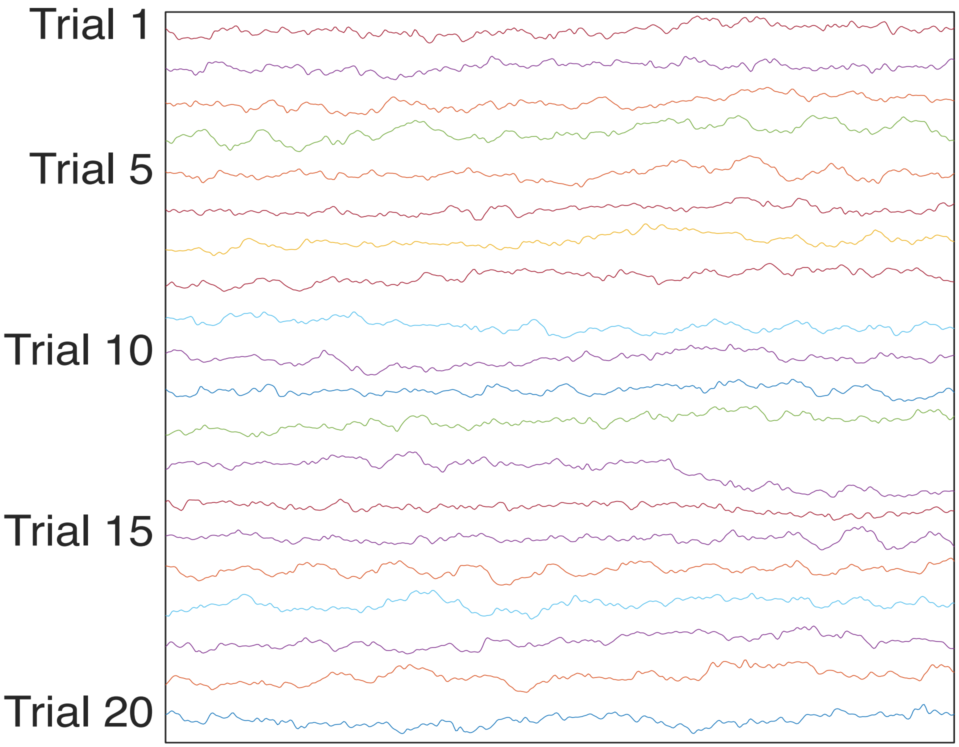


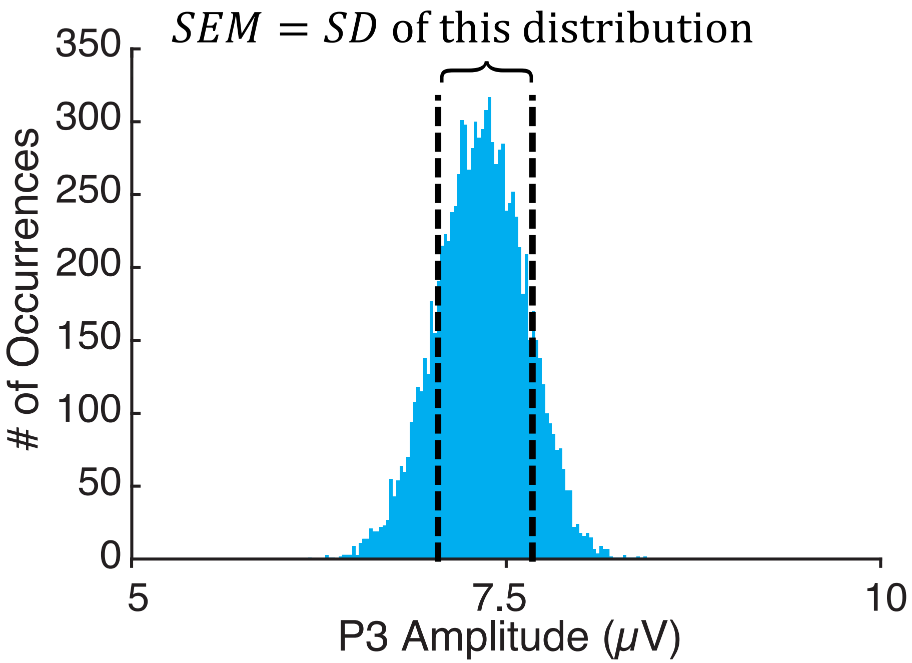


Distribution of means from 10,000 repetitions of the experiment

- In theory, we have an infinite population of single-trial EEG epochs for a given subject
- We could get the standard error of some measure (e.g., P3 peak latency) by running 10,000 sessions, each with a different random sample of trials
- For each session, we would make an averaged ERP waveform f and get the P3 peak latency score
- The standard error would be the SD of these scores

Essence of Bootstrapping



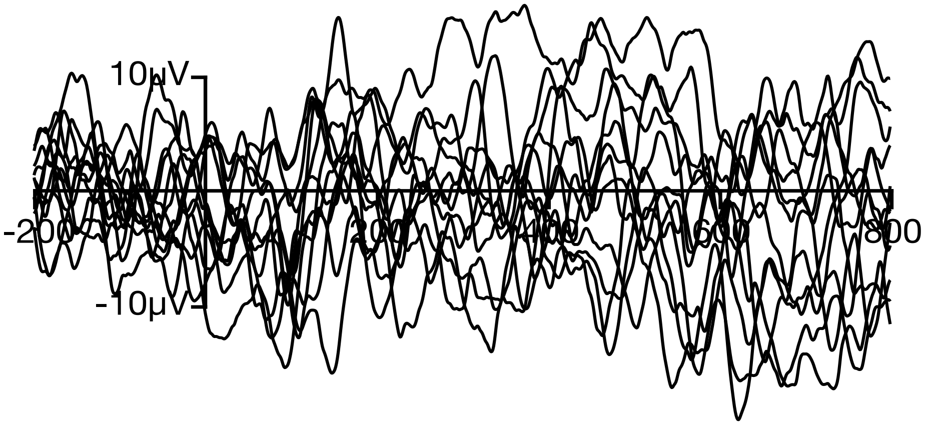


Distribution of means from 10,000 repetitions of the experiment

- Instead, we have a fixed number of trials (e.g., 20)
- We can simulate 10,000 sessions by sampling randomly with replacement from our 20 trials
  - E.g., Trials 1, 3, 3, 4, 5, 6, 9, 9, 9, 11, 13, 14, 14, 14, 14, 15, 15, 19, 20, 20
- For each simulated session, we would make an averaged ERP waveform f and get the P3 peak latency score
- The standard error would be the SD of these scores
- Bootstrapping sounds crazy, but it works and is widely used

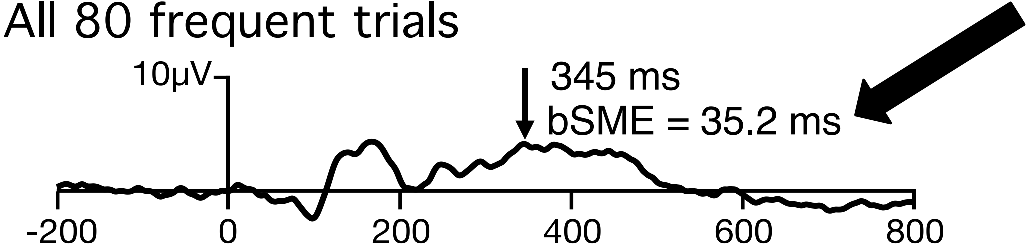
Bootstrap Example: P3 Peak Latency

Frequent (80 trials)

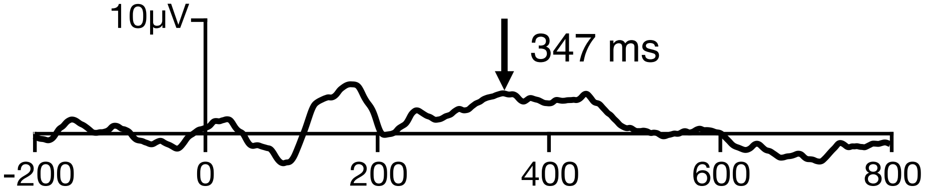


To compute bSME for peak latency, we make 10,000 bootstrapped averages, measure the peak latency from each average, and take the SD of these 10,000 latency values.

All 80 frequent trials

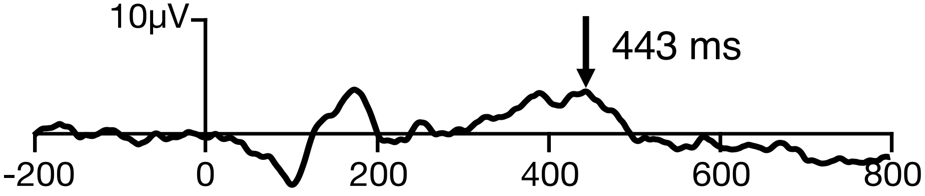


80 random frequent trials



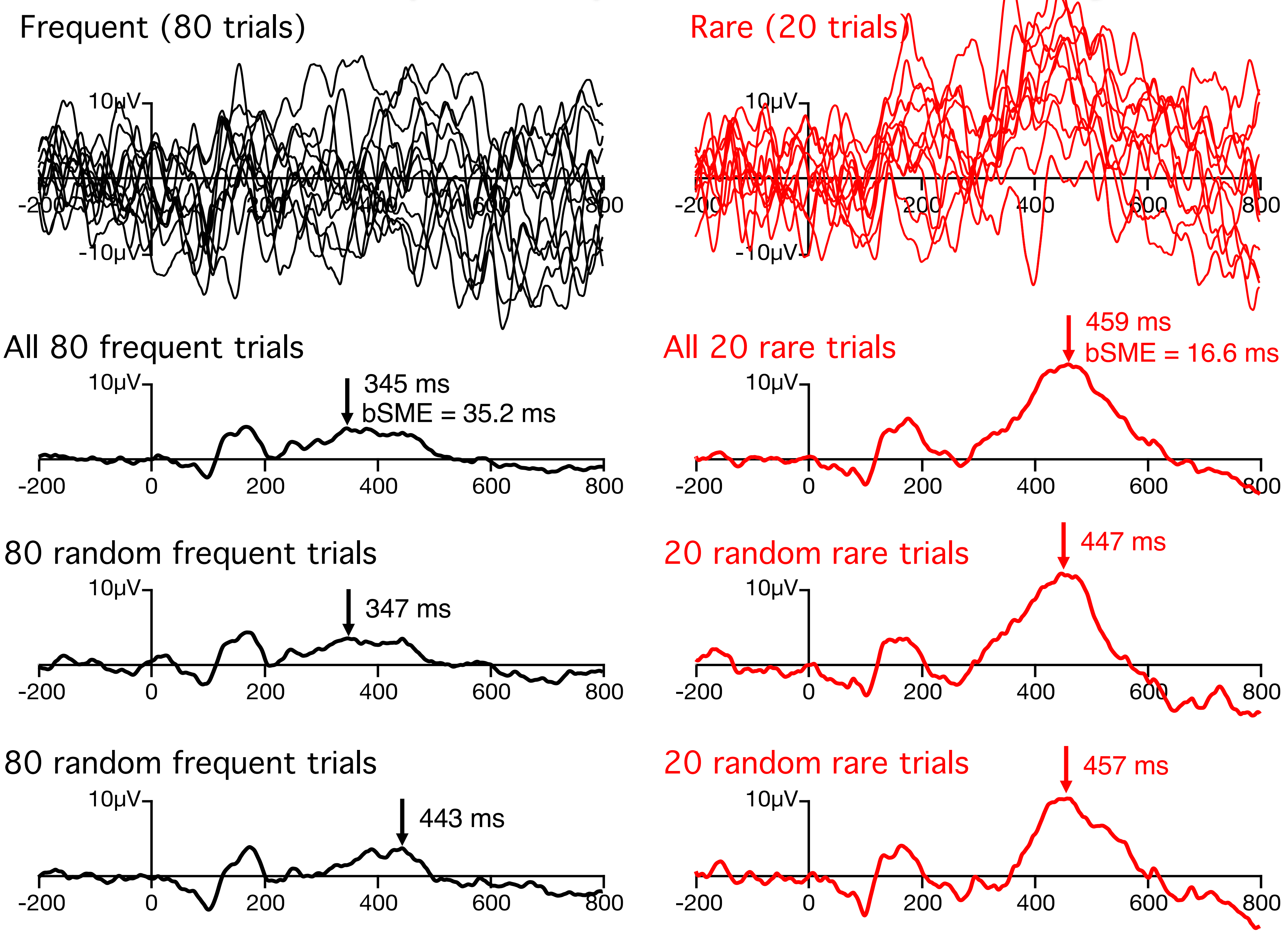
The 80 trials in this average were selected at random with replacement from the 80 available trials

80 random frequent trials



This average is from a new set of 80 trials selected at random with replacement from the 80 available trials

Bootstrap Example: P3 Peak Latency



Frequent (80 trials)

Rare (20 trials)

All 80 frequent trials

345 ms
bSME = 35.2 ms

All 20 rare trials

459 ms
bSME = 16.6 ms

80 random frequent trials

347 ms

20 random rare trials

447 ms

80 random frequent trials

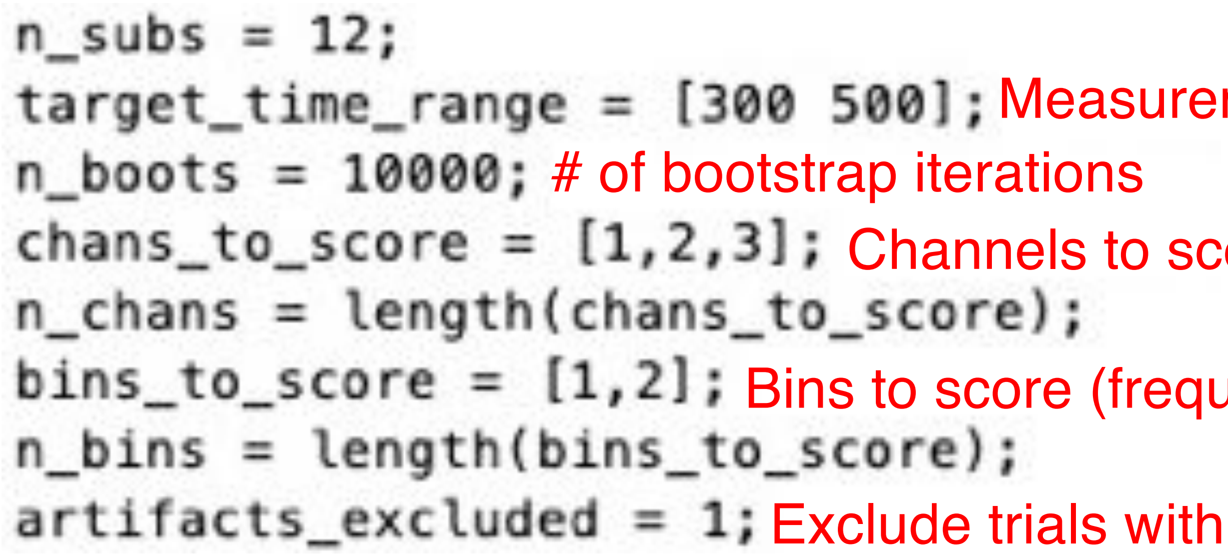
443 ms

20 random rare trials

457 ms

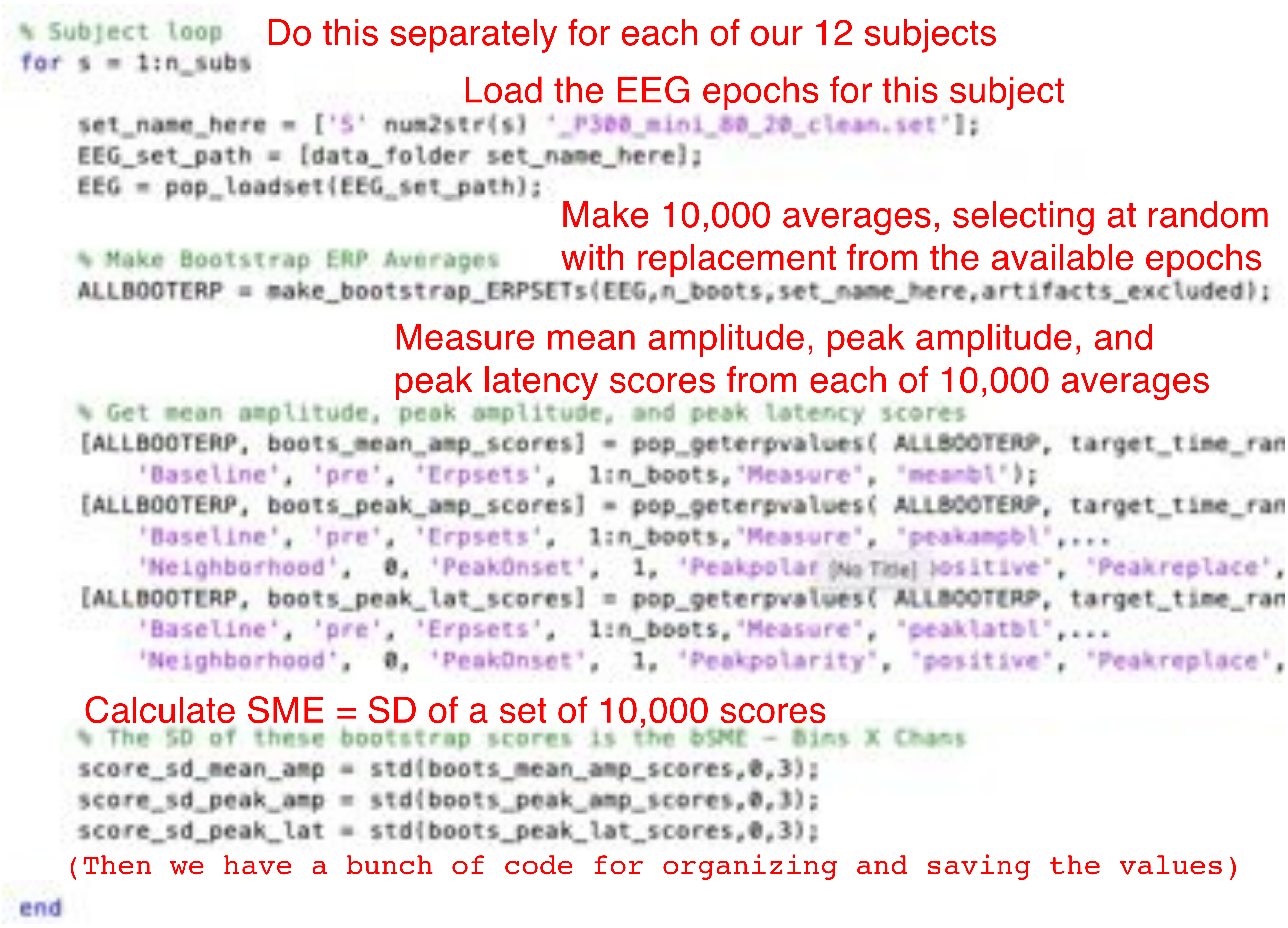
SME_demo_3_bSME_peak_amp_peak_latency.m

(Begins with some housekeeping)



```
n_subs = 12;
target_time_range = [300 500]; Measurement Window
n_boots = 10000; # of bootstrap iterations
chans_to_score = [1,2,3]; Channels to score (Fz, Cz, Pz)
n_chans = length(chans_to_score);
bins_to_score = [1,2]; Bins to score (frequent, rare)
n_bins = length(bins_to_score);
artifacts_excluded = 1; Exclude trials with artifacts
```

SME_demo_3_bSME_peak_amp_peak_latency.m



```
% Subject loop
for s = 1:n_subs
```

Do this separately for each of our 12 subjects

Load the EEG epochs for this subject

```
    set_name_here = ['S' num2str(s) '_P300_mini_80_20_clean.set'];
    EEG_set_path = [data_folder set_name_here];
    EEG = pop_loadset(EEG_set_path);
```

Make 10,000 averages, selecting at random with replacement from the available epochs

```
    % Make Bootstrap ERP Averages
    ALLBOOTERP = make_bootstrap_ERPSETs(EEG,n_boots,set_name_here,artifacts_excluded);
```

Measure mean amplitude, peak amplitude, and peak latency scores from each of 10,000 averages

```
    % Get mean amplitude, peak amplitude, and peak latency scores
    [ALLBOOTERP, boots_mean_amp_scores] = pop_geterpvalues( ALLBOOTERP, target_time_ran
        'Baseline', 'pre', 'Erpsets',  1:n_boots,'Measure', 'meanbl');
    [ALLBOOTERP, boots_peak_amp_scores] = pop_geterpvalues( ALLBOOTERP, target_time_ran
        'Baseline', 'pre', 'Erpsets',  1:n_boots,'Measure', 'peakampbl',...
        'Neighborhood',  0, 'PeakOnset',  1, 'Peakpolar (No Time) ositive', 'Peakreplace',
    [ALLBOOTERP, boots_peak_lat_scores] = pop_geterpvalues( ALLBOOTERP, target_time_ran
        'Baseline', 'pre', 'Erpsets',  1:n_boots,'Measure', 'peaklatbl',...
        'Neighborhood',  0, 'PeakOnset',  1, 'Peakpolarity', 'positive', 'Peakreplace',
```

Calculate SME = SD of a set of 10,000 scores

```
    % The SD of these bootstrap scores is the bSME - Bins X Chans
    score_sd_mean_amp = std(boots_mean_amp_scores,0,3);
    score_sd_peak_amp = std(boots_peak_amp_scores,0,3);
    score_sd_peak_lat = std(boots_peak_lat_scores,0,3);
```

(Then we have a bunch of code for organizing and saving the values)

```
end
```

What’s a “Good” SME Value?

- “It depends”
- Relative differences between subjects or between channels

ERP CORE Experiments (http://erpinfo.org/erp-core)

P3 N400 MMN N2pc N170 ERN+LRP

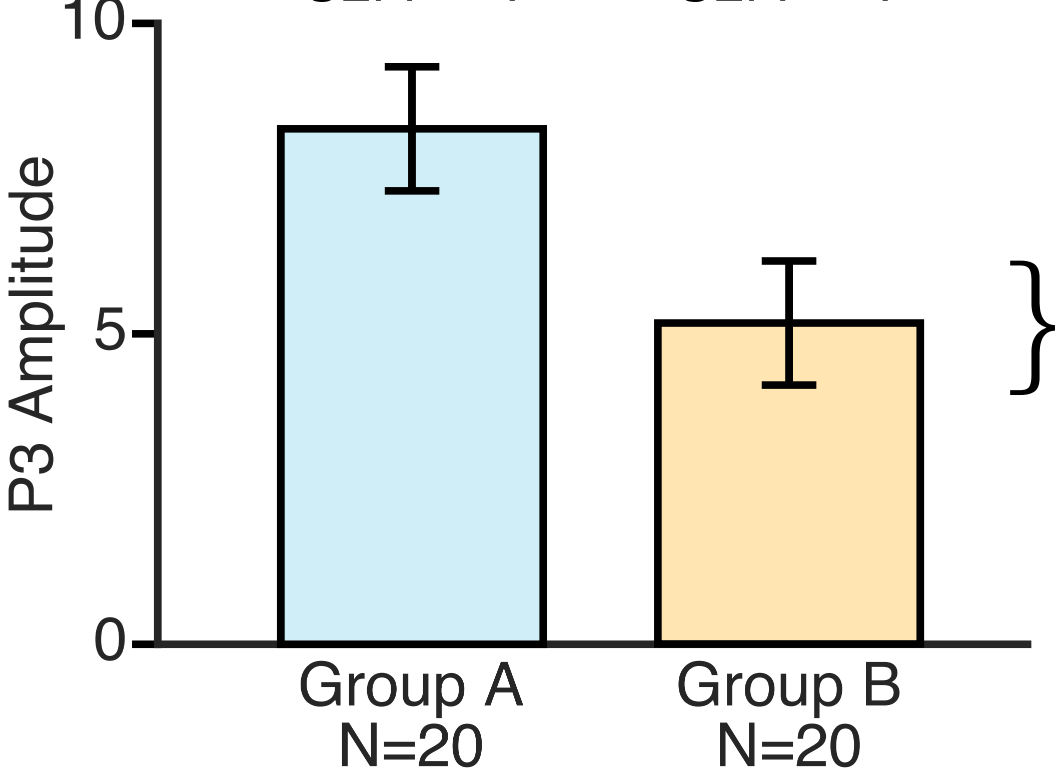
Relating SME to Effect Size & Statistical Power

Mean = 8
SD = 4.47
SEM = 1

Mean = 5
SD = 4.47
SEM = 1

Effect Size (Cohen's d) =
(8 – 5) / 4.47 = 0.67

Power = 0.54



How much of the variability across subjects reflects measurement error?

How much reflects true differences among subjects?

How much bigger would our effect be if we reduced noise in the EEG by 50%?

How would our power change if we reduced the number of trials by 30%

You can answer these questions by computing SME for each subject and combining those values into RMS(SME)

How Could You Use SME?

- Within a lab, SME could be used to…
  - Find subjects who should be excluded and channels that should be interpolated
  - Rigorously test whether new recording and analysis procedures actually improve data quality
  - Choose optimal parameters for signal processing
- If every paper reported RMS(SME), we could…
  - Have objective evidence that the data from a given study are unusually noisy, making the results less believable
  - Quantitatively assess how data quality varies among different experimental paradigms and different subject populations
  - Determine which recording and analysis procedures lead to the most reliable scores

My Dream

In 10 years, every new ERP paper reports a measure of data quality