Supplementary Materials for

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

Steven J. Luck, Andrew X. Stewart, Aaron M. Simmons, and Mijke Rhemtulla

Contents

S1. Psychometric Reliability as a Measure of Data Quality

S2. Plotting the SEM at Each Time Point in an Averaged ERP Waveform

S3. Description of Example Oddball Study

S4. Overview of Bootstrapping

S5. Example SME Values

S6. SME and the Signal-to-Noise Ratio

S7. The Assumption of Independence

S8. A Potential Criterion for Defining Extreme SME Values

S9. References for Supplementary Materials

S1. Psychometric Reliability as a Measure of Data Quality

Some studies have quantified ERP data quality using correlation-based measures of reliability (e.g., Cronbach's alpha, split-half reliability). We will call these *psychometric measures of reliability* to distinguish them from metrics of reliability used in other fields (see Brandmaier et al., 2018). In this section, we describe three limitations of these measures in the context of ERP data quality.

One key limitation of psychometric reliability (as it is usually computed) is that it provides a single value for an entire group rather than providing a separate index of data quality for each participant. Thus, it does not meet one of the three key criteria for metrics of data quality described in Section 1.1 of the main text.

A second limitation of psychometric reliability is that it is influenced by factors beyond the quality of an individual participant's data. As discussed in Section 6.2 of the main text, psychometric reliability is typically defined as the proportion of total variance across participants that is due to true score variance (and therefore not due to measurement error). Consequently, if the true score variance decreases but the measurement error remains the same, psychometric reliability decreases (i.e., a lower proportion of the total variance is the result of true score variance). For example, if you sample from a more homogeneous population, this will decrease the true score variance, and this will in turn decrease the reliability. Thus, psychometric measures of reliability reflect the properties of the sample of participants and not just the quality of the single-participant data.

A third limitation of psychometric reliability is that its relationship to statistical power is not straightforward for comparisons of group or condition means (e.g., in *t* tests and ANOVAs). For example, Hedge et al. (2018) showed that many widely used behavioral paradigms yield low psychometric reliability (because of low true score variance) but also produce high statistical power for within-subjects effects. Conversely, Thigpen et al. (2017) found that the number of trials needed to obtain acceptable reliability in an ERP paradigm was far lower than the number of trials needed to obtain satisfactory effect sizes for within-subjects comparisons. Together, these studies indicate that when group or condition means are being compared, low reliability may be accompanied by large effect sizes, and high reliability may be accompanied by small effect sizes. By contrast, the SME can be directly related to effect sizes in these cases (see Section 6 of the main text). Thus, although psychometric reliability plays a straightforward and important role in correlational studies of individual differences, it is not well suited as a universal metric of data quality for averaged ERPs.

S2. Plotting the SEM at Each Time Point in an Averaged ERP Waveform

The $\hat{SEM}$ is sometimes shown for each time point in a grand average ERP waveform. This is illustrated in Figure S1A, which shows the data from the oddball trials in a simple oddball experiment with 80 standard trials and 20 oddball trials for each participant (see Section S3 for a description of the experimental design). The averaged ERP waveforms from each individual participant are shown at the top, and the grand average waveforms across participants are shown at the bottom. The grand average at a given point in time is simply the mean of the single-participant averaged ERP waveforms at that point in time, and the $\hat{SEM}$ for that point in time can be computed with Equation 1: we simply calculate the $\hat{SD}$ of the single-participant voltages at that point in time and divide by the square root of the number of participants. The shading

around each grand average ERP waveform represents this $\hat{SEM}$ value. We call this the *single-point group SEM*.

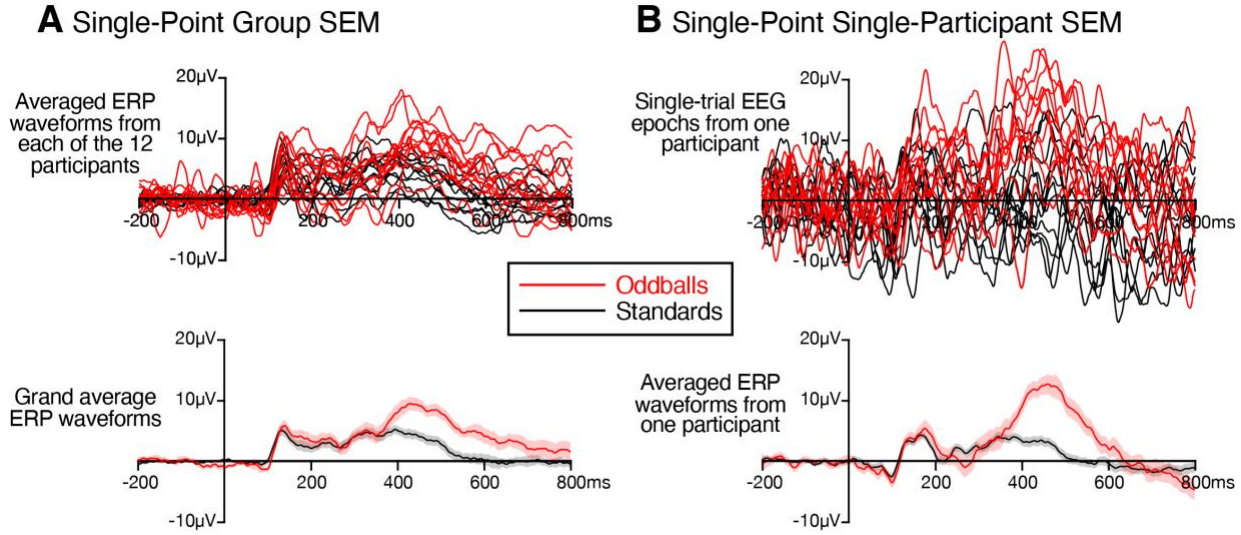

Figure S1. Example of two ways of conceptualizing the standard error of the mean (SEM) for ERP waveforms in an oddball experiment. (A) Single-point group SEM. The averaged ERP waveforms from the standard and oddball trials are shown for each of the 12 participants at the top, and the grand averages of these ERP waveforms are shown at the bottom. The shaded areas around the grand averages show the range of voltages that fall within ±1 $\hat{SEM}$ of the mean at each time point. For example, the $\hat{SEM}$ at 500 ms was estimated by measuring the voltage at 500 ms in each of the single-participant oddball ERPs and applying Equation 1 to these values. (B) Single-point single-participant SEM. A subset of the single-trial EEG epochs from a single participant are shown at the top, and the averaged ERP waveforms for this participant are shown at the bottom. The shaded region shows ±1 $\hat{SEM}$, where the $\hat{SEM}$ is estimated by applying Equation 1 to the single-trial voltages at a given time point. Note that the $\hat{SEM}s$ are much smaller in (B) than in (A), and this is partly because the number of trials (*N* in the denominator of Equation 1 for the data in B) is much larger than the number of participants (*N* in the denominator of Equation 1 for the data in A).

The single-point group SEM can be useful as a visualization tool, but it is not a good metric of data quality because it is influenced by both variability due to noise in the single-participant ERP waveforms (because noisier single-participant data will lead to more variability among participants in their averaged ERP waveforms). In addition, it does not meet the first two criteria described in Section 1.1 of the main manuscript: it tells us about an entire group rather than an individual participant, and it tells us about individual time points rather than our dependent variable (which might be the time-window mean amplitude or the peak latency between 300 and 500 ms).

We can solve the first of these problems by computing the $\hat{SEM}$ at each time point in a single-participant averaged ERP waveform rather than in a grand average. This is illustrated in Figure S1B, which shows the single-trial EEG epochs from a single participant (top), along with the average across trials (bottom). To obtain the $\hat{SEM}$ at a given time point using Equation 1, we simply find the voltage at that time point on each trial, calculate the $\hat{SD}$ across trials, and divide by the square root of the number of trials. We call this the *single-point single-participant SEM*.

Unlike the single-point group SEM shown in Figure S1A, the single-point single-participant SEM shown in Figure S1B reflects trial-to-trial variability with no influence of subject-to-subject variability, and it could therefore be used as a metric of data quality. However, it does not satisfy the second of our three criteria, because it does not reflect the expected error for the actual amplitude or latency score that we will put into our statistical analyses (unless we happen to be using the voltage at a single time point as our dependent variable, which would be quite unusual). For example, if we quantified P3 amplitude as the time-window mean amplitude from 300-500 ms, the error in this score is not the average of the single-point SEM values between 300 and 500 ms. In particular, high-frequency noise would have a large impact on the single-point SEM values but would have relatively little impact on the time-window mean amplitude from 300-500 ms (see Figure 1B).

S3. Description of Example Oddball Study

Some of the examples in this paper use data from a previously published oddball experiment (Kappenman & Luck, 2010). In this experiment, 12 neurotypical adult participants saw a sequence of alphanumeric characters and pressed one of two buttons to indicate whether a given stimulus was a letter or a digit. One of these two categories occurred frequently (*standards*, 80% of stimuli) and the other occurred infrequently (*oddballs*, 20% of stimuli). This experiment included an unusually large number of trials, and to simulate a more typical number of trials, we used only the first 20 artifact-free oddballs and the first 80 artifact-free standards per participant. Note that Cohen and Polich (1997) concluded that 20 trials is sufficient to obtain a stable measure of P3 amplitude (but see Boudewyn, Luck, Farrens, & Kappenman, 2018). We also provide data in which all the trials are included at https://doi.org/10.18115/D58G91, and trials with artifacts are marked so that they can be explicitly excluded during averaging (and during calculation of the $\widehat{SME}$). This experiment also manipulated the electrode impedances to influence the data quality. Here, we show only the data from the low-impedance electrode sites.

S4. Overview of Bootstrapping

Bootstrapping is a simple but powerful approach for estimating standard errors, but it seems a little mystical at first. Here, we provide a simple overview. We begin by describing how it would work in the case of reaction time (RT), and then we describe how to extend it to ERPs.

A Reaction Time Example

Imagine that we are conducting an experiment in which we obtained a single-trial RT value on each of 20 trials for a given participant, and we want to know the standard error of the mean (SEM) for this participant. We could just use the analytic SEM formula (Equation 1 in the main text), but we could also use bootstrapping. And once we understand how bootstrapping works for this simple case, we can extend it to other scores for which Equation 1 does not apply.

First, however, let's review what the SEM actually *means*, which is most naturally expressed in terms of the empirical approach to estimating the SEM. Figure S2-A illustrates the empirical approach, in which we assume that there is an underlying infinite distribution of RTs for a given participant. Every time we run the experiment with this participant, we are sampling 20 trials from this distribution and computing the mean of those 20 trials. We would repeat the

same experiment 10,000 times[1] with this participant, each time getting 20 single-trial RTs and computing the mean RT of these 20 trials. We would save each of these mean RTs so that we can construct the *sampling distribution* of the mean RT, which is shown as the histogram at the top of Figure S2-A. The $\widehat{SEM}$ is defined as the $\widehat{SD}$ of the 10,000 mean RT values. More generally, the standard error of a measure is the SD of the sampling distribution for that measure.

It is not practical to repeat an experiment over and over again with a given participant, and bootstrapping uses a clever trick that allows us to simulate repeating the experiment multiple times. The trick is to use the observed set of 20 RTs as an approximation of the infinite distribution of RTs. To simulate a single experiment, we would randomly sample 20 single-trial RTs *with replacement* from that set of 20 observed single-trial RTs. We would repeat this procedure many times (e.g., 10,000 times) to simulate many experiments, saving the mean RT for each simulated experiment. In other words, whereas the empirical approach involves randomly sampling 20 trials from a hypothetical infinite distribution of single-trial RTs for each of 10,000 actual experiments, we instead randomly sample 20 trials from the set of 20 trials we actually have from this participant and then compute the mean RT for each simulated experiment.

To avoid getting the same set of 20 trials (and therefore the same sample mean) on each of the iterations, we must sample *with replacement* from our set of 20 observed single-trial RTs. On any given iteration, some of the 20 trials would not get selected, and others would selected twice or even more than twice. However, if we believe that we've selected 20 trials at random from an infinite set of RTs when we collected our actual data[2], then these 20 trials will reflect the properties of the infinite set of RTs. As a result, if we sample 20 trials *with replacement* from the observed set of 20 RTs, the result will be an approximation of what we would get if we sampled 20 brand-new trials from the infinite population of RTs. It's only an approximation, but in practice it works well as long as we have a reasonable number of observed trials. A minimum of 8 trials is recommended to avoid identical random samples on different iterations (Chernick, 2011).

As illustrated in Figure S2-B, the process of estimating the SEM with bootstrapping is exactly the same as the empirical approach, except that each of our 10,000 iterations involves sampling 20 trials with replacement from the 20 observed single-trial RTs (instead of sampling from the hypothetical infinite population of RTs). That is, for each iteration, we sample 20 trials from our 20 observed single-trial RTs, get the mean of those 20 trials, and save that mean value to get the sampling distribution. This gives us a sampling distribution composed of 10,000 mean RTs. The $\widehat{SD}$ of these 10,000 mean RTs is then an estimate of the standard error of the single-participant mean RT.

You might think that you could just sample 10 of the 20 trials on each of the 10,000 iterations. However, if we took the mean of 10 trials instead of the mean of 20 trials, the set of 10,000 means would be more variable than the mean of 20 trials. Thus, the number of trials

---

[1] We are using 10,000 replications in our examples, but there is nothing special about this particular number (although it is commonly used in bootstrapping).

[2] When trials are sampled over time from an individual participant, the samples are not independent, and there may be sequential dependencies in the samples (e.g., as a result of learning or fatigue). Both the bootstrapped SEM and the analytic SEM assume that the samples are independent, so the procedure described here is only approximate. This is described in more detail in the Section S6.

selected on each iteration must be exactly the same as the number of trials in our observed set of single-trial values.

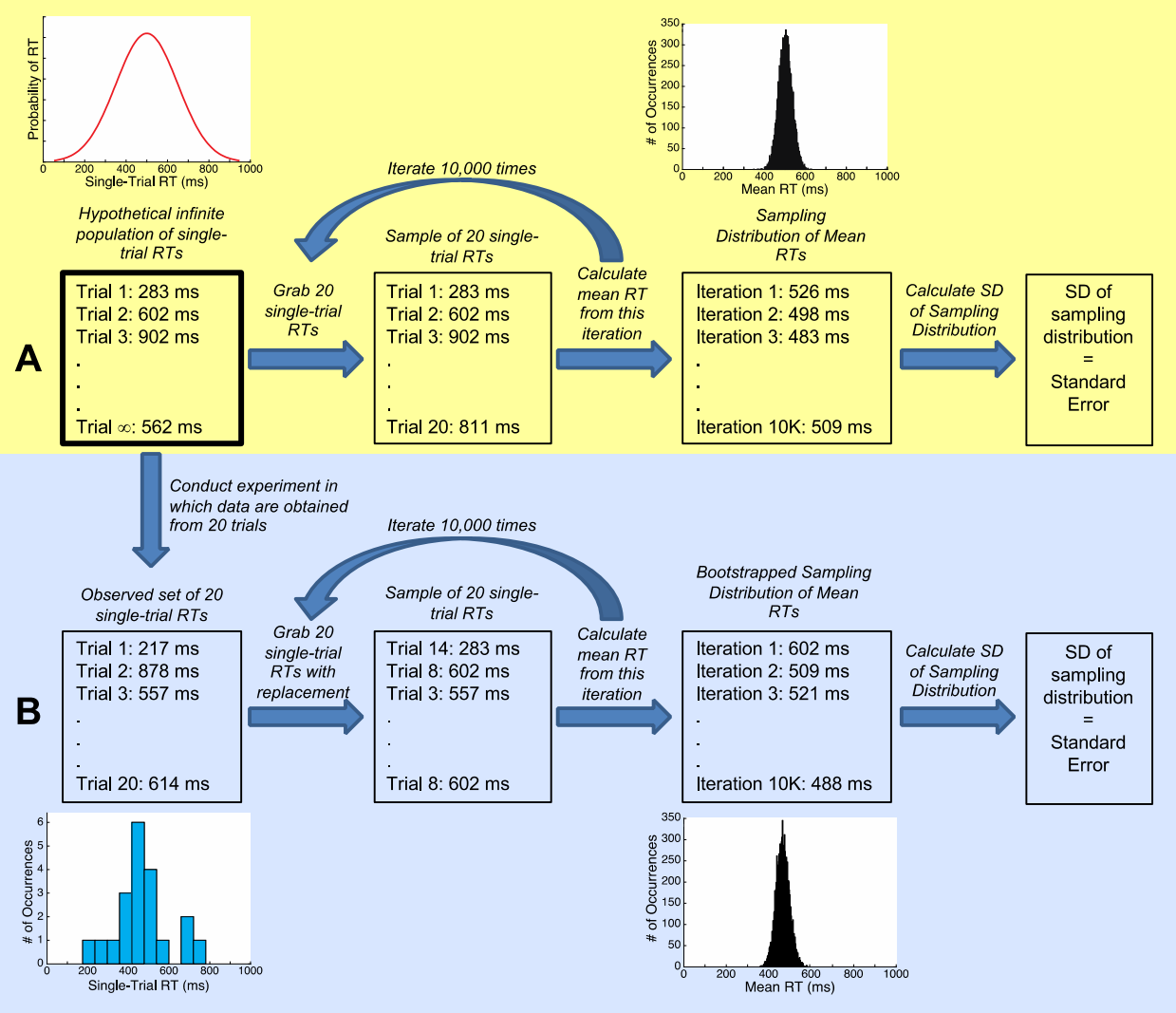

Figure S2. Example of using the empirical approach (A) and bootstrapping (B) to estimating the standard error of the mean of 20 reaction times.

Although this may seem like an odd approach, it has been very well established and is widely used in many areas of science (Boos, 2003; Efron & Tibshirani, 1994). Moreover, when applied to the mean across trials it provides an estimate of the same standard error that we would get by using Equation 1. That is, if we use bootstrapping to get 10,000 mean RTs, and we take the $\widehat{SD}$ of these 10,000 mean RTs, we will get approximately the same value (on average) that we would get by taking the 20 observed single-trial RTs and using Equation 1 to estimate the standard error of the mean RT. These are just two different ways of estimating the standard error from the same data. Indeed, bootstrapping is more directly related to actual meaning of the standard error, because it is the $\widehat{SD}$ of a sampling distribution.

Whereas the analytic SEM formula (Equation 1 in the main text) can be applied only to means, bootstrapping can provide an estimate of the standard error of other measures. For

example, bootstrapping could be used to estimate the standard error of the median. To do this, you would sample 20 RTs with replacement from the observed set of 20 RTs and take the median of these 20 RTs (instead of the mean). You would then repeat this 10,000 times, calculating the median of a new random sample of 20 RTs on each iteration, and then compute the $\widehat{SD}$ of these 10,000 median RTs. This would give you the standard error of the median[3]. You can use bootstrapping to estimate the standard error of virtually any value that is obtained by combining data from multiple trials.

How Bootstrapping Works with ERPs

Bootstrapping can also be used to estimate the standard error of estimate for virtually any ERP amplitude or latency score. First, consider how it would work if we used the time-window mean amplitude from 300-500 ms to score P3 amplitude for the oddballs in an experiment with 20 oddball trials. As shown in Figure S3-A, the empirical approach to estimating the SEM involves getting the single-trial EEG epochs for 20 oddball trials from an infinite population of single trials, averaging them together to get an averaged ERP waveform, computing the time-window mean amplitude from 300-500 ms from this averaged ERP waveform, and then saving this score. We would then repeat this 10,000 times to get 10,000 P3 amplitude scores. The $\widehat{SD}$ of these 10,000 scores would be the standard error of the time-window mean P3 amplitude.

Because we do not actually have access to an infinite population of single trials, we can sample 20 trials with replacement from the set of 20 trials that we actually recorded, as illustrated in Figure S3-B. That is, we take our 20 single-trial EEG epochs, randomly sample 20 of them *with replacement* (so that we don't just get the same 20 trials every time), make an averaged ERP waveform from these 20 trials, and calculate the time-window mean voltage from 300-500 ms in this waveform. We then repeat this 10,000 times, taking a new random set of 20 single-trial EEG epochs to create the averaged ERP waveform for each iteration, and then compute the $\widehat{SD}$ of the scores obtained from each of the 10,000 averaged ERP waveforms. This gives us the standard error of the time-window mean amplitude.

For the time-window mean amplitude score, we could instead use Equation 1 to estimate the SME, which would be far simpler and faster than bootstrapping. In practice, bootstrapping would be used for other types of scores.

For example, we could score the P3 peak latency (e.g., the latency of the maximum voltage between 200 and 800 ms) from the averaged ERP waveform in each iteration of the bootstrap procedure. The end result would then be the standard error of the peak latency. Moreover, bootstrapping can be applied to scores that require multiple processing steps between averaging across trials and obtaining a score. For example, if filtering or averaging across channels is performed after averaging but before scoring, Equation 1 cannot be used to estimate the SME of the time-window mean amplitude, and bootstrapping would be necessary. Another example is scores that must be obtained from combinations of waveforms (e.g., the onset time of a component measured from a difference wave).

[3] There is also an analytic solution for estimating the standard error of the median from a set of single-trial values. However, analytic solutions are not available for all measures, and we know of no analytic equation for estimating the standard error of common ERP measures such as peak amplitude and peak latency.

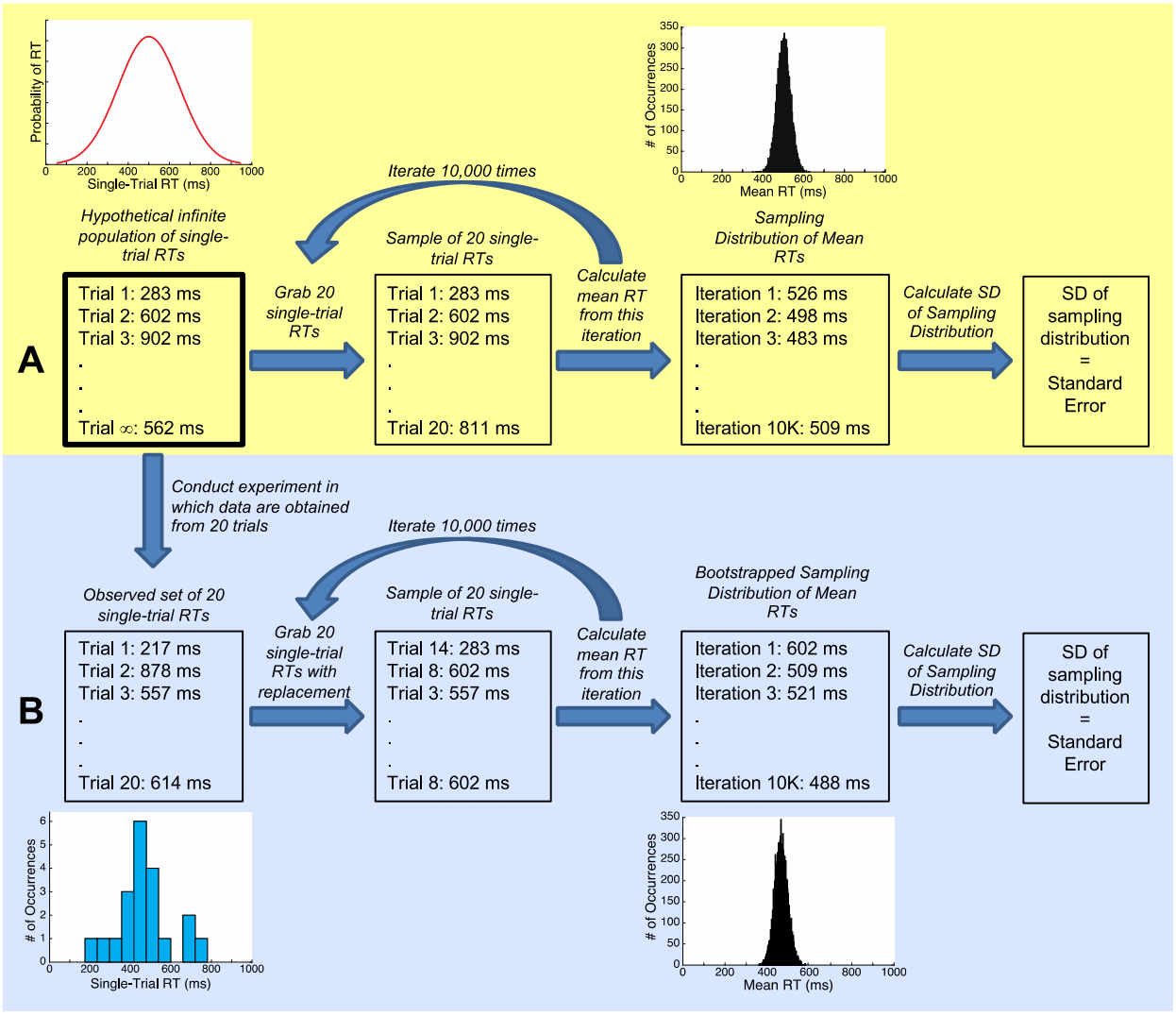

Figure S3. Example of using the empirical approach (A) and bootstrapping (B) to estimating the standard error of the mean of the P3 amplitude score from an ERP waveform constructed by averaging together 20 single-trial EEG epochs.

S5. Example SME Values

To provide concrete examples, we obtained $\hat{SME}$ values for the amplitude and latency of the P3 wave in each participant in the oddball study described in Section S3. Specifically, we obtained both the analytic and bootstrapped $\hat{SME}$ values for the time-window mean amplitude score, and we obtained bootstrapped $\hat{SME}$ values for peak amplitude and peak latency scores. The time window was 300–500 ms for all scores. We used 10,000 iterations for each bootstrapped $\hat{SME}$ calculation.

The single-participant scores and $\hat{SME}$ values are provided in Table 1 of the main paper, along with group means, group standard deviations, and RMS($\hat{SME}$) values. We would like to stress that these values are from a small number of participants in a single experimental paradigm, and they are not intended to serve as typical or normative values. However, they are

useful in illustrating some important properties of the SME. For example, Table 1 shows that the analytic and bootstrapped SME values are nearly identical to each other for the time-window mean amplitude.

The data in Table 1 also illustrate how RMS($\widehat{SME}$) values can be compared to sample $\widehat{SD}$ values when assessing the aggregate data quality in an experiment. For the standards, the group mean of the time-window mean amplitude scores was 4.38 μV ($\widehat{SD}_{Total}$ = 2.44) with an RMS($\widehat{SME}$) of 1.13 μV for the standards. For the oddballs, the group mean was 7.18 μV ($\widehat{SD}_{Total}$ = 3.78) with an RMS($\widehat{SME}$) of 2.15 μV. Thus, the RMS($\widehat{SME}$) values were approximately half as large as the $\widehat{SD}_{Total}$ values. Using Equation 7, we can also estimate $\widehat{SD}_{True}$ for these data, yielding a value of 2.16 μV for the standards and 3.11 μV for the oddballs.

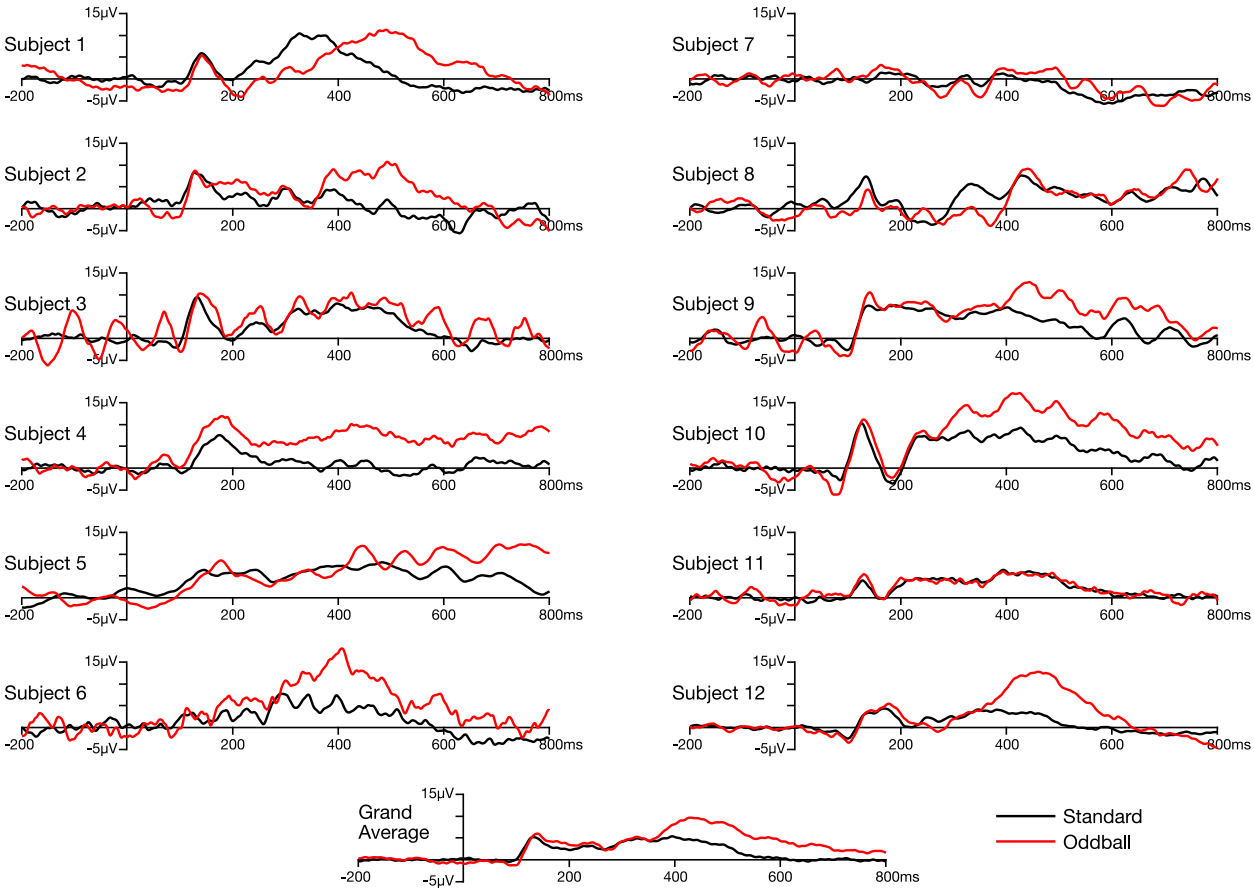

Figure S4. ERP waveforms from the parietal electrode sites, averaged separately across either the standard or oddball trials. Waveforms are shown for each participant as well as for the grand average across participants.

The fact that the RMS($\widehat{SME}$) values were lower than the $\widehat{SD}_{True}$ values indicates that the contribution of measurement error to the observed variability in scores across participants was not as great as the contribution of true differences among participants. The power calculator

provided by Baker et al. (2020) would allow you to determine the extent to which improving the data quality (e.g., by increasing or decreasing the number of trials) would increase the estimated statistical power given these RMS($\hat{SME}$) values[4].

Table 1 also shows that the RMS($\hat{SME}$) values are larger for the peak amplitude than for the time-window mean amplitude. This is consistent with previous claims that peak amplitude measures are more noise-sensitive than mean amplitude measures (Clayson, Baldwin, & Larson, 2013; Luck, 2014).

For both time-window mean amplitude and peak amplitude, the $\hat{SME}$ values are approximately twice as great for the oddballs as for the standards. As described in Section S6, this is consistent with the idea that the signal-to-noise ratio is related to the square root of the number of trials, combined with the fact that there were four times as many trials for the standards as for the targets.

Figure S4 shows the averaged ERP waveforms for each individual participant so that the visual appearance of the waveforms can be compared with the single-participant $\hat{SME}$ values in Table 1. Interestingly, the waveforms that "look" noisiest are not the waveforms with the largest $\hat{SME}$ values. For example, the waveforms look smoother for Subject 1 than for Subject 2, but the $\hat{SME}$ value for the time-window mean amplitude was worse (greater) for Subject 1. Similarly, the waveforms from Subject 3 look much noisier than the waveforms from Subject 5, but Subject 3 had a much better (smaller) $\hat{SME}$ value than Subject 5.

Why might there be this dissociation between subjective appearance of noise in the waveforms and the objective $\hat{SME}$ values? Previous research (Kappenman & Luck, 2010; Tanner, Morgan-Short, & Luck, 2015) suggests that low-frequency drifts have a large impact on statistical power for slow components like the P3 wave. However, it is difficult to see low-frequency noise in averaged ERP waveforms (although some drift can be seen in the baseline period of Subject 1's waveforms). To make it possible to visualize differences in low-frequency noise, Figure S5 shows the first ten single-trial EEG epochs for the standards in Subjects 1, 2, 3, and 5. Subjects 1 and 5 clearly had much larger low-frequency drifts in their data than Subjects 2 and 3, and this explains why the SME values were worse for Subjects 1 and 5. In other words, the low-frequency drifts for Subjects 1 and 5 created substantial trial-to-trial variability in the P3 time window, making it difficult to estimate the true P3 amplitude for these subjects from their averaged ERP waveforms.

Although these are merely anecdotal examples, they demonstrate how SME values may provide valuable objective information that is not obvious from visual inspection of the waveforms. Moreover, the $\hat{SME}$ values quantify the precision of the actual scores that are used in the statistical analyses of a study, whereas the noise that is most visually salient in the waveforms does not always have much impact on the statistical power of these analyses.

[4] Note, however, that a statistical comparison between the standards and oddballs would be a within-subjects comparison, so the covariance between conditions would also be important. As a result, it may be valuable to score the P3 amplitude from oddball-minus-standard difference waves and use bootstrapping to estimate the SME for this difference score.

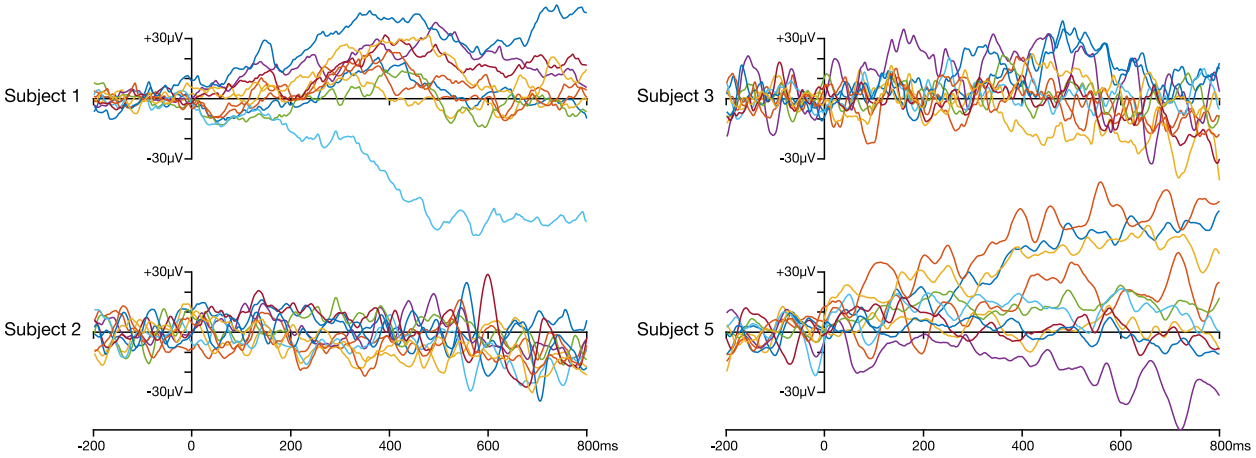

Figure S5. Single-trial EEG epochs from four individual participants. For each participant, the EEG epochs are shown for the first ten standard trials. The waveform colors are arbitrary.

S6. SME and the Signal-to-Noise Ratio

ERP methodology papers and books often note that the signal-to-noise ratio (SNR) of an ERP waveform increases as a linear function of the square root of the number of trials being averaged together (Luck, 2014; Picton, 2011; Regan, 1977; Thigpen et al., 2017). However, it is not always clear what is meant by signal and what is meant by noise. We can now provide a precise and flexible definition of the SNR by defining the "signal" as the amplitude or latency score we are using as our dependent variable (measured from a single-participant averaged ERP waveform) and the "noise" as the measurement error of this score (i.e., the SME for that score). We can express this simply as

$$SNR = \frac{Score}{SME}$$

which can be estimated as

$$\hat{SNR} = \frac{\hat{Score}}{\hat{SME}}$$

If our score is simply the voltage at a single time point in an averaged ERP waveform, then this voltage is the "signal," and the "noise" is the standard error of this signal and can be estimated using Equation 1 from the main text (as in Figure 1B). Because the denominator of Equation 1 is the square root of the number of observations (in this case, the number of trials), the SME will decrease linearly as the square root of the number of trials increases. This is why the SNR increases as a linear function of the square root of the number of trials.

The same is true if we use the time-window mean amplitude to score the amplitude of an ERP component. In this case, the signal is the time-window mean amplitude, and the noise is the SME of this score. If we use Equation 1 to estimate the SME of this score, then the denominator is again the square root of the number of trials, and the SNR will increase as a linear function of the square root of the number of trials (all else being equal).

In this approach, the "noise" portion of the SNR reflects the unreliability of the score of interest. This contrasts with other ways of computing the SNR in which the "noise" is quantified from the variability of the voltage during the prestimulus baseline period (Thigpen et al., 2017).

When bootstrapping is used to estimate the SME, our definition of SNR can be extended to virtually any type of ERP score, such as peak latency. For example, the observed peak latency of the P3 wave on oddball trials was 459 ms for the participant shown in Figure 6, and the $\widehat{SME}$ was 16.6 ms, so the $\widehat{SNR}$ for this peak latency measure was 459/16.6 or 27.6:1. However, because we are no longer using Equation 1, we cannot assume that the SNR will increase as a linear function of the square root of the number of trials. This is not a shortcoming of using SME to quantify measurement error; it is a consequence of measuring a score from the averaged ERP waveform that is not equal to the average of the single-trial scores. Without using the SME (or something similar) to quantify the measurement error, we would have no way of quantifying the SNR for such scores.

S7. The Assumption of Independence

Both the analytic SEM equation and bootstrapping assume that the individual trials are independent of each other, but this assumption will typically be violated because the trials are obtained consecutively (or nearly consecutively) from an organism that changes systematically over time. For example, factors such as mind-wandering may cause the ERPs to be more similar on consecutive trials than on nonconsecutive trials, and factors such as learning and fatigue may change systematically over the course of a recording session. Here, we address the likely impact of violating this assumption and methods for overcoming the lack of independence.

There are two major ways that violations of the assumption of independence can distort the $\widehat{SME}$, but they will have opposite effects. First, if the series of values across trials is positively autocorrelated, this will lead to an underestimate of the standard error (Bence, 1995). Such an autocorrelation could occur in amplitude measurements as a result of low-frequency drifts in the EEG. However, baseline correction should eliminate most of this autocorrelation, and high-pass filters will further reduce this autocorrelation. There may also be autocorrelation of latency values if, for example, participants alternate between periods of attentiveness (producing short latencies) and inattentiveness (producing long latencies). However, these autocorrelations are like to be small relative to the other sources of noise in the single-trial EEG epochs. It will be important for future research to examine the degree of autocorrelation for a variety of ERP scores and tasks. If significant autocorrelations are present, they can be quantified and an adjustment factor can be applied (Bence, 1995).

The second possibility is that systematic changes may occur that masquerade as noise. For example, imagine that the peak latency of the P3 wave always gets gradually faster over the course of an experiment, and the latency is therefore 50 ms earlier by the last 20 trials compared to the first 20 trials of an experiment. This will tend to cause the $\widehat{SME}$ to overestimate the amount of measurement error. Imagine that we used the empirical approach to estimating the standard error of the peak P3 latency in a given participant, in which we repeat the experiment 10,000 times and compute the $\widehat{SD}$ of the latency values across those 10,000 experiments from this participant. If the participant exhibits this systematic change in P3 latency in every replication of an experiment, then this would not create variability across replications. However, both Equation 1 and bootstrapping would treat this systematic trial-by-trial variation as random noise, and the estimated standard error would therefore be inappropriately large. It seems likely that this systematic trial-by-trial variation would be relatively modest compared to the other sources of noise that impact the ERPs, but that would need to be established empirically. If this violation of the assumption of independence does turn out to have a meaningful impact on

estimates of the SME, the impact could potentially be minimized in the bootstrapping process by making sure that each set of trials chosen for a given iteration contained an equal number of trials from, for example, each quarter of the session.

Similarly, imagine that we conduct an oddball experiment using X as the oddball and O as the standard in some trial blocks, but using O as the oddball and X as the standard in other blocks. Further, imagine that the P3 peak latency is slightly faster for the Xs than for the Os. Ordinarily, we would collapse across the Xs and Os, making one averaged ERP waveform for all of the oddball trials and another for all of the standard trials (see Luck & Gaspelin, 2017 for the rationale behind collapsing across factors such as this). However, if we obtained the $\hat{SME}$ from the entire set of trials, the different latencies for the different stimuli would cause variations in the P3 latency scores that would seem like noise but are actually systematic, and our $\hat{SME}$ would be inappropriately large. Again, this kind of systematic variability is likely to be small relative to the sources of random variability, so it would probably have only a small effect on the $\hat{SME}$. However, this small error in the estimate of the SME could presumably be eliminated by ensuring that every bootstrap iteration contained equal numbers of trials from the two stimuli.

S8. A Potential Criterion for Defining Extreme SME Values

Here we provide a potential criterion for excluding participants with extreme $\hat{SME}$ values. The goal is to determine whether the $\hat{SME}$ for a given participant is so large that including this participant would be expected to decrease the precision of our estimate of the group mean for the score of interest. In other words, this criterion states that a participant should be excluded if the inclusion of that participant would be expected to increase the standard error of estimate of the group mean. This approach assumes that the ultimate statistical tests will involve comparisons of group means, in which case larger effect sizes and greater statistical power would be expected if the standard errors of the group means are smaller.

Ordinarily, excluding a participant would increase the standard error of the group mean by decreasing the number of participants (*N*) in the denominator of Equation 1. Consequently, one would not want to exclude a participant unless the $\hat{SME}$ for that participant was so extreme that it outweighed the decrease in *N*. To determine whether an $\hat{SME}$ is this extreme, we first need to define the standard error of the group mean in a way that makes the contribution of the $\hat{SME}$ explicit. This can be done by estimating the standard error of the group mean using the estimate of $\hat{Var}_{Total}$ in Equation 5, which specifies the contribution of the $\hat{SME}$ values. Specifically, we can take the square root of this estimate of $\hat{Var}_{Total}$ to convert it into a standard deviation, and we can then use this standard deviation as the numerator of Equation 1 to estimate the standard error of the group mean ($\hat{SEM}_{Group}$):

$$\hat{SEM}_{Group} = \sqrt{\hat{Var}_{True} + MS\left(\hat{SME}\right)}/\sqrt{N} \qquad \text{(Equation S1)}$$

To determine whether the $\hat{SME}$ for a given participant is so extreme that the participant should be excluded, one could compute $\hat{SEM}_{Group}$ with versus without the $\hat{SME}$ from that participant (i.e., including versus excluding $\hat{SME}$ for that participant when computing $MS\left(\hat{SME}\right)$, and adjusting *N* accordingly). If $\hat{SEM}_{Group}$ is better (smaller) without a given

participant, then the group mean would be expected to be closer to the true group mean if that participant were excluded.

There are two caveats to this approach. First, it assumes that the SME is independent of the true score (i.e., that participants with noisier data do not differ in their true amplitude or latency scores from participants with cleaner data). This could be assessed by determining whether the $\hat{SME}$ values in a given study are correlated with the scores. If the $\hat{SME}$ values are, in fact, correlated with the scores, then excluding participants with extreme $\hat{SME}$ values would likely bias the results and could lead to incorrect conclusions.

A second caveat is that $\hat{SME}$ is only an estimate, and this estimate will be less precise when the number of trials is small (which is often the case in studies with extremely noisy data). As a result, some participants might be excluded unnecessarily because their estimated SME values are substantially higher than their true SME values. Given the potential risks of this (or any) criterion for excluding participants, it would be prudent to perform extensive simulations or evaluate the criterion across a broad range of previous data sets before applying it to new research.

S9. Reporting SME Values

In this section, we provide some recommendations for how to report $\hat{SME}$ values in publications. However, these recommendations are necessarily tentative given the broad range of ERP study designs and the newness of the metric. Our recommendations reflect the idea that the nature of the reported $\hat{SME}$ values should match the nature of the primary statistical analyses so that they are informative about the measurement error that actually impacts the effect sizes and statistical power in these analyses.

In a simple between-subjects design with one factor (e.g., P3 peak latency in two groups of participants), where the statistical analysis would be something like an independent samples *t* test, you could simply report the RMS($\hat{SME}$) values for each group (and possibly include a table with the single-participant values in supplementary materials).

In a simple within-subjects design with one factor (e.g., P3 peak latency for rare and frequent trials), where the statistical analysis would be something like a paired *t* test, the key factor in determining power is the variability of the difference between the two cells of the design. Indeed, a paired *t* test is identical to making a difference score for each participant and entering those values into a one-sample *t* test that compares the mean difference score against zero. Thus, you would want to quantify the $\hat{SME}$ of the difference between the scores. This would require bootstrapping: On each bootstrap iteration, you would get the two scores and the difference between them, and then you would calculate the $\hat{SME}$ for this difference score[5]. One of the demonstration scripts at https://doi.org/10.18115/D58G91 provides an example of computing this $\hat{SME}$ value. You could then report the RMS($\hat{SME}$) value for the difference score (and possibly include a table with the single-participant values in supplementary materials).

Designs with multiple factors are more complicated. One such design would be a 2-way ANOVA with one within-subjects factor with two levels (e.g., rare vs. frequent) and one between-subjects factor with 2 or more levels (e.g., patients vs. controls). The interaction term in

[5] Note that, unless your score is the time-window mean amplitude, the difference between the two scores is not the same thing as obtaining the score from a difference wave.

this ANOVA is equivalent to obtaining a difference score for each participant for the within-subjects factor and conducting a one-way between-subjects ANOVA (or an independent-samples *t* test) on the difference scores. Thus, it would typically be sensible to obtain the difference scores and report the RMS($\hat{SME}$) value for this difference score separately for each group.

Another common design would have a within-subjects factor with two levels (e.g., rare vs. frequent) and another within-subjects factor with two or more electrode sites (e.g., Fz, Cz, and Pz). If you are mainly interested in the main effect of the factor with two levels, you could simply average across electrode sites prior to computing the $\hat{SME}$ values. Indeed, we recommend not including electrode site as a separate ANOVA factor unless it is essential for testing your scientific hypotheses (Luck & Gaspelin, 2017). If you absolutely must test the trial type × electrode site interaction, and you can limit yourself to two electrode sites (or two clusters of electrode sites), you could compute a double difference score (e.g., rare-minus-frequent at Pz minus rare-minus-frequent at Cz). The two-way interaction in the main ANOVA would be equivalent to a one-sample *t* test against zero for this double difference score, so the $\hat{SME}$ values for this score would be a sensible way of quantifying the measurement error that is relevant for this interaction.

When a within-subjects factor must include more than 3 levels, there is no simple way to compute a single $\hat{SME}$ value that corresponds directly to this factor. Future work may be able to come up with a solution for such cases. In the meantime, researchers can simply report the $\hat{SME}$ values for each level of the factor.

S10. References for Supplementary Materials

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2020). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *ArXiv:1902.06122 [q-Bio, Stat]*. Retrieved from http://arxiv.org/abs/1902.06122

Bence, J. R. (1995). Analysis of short time series: Correcting for autocorrelation. *Ecology*, *76*, 628–639.

Boos, D. D. (2003). Introduction to the bootstrap world. *Statistical Science*, *18*, 168–174.

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How Many Trials Does It Take to Get a Significant ERP Effect? It Depends. *Psychophysiology*, *55*, e13049.

Brandmaier, A. M., Wenger, E., Bodammer, N. C., Kühn, S., Raz, N., & Lindenberger, U. (2018). Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). *ELife*, *7*. https://doi.org/10.7554/eLife.35718

Chernick, M. R. (2011). *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons.

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology, 50*, 174–186.

Cohen, J., & Polich, J. (1997). On the number of trials needed for P300. *International Journal of Psychophysiology, 25*, 249–255.

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC press.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*, 1166–1186.

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, *47*, 888–904.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique, Second Edition*. Cambridge, MA: MIT Press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*, 146–157.

Picton, T. W. (2011). *Human Auditory Evoked Potentials*. San Diego: Plural Publishing.

Regan, D. (1977). Evoked potentials in basic and clinical research. In *EEG Informatics: A Didactive Review of Methods and Applications of EEG Data Processing* (pp. 319–346). North-Holland: Elsevier.

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*, 997–1009.

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*, 123–138.