Getting a precise RT estimate for single items in a reading aloud task

Xenia Schmalz

Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy,
University Hospital, LMU Munich

Abstract

In single-word reading aloud studies, the researcher normally analyses the Reaction Time (RT): the time between target presentation and the onset of the participant's response. RT measures are inherently noisy. For noisy measures, a large number of observations are sometimes required to obtain reliable estimates. Reducing the noisiness of a measure by including many observations leads to more precise estimates of an effect, and to increased power, as it diminishes unsystematic variance and thus leads to higher standardised effect sizes. Here, a set of simulations aims to establish how the variability of the observed mean RT for a single item changes as a function of the number of participants who respond to this item. A rule-of-thumb cannot be established, because the researcher needs to decide on a range of expected values which is acceptable for their design and research question.

Getting a precise RT estimate for single items in a reading aloud task

For Registered Reports, grant applications, ethics applications, and similar documents, researchers are expected to provide a power calculation. Calculating power requires an effect size estimate. Sometimes, this is problematic, because the design may be novel, which makes it difficult to estimate a realistic effect size: even if when pilot data exists, the observed effect size is variable when the sample size is small (Whitehead, Julious, Cooper, & Campbell, 2016). Effect sizes can be estimated from previous studies; however, the presence of publication bias and questionable research practices leads to systematic over-estimation of the true effect size (Vasishth, Mertzen, Jäger, & Gelman, 2018). The design of the study might be complex, and standard and user-friendly software may not provide the option to calculate the power for a complex design. For exploratory studies, it is not known *a priori* which statistical tests will be conducted. Furthermore, calculating a formal power analysis for exploratory studies is problematic: it has been argued that *p*-values become uninterpretable in an exploratory context. Thus, power analyses, which find the long-term probability of getting a *p*-value in a specific range, are also meaningless (Nosek & Lakens, 2016). Regardless of these issues, sample size justifications are important for ethics and funding applications.

In justifying our sample size, an alternative to a power analysis is to plan for a certain degree of precision (Kelley, Maxwell, & Rausch, 2003; Schönbrodt & Perugini, 2013). For estimating precision, we use an *a priori* expectation of the standard deviation to calculate a confidence interval that guarantees that, in the long run, our observed estimate is within an acceptable bound. There is freedom in deciding the width of the confidence interval (e.g., 80%, 90%, 95%), and one needs to have an estimate of the standard deviation. The aim of the current simulations is to explore the precision of reaction time (RT) estimates on the item of single words in a typical reading aloud study.

In a reading aloud study, a number of participants see a number of words, and are asked to read it aloud as accurately and quickly as possible. The variable which is

analysed is RT: the number of milliseconds between the appearance of the item and the onset of the vocal response. The items are generally chosen to vary in some linguistic characteristic, and subsequent statistical analyses would be conducted to see if the linguistic characteristics affect the RT.

In most cases, the data would be analysed using a Linear Mixed Effect model, where item- and participant-level characteristics can be included as predictor variables. More information about calculating power and required sample sizes for Linear Mixed Effect models can be found elsewhere (Brysbaert & Stevens, 2018; Westfall, Kenny, & Judd, 2014) and; and a corresponding app can be found at <https://jakewestfall.shinyapps.io/crossedpower/>. Here, the focus is a different question: If we look at a single items, how many participants do we need to obtain stable estimates?

On the surface, the logic behind this question is very simple. For each item, one can calculate the average RT, across N participants. As N increases, the observed average should approach a hypothetical true population value. To assess which item-level characteristics affect RTs, it is advantageous to have as precise an estimate as possible. If only a few participants respond to each item, the average observed RT is likely to vary extensively across samples of the same size. This high between-sample variability is reflective of an imprecise estimate. In an ANOVA or regression analysis, where the dependent variable is the item-level average RT, imprecise estimates reduce the size of the effects, because they contain a smaller signal-to-noise ratio than more precise estimates.

Analyses

A larger number of observations should result in a more stable RT estimate. As a first step, one can visualise the likely trajectory of a given observed average RT, based on observed characteristics for various words, in the large-scale Lexicon projects. The English Lexicon Project (Balota et al., 2007) has responses for thousands of items, with up to 35 responses per item.

Figure 1 plots the expected trajectories of average RTs, as a function of N , for the word “vanishes”. The English Lexical Project, contains 35 valid and correct reading aloud responses for this item. The average reading aloud RT is $743.4ms$ ($SD = 345.3$). Based on the mean and SD, one can simulate the likely trajectories of the observed average RTs at different values of N . Using the item’s mean and SD, one can simulate a normal distribution, and draw a single value from it: The single value provides an RT for $N = 1$. Then the next value is draw from the same normal distribution, and the average is calculated between the first and second values. This provides an average RT for $N = 2$. This procedure can be repeated, while always plotting the observed average RT for each N . Figure 1 shows such a plot for 35 participants: each line represents a “walk” as described above, where the average RT approaches the RT which we specified as a parameter for our normal distribution. Then, the whole procedure is repeated to simulate more “walks”. Figure 1 shows 100 such “walks”.

FIGURE 1 ABOUT HERE

As expected, the initial average RTs tend to be very variable: if we were to stop our simulated data collection at $N = 5$, we have a relatively high probability to obtain estimates ranging from $400ms$ to $1200ms$. As the simulated data collection progresses (i.e., as N increases), the variability between the “walks” diminishes, and at $N = 30$ we would expect the observed average RT for the word “vanishes” to lie somewhere between $600ms$ and $1,000ms$.

Analytically, the variability at different values of N can be quantified as confidence intervals: the proportion of times that we expect the average RT to lie within the interval, in the long run. The width of the confidence intervals depends (1) on the confidence level that one would like to achieve (fixed here at 95%), (2) the population standard deviation (σ), and (3) the number of participants. The σ parameter is unknown, but one can get a plausible range of σ -values through data from the English Lexicon Project. Figure 2 shows a box-plot, reflecting the distribution of the SDs for each item, after removing observations with $RTs < 250ms$, which are likely to be miscoded.

FIGURE 2 ABOUT HERE

The distribution of SDs is left-skewed, with reflects a disproportional number of large values. Given the large range of observed SDs, taking a single point (e.g., an average or median) would be likely to under- or over-estimate σ under many realistic conditions. However, one can define a range of reasonable σ values, and conduct a sensitivity analysis to assess how the confidence interval differs both as a function of N and across different values of σ .

For the item-level SDs of the English Lexicon Project, the median is $SD_{median} \approx 200$, the 20% quantile, $SD_{lower} \approx 130$, the 80% quantile, $SD_{upper} \approx 350$, and a pessimistic estimate can be obtained by taking the location of the upper bar in the boxplot in Figure 2, $SD_{pessimistic} \approx 600$. For each of these SD estimates, one can calculate the 95% confidence range for different values of N , with the formulae:

$$CI_{upper} = 1.96 \cdot \frac{\sigma}{\sqrt{N}}$$

$$CI_{lower} = CI_{upper} \cdot (-1)$$

To calculate the expected range of average RTs, one would add these values to the average RTs. However, here we are concerned with the deviations from any hypothetical mean, therefore we can simply focus on the upper bound; the expected deviation is therefore $CI_{upper} \cdot 2$. Figure 3 shows CI_{upper} as a function of N for the four different SD estimates (low, median, high, and pessimistic).

FIGURE 3 ABOUT HERE

Many single-word studies contain between 20 and 30 participants per item (e.g., many of the large-scale Lexicon projects, see Table 1 of Keuleers, Lacey, Rastle, and Brysbaert (2012)). For this sample size ($N = 30$), the expected range of deviation ($CI_{upper} \cdot 2$) is $93ms$ in the most optimistic-case scenario, $143ms$ in the median-SD scenario, $250ms$ for the high estimate, and $428ms$ for the pessimistic scenario. Increasing N to 50 participants, the expected range of deviation is $72ms$ for the low estimate, $110ms$ for the median estimate, $194ms$ for the upper estimate, and $332ms$ for

the pessimistic estimate. For 100 participants, the range reduces to $50ms$, $78ms$, $137ms$, and $235ms$, respectively. Thus, in the median-SD scenario, increasing the sample size from 30 to 50 already more than halves the expected range of values.

Discussion

The current set of simulations aims to provide an intuitive understanding of how the precision of an RT estimate changes with an increasing number of observations. The expected ranges, calculated above, will differ depending on the level of confidence that the researcher wants to achieve. The width of an acceptable range also depends on the researcher's best judgment. In deciding on an acceptable range, researchers will need to balance precision against practical considerations, such as the feasibility of recruiting a given number of participants.

The current paper does not encourage the use of any rules-of-thumb. Rather, it aims to raise awareness of an important issue. Imprecise measurements introduce noise: thus, even if, in an analysis, the researcher calculates an average of many items which fall into the same condition, imprecise estimates on the level of the individual items will lead to decreased precision of the condition's average RT estimate, increase the variance, and thus lead to smaller standardised effect sizes, which in turn leads to a smaller probability of finding a significant effect.

As a limitation of the current analyses, the assumption that there is a true value for the average RTs is unreasonable. For example, familiarity with a given word will vary across participants: a psychology student is likely to respond faster to words that they encounter in their daily life, such as "depression", "diagnosis", "comorbidity", than someone who does not encounter these words on a regular basis (e.g., an economics student). Thus, the true RT is more likely to be a distribution rather than a single point.

In summary, researchers may rely on the current simulations to gain an idea about the extent to which their item-level observed RTs are likely to vary around a hypothetical population average, as a function of the number of participants. The

current simulations may also be useful as an example for future research aiming to simulate the precision of an effect or estimate of interest.

Open Code

Link to the analyses and simulations (in Python): <https://osf.io/mrnzj/>

References

- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The english lexicon project. *Behavior research methods*, *39*(3), 445–459.
- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1).
- Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation & the health professions*, *26*(3), 258–287.
- Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior research methods*, *44*(1), 287–304.
- Nosek, B. A., & Lakens, D. (2016). Registered reports: A method to increase the credibility of published reports. *Social Psychology*, *45*.
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612.
- Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175.
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020.
- Whitehead, A. L., Julious, S. A., Cooper, C. L., & Campbell, M. J. (2016). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical methods in medical research*, *25*(3), 1057–1073.

Figure 1. Convergence of 100 “walks” to a simulated average as a function of the number of observations.

Figure 2. Distribution of SDs for all items in the ELP.

Figure 3. Upper Confidence Interval as a function of N, for different estimates of SD.