

Supplementary material: power analysis

Owing to participant availability given our inclusion criteria, we were able to recruit 10 participants for this study. See Appendix S1 for details about the linguistic situation, variability, and sample size in multilingual settings. As part of the peer review process, we carried out post-hoc power analyses using the `mixedpower` package (Kumle, Võ, & Draschkow, 2021) in R. Although there is a body of literature that debates the usefulness of post-hoc power analyses in interpreting the findings of a completed study where statistical analyses have already been performed (Dziak, Dierker, & Abar, 2020; Gelman, 2019; Lakens, 2021; Lenth, 2007), we have included this material as a discussion around considerations for power calculation in bilingualism research.

The aim of the analysis was to explore the power of our experiment design to detect the effects we are investigating. The package simulates responses based on a given dataset and runs the specified analyses on these new data, estimating power by calculating how frequently statistical significance is obtained in the simulated data. The target of the analysis was to calculate the power for detecting an effect for our main variables of interest: vowel*context (to test the hypotheses L2-vowels and Asymmetry), and vowel*context*task (to test the hypothesis Paradigm).

Power is a function of sample size and effect size. Thus, a key issue in power analyses is to estimate the expected effect size for the treatment/factor of interest (Brysbaert & Stevens, 2018; Kumle et al., 2021). This may be expressed by a variety of measures, including Cohen’s d , η^2 , and model estimates (β). Using observed effect sizes from the data for power analyses has been shown to be meaningless (Gelman, 2019; Hoenig & Heisey, 2001), since such “observed power” is a function of the p-value. Therefore, we need a reasonable estimate of effect size. One way to do this is to use effect sizes reported in other existing literature (however, see Brysbaert and Stevens (2018) who recommends against using effect sizes from

single published studies, as these tend to be inflated), or meta analyses. An alternative approach is to define a smallest effect size that is meaningful to us (smallest effect size of interest; SESOI (Kumle et al., 2021)), and test whether the study has enough power to detect such an effect, if it were to exist. This value needs to be theoretically motivated. To the best of our knowledge, there is no existing literature on typical effect sizes for spectral differences in cross-language vowel production, and not many studies that report effect sizes. Only more discussion in this field will lead to a consensus on what is considered a typical or minimally meaningful cross-language effect. Similarly, while differences between language-switching paradigms have been informally discussed in previous literature, we are not aware of any study that has measured these differences in a controlled experimental setup, and thus no reported effect sizes for between-paradigm differences in phonetic transfer to act as a guideline.

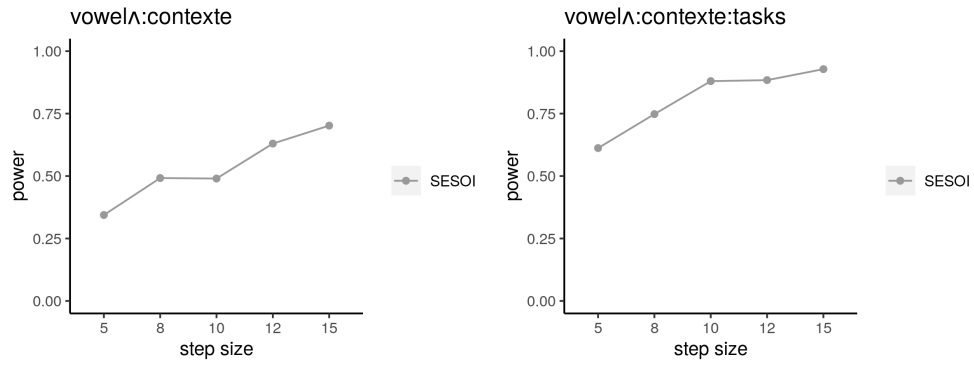
Cohen (1988) classifies effect sizes (Cohen’s d values) as small (≤ 0.2), medium (0.5), and large (≥ 0.8). Although the usefulness of this classification has been debated in the literature, these are often used as ballpark values for power analyses in psychological studies (for a discussion, see Brysbaert and Stevens (2018)). However, this approach does not generalize well to mixed-effects analyses, particularly for complex models with multiple random effects and interactions, as is typical in psycholinguistic and bilingualism research. “Determining the SESOI for (G)LMMs is difficult in a simulation-based approach where effect sizes are indicated through the model’s unstandardized beta coefficients...relating effect sizes to beta coefficients in complex models is far from trivial and the authors therefore refrain from making specific recommendations. (Kumle et al., 2021)”.

The results of a power analysis are only as meaningful as the assumptions that go into it. To avoid making arbitrary assumptions, we followed the approach in Brysbaert and Stevens (2018) (and discussed in Kumle et al. (2021)) of extracting the effect sizes from the study of interest and directly modifying all the beta coefficients by a fixed percentage. Note that

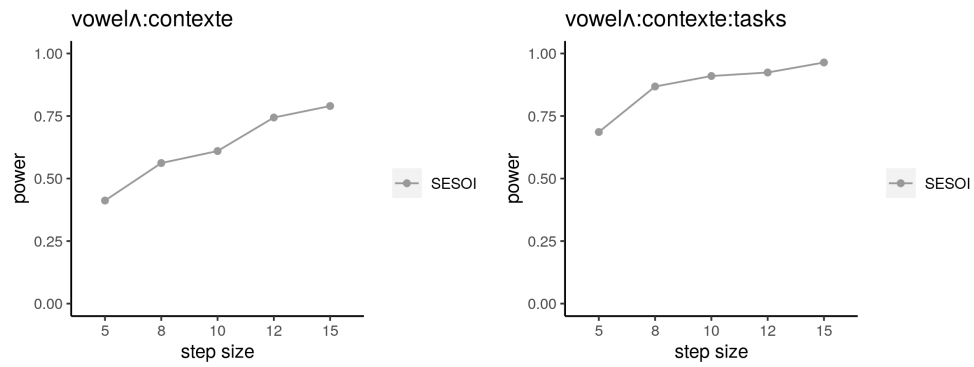
we are not assuming that doing so gets us closer to a “true effect size”. Instead, we run simulations with a range of effect sizes and report the power of our study to plausibly detect an effect of that size. The results are shown in the plots in figures ?? and ?. Each panel presents power analysis at a different effect size, and is based on 500 simulations. The plots show power (y-axis) over a range of sample sizes (x-axis), given the assumed effect size.

As expected, power increases as a function of sample size for all effects. However, of greater interest to us is the differences between the power curves in the four panels (a,b,c,d). These demonstrate how results of the power analysis vary as a function of assumed effect size. Note the the range of effect sizes used here is fairly narrow, varying between 15% smaller and 15% greater than the observed effect sizes. These differences demonstrate the importance of theoretically justified effect size estimates before attempting to make inferences based on power analyses. This applies equally to aposteriori power analyses done to make decisions about sample sizes prior to data collection (to reiterate, post-hoc power analyses are best avoided). This emphasizes the importance for bilingualism studies to report effect sizes as a part of their results, which can facilitate discussions and meta-analyses that eventually lead to such norms being established in the field. ¹

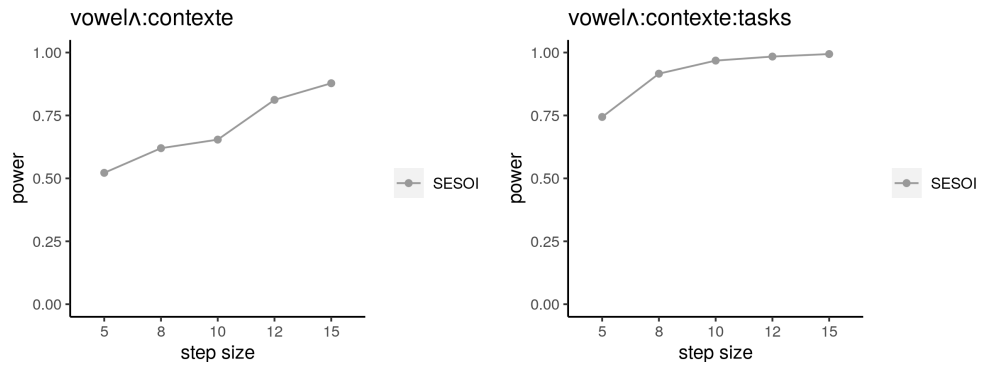
¹Thanks to the anonymous reviewer at BLC for initiating a fruitful discussion around this.



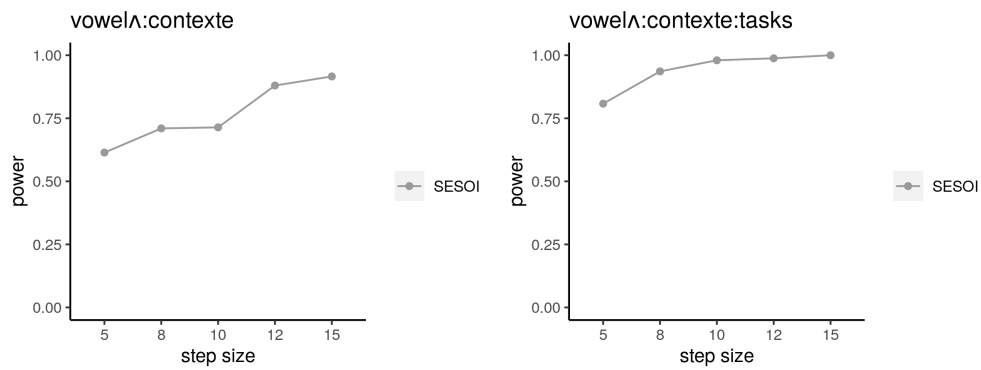
(a) Effect size = observed effect size - 15%



(b) Effect size = observed effect size - 5%

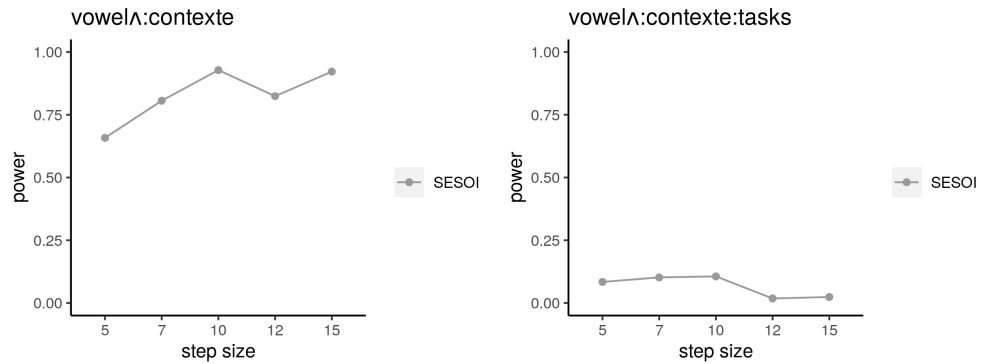


(c) Effect size = observed effect size + 5%

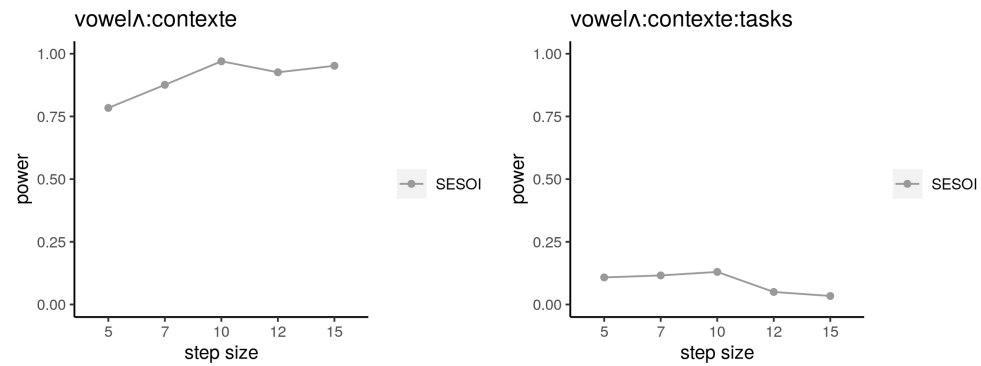


(d) Effect size = observed effect size + 15%

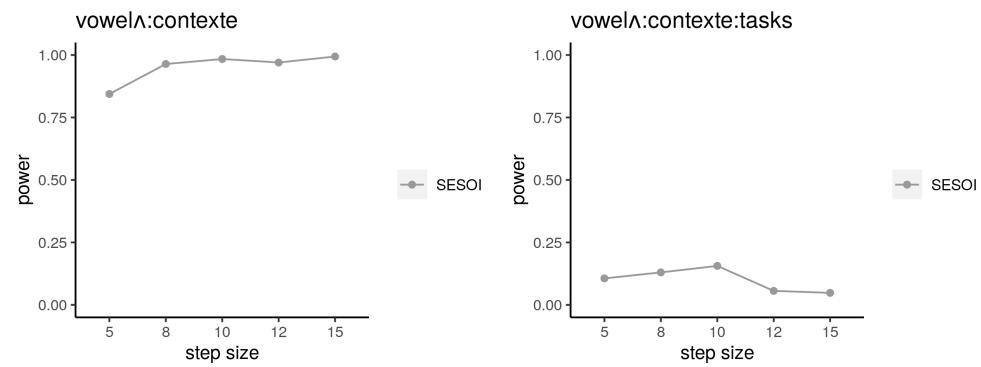
Figure 1: Power analysis for fixed effects of interest at varying sample sizes: F1



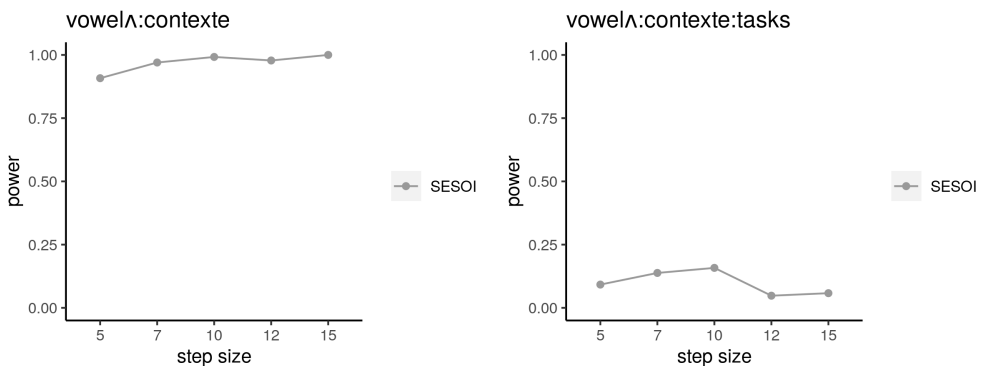
(a) Effect size = observed effect size - 15%



(b) Effect size = observed effect size - 5%



(c) Effect size = observed effect size + 5%



(d) Effect size = observed effect size + 15%

Figure 2: Power analysis for fixed effects of interest at varying sample sizes: F2

References

- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1(1).
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.
- Dziak, J. J., Dierker, L. C., & Abar, B. (2020). The interpretation of statistical power after the data have been gathered. *Current Psychology*, 39(3), 870–877.
- Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of surgery*, 269(1), e9–e10.
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 19–24.
- Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in r. *Behavior research methods*, 53(6), 2528–2543.
- Lakens, D. (2021, Jan). Sample size justification. PsyArXiv. Retrieved from psyarxiv.com/9d3yf doi: 10.1525/collabra.33267
- Lenth, R. V. (2007). *Post hoc power: tables and commentary*. Iowa City: Department of Statistics and Actuarial Science, University of Iowa, 1–13.