A SIMULATED ANALYSIS OF REPEATED MEASURES IN VOT: WE NEED MORE TOKENS!

Kyle Parrish

Rutgers University
kyle.parrish@rutgers.edu

ABSTRACT

In speech research, studies often make use of repeated measures, in which multiple productions of a given sound are taken from the same participant and incorporate them into multilevel models. For instance, in studies of voice onset time (VOT), researchers often elicit productions of stop consonants from multiple words produced by the same participant. It is unclear, however, how many repetitions of the same segment are necessary and how researchers go about choosing this number. The present study used reported data from previous literature to simulate an underlying distribution of 1000 points for each stop consonant in American English (/p/, /t/, /k/, /b/, /d/, /g/) in order to determine how many of these tokens were necessary so that the random sample would be practically equivalent sample to the full simulated distribution. The results suggest that at least 60 tokens are necessary for all 6 stop consonants to achieve a practically equivalent sample (equivalence bounds: Cohen's d = +/- .4). All materials used to create this paper, the application and the R code used to run the simulation can be found at https://osf.io/k8534/.

Keywords: Repeated Measures, VOT, simulation.

1. INTRODUCTION

Linguistic and psychological research of human subjects is often concerned with characterizing the behavior of a participant in distinct experimental conditions. It is generally agreed that, in most cases, it is necessary for experimenters to collect a sample of participants from a hypothetical underlying distribution, since the resources involved make it impossible to collect data from every possible subject meeting some given group criteria. For example, if one wanted to determine how the segment /a/ is produced in American English, it would likely not be feasible to collect tokens of /a/ from every American English speaker and only a sample of these speakers could be collected due to resource constraints. In general, the number of participants necessary for a faithful sample are determined by power analysis, where a researcher specifies an effect size, power level, significant threshold and finds the minimum number of participants needed to reliably find that effect.

In addition to knowing the appropriate number of participants for a study, the number of repetitions of a particular sound by each of those participants is another factor which has received much less attention. Following the same logic as sampling from a distribution of participants, it is also the case that researchers cannot collect all productions of a given sound from a subject, and must collect a sample. While most studies in speech research utilize repeated measures, it is unclear how the number of productions of a given segment are justified.

The quantity of repeated measures has varied in the literature. For instance, in their tutorial for multilevel models, Baayen and colleagues [2] simulated three tokens per condition in their multilevel model. Although their data set was simulated to demonstrate the nested structure of data and that data points from the same individual are not independent, the decision of three repetitions per token in this paper appears to be arbitrary and was not specifically justified.

In speech research, the quantity of repeated measures has also varied. Given their abundance, VOT studies provide a good example of the variation in repeated measures in speech research, and are the focus of the present paper. To investigate the use of repeated measures in recent VOT studies, a brief analysis of relevant studies in VOT was carried out focusing on the Journal of Phonetics and Google Scholar. Using both websites, a search for the term "VOT production" sorted by relevance revealed that the six most relevant articles range from 3 to 50 tokens per segment and condition [6], [1], [18], [11], [9], [4]. To be clear, these studies ranged in how they divided conditions. In some cases, the all stop consonants were presented in a single vocalic context [18], while others elicited stops with many vowel sounds [4]. Here, the focus of repeated measures is the quantity of tokens per segment per language (in a bilingual context), thus the same stop in the same language, but in a different vocalic context were counted as belonging to the same condition. For example, Chodroff and Wilson [4] analyzed all 6 stops in American English and included 5 repetitions of each stop in 10 vocalic contexts, for a total of 50 productions of each stop per participant. It is important to note, however, that there is evidence that repetitions of the same word by the same participant cease to be independent of one another [22]. As a result, the underlying distribution may not be a normal distribution, which is a key assumption of the current work. Ideally, repetition of the same word should be avoided.

Table 1 shows a breakdown of the top studies returned by the search and their tokens per condition.

Table 1: Most relevant studies by filter in the Journal of Phonetics

Study	Repetitions
Olson (2013)	3
Antoniou (2011)	4
Hussain 2018	5
Fish (2017)	9 – 12
Gorba & Cebrian (2022)	10
Chodroff & Wilson (2017)	45 – 50

In addition to the studies from the Journal of Phonetics, the top 10 studies from the Google Scholar search for "VOT production" were also analyzed [23], [13], [8], [10], [7],[12], [21], [17], [15], [20]. The mean number of productions of a given sounds in this subset was 23.9 (sd = 28.6). This high standard deviation was due to the Nielsen study [17], which elicited 100 productions of a given sound. Omitting this study, the mean tokens were 15.4 (sd = 10.8; range 3-35).

The source of the variation in the number of repeated measures in the literature is unclear, but could be made more consistent by carrying out a power analysis, similarly to how sample size is justified. The power of a sample refers to the probability of detecting an effect when it exists [5]. A power analysis consists of four parts, an significance threshold (typically .05), a power level (typically .8), a sample size and an effect size. Given three of these four, a power analysis calculates the missing number. For example, with the significance threshold at .05 and the power level of .8, a power analysis reveals that about 80 participants are needed per group to reliably detect a small effect (Cohen's D = .4). Although speech research is often concerned with finding differences in groups or conditions, it is also possible and useful to conduct a power analysis to determine the probability of detecting practical equivalence within a particular effect size. In this case, a threshold for practical equivalence can be specified, along with a number of samples and significance threshold. As a result, the power analyses of the present study refer to the probability of detecting practical equivalence, rather than a particular effect size, when it exists.

Table 2: Most relevant studies of "VOT production" on Google Scholar

Study	Repetitions
Zlatin & Koenigsknecht (1976)	35
Kupske (2017)	18
Gandour & Dardarananda (1984)	30
Gabriel et al. (2018)	3
Harada (2003)	9
Khattab (2002)	17
Ryalls et al. (1997)	9
Nielsen (2011)	100
Llama et al. (2016)	6
Riney et al. (2007)	12

The present study implements this approach using the reported means and standard deviations in Chodroff and Wilson (2017) [4]. The present study carries out several power analyses on large simulated data sets of a 6 stop consonants in English to determine how many tokens are necessary to consistently generate a sample that reliably falls within a small effect size of the simulated underlying distribution. That is, if we consider the speaker to be producing VOT from an underlying distribution of possible values, we can analyze how many samples from that distribution are needed to achieve a particular level of precision when we characterize an individual's stop production. In particular, the present study is guided by the following research question:

RQ1: For each of the 6 stop consonants in English, how many tokens are necessary to produce a sample that is practically equivalent (falls within Cohen's d = +/- .4 on the test of equivalence when p > .05) to the underlying sampling distribution (a simulated data set of 1000 points).

It was predicted that, just like power analyses of needed participants, that more samples would be associated with higher power.

2. METHODS

All data used for the present study were simulated using the rnorm function in R. First, for each stop

(/p/,/t/,/k/,/b/,/d/,/g/), an underlying distribution of 1000 points each was generated. The means and standard deviations of each distribution came from the from the literature (see Table 3) [4] for all 6 top consonants. This underlying distribution was generated to serve as a the representation of all possible realizations of a given consonant which were each equally probable to be produced by a given individual, although the true mean and standard deviation of this distribution will never be known.

Table 3: Means and Standard Deviations for each Stop Consonant in American English from Chodroff & Wilson (2017)

Segment	Mean (SD)
/p/	89(27)
/t/	98(28)
/k/	99(24)
/b/	13(5)
/d/	21(7)
/g/	28(10)

Then, using the generated underlying distribution a loop randomly selected a given number of tokens from the total of 1000. The sample sizes analyzed were 10, 20, 30, 40, 50, 60, 70, and 80 which were looped 100 times each. For example, if the sample size was 10, then the function chose 10 random rows of the total of 1000 per iteration. Following each selection of the random tokens, a test of equivalence [14], and a Welch's t-test were carried out between the random sample (e.g., the 10 random tokens of that iteration) and underlying distribution (all 1000 simulated data points). This process, which generated 48000 total data points (100 simulations x 8 sample sizes x 6 segments) was repeated 10 times. Then, the percentage of equivalent samples per segment per sample size per repetition was calculated. Essentially, this process generated 10 power analyses in each possible case of segment and sample size.

The test of equivalence used equivalence bounds of +/- d = .4, or a less than a small effect in language research based on a recent study [19]. In other words, the present study considered a random sample from a participant's greater distribution to be practically equivalent when the difference between the sample and all their productions was less that a small effect. The data were coded for whether or not the test of equivalence or t-test were was significant for a given iteration (1 for yes/0 for no).

3. RESULTS

Figure 1 shows the percentage of practically equivalent results in each sample size for each segment. Each box in the plot is made up of 10 total data points, where each of them is the total number of practically equivalent samples out of 100. The figure suggests that at least 60 samples of a given segment are needed to achieve a sample that is within d = +/- .4 at least 80 percent of the time.

10 tokens per segment was the least successful at producing a practically equivalent sample, as all segments were practically equivalent 0 percent of the time. Table 4 reports the percentage of significant tests of equivalence for each segment and each sample size. Again, for each segment, at least 60 tokens are necessary to produce a practically equivalent sample at least 80 percent of the time.

Figure 1: Quantity of practically equivalent instances per sample size

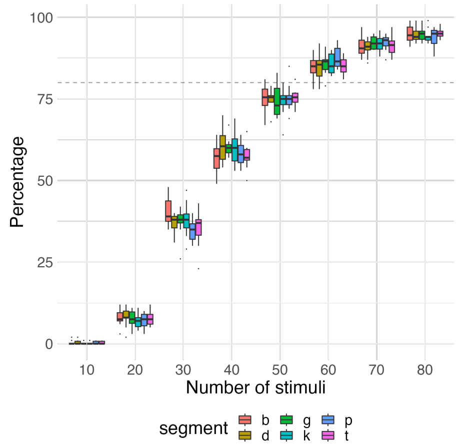

Table 4: Percentage of significant Tests of Equivalence per sample size and segment

n	p	t	k	b	d	g
10	0	0	0	0	0	0
20	0.08	0.08	0.07	0.07	0.07	0.08
30	0.4	0.37	0.37	0.38	0.35	0.35
40	0.57	0.61	0.6	0.6	0.58	0.58
50	0.75	0.74	0.75	0.74	0.75	0.76
60	**0.84**	**0.85**	**0.86**	**0.86**	**0.88**	**0.85**
70	**0.91**	**0.91**	**0.92**	**0.92**	**0.92**	**0.91**
80	**0.95**	**0.95**	**0.95**	**0.94**	**0.94**	**0.95**

3.1. False Positive results

In frequentist analysis, the significance threshold (alpha: typically .05 in linguistic research), refers to the false positive rate. Given that our significance threshold was .05, it is expected that around 5 percent of the total t-tests would be (falsely) positive. In total, 1903 of 48000 t-tests were significant (3.9 percent). Each of these iterations also had a significant test of equivalence, and there were 1900 instances of a significant t-test and an insignificant test of equivalence and 3 cases of a significant t-test and a significant test of equivalence. Figure 2 visualizes the pooled quantity of false positive tests in the whole data set by sample size. The figure suggests that the false positive rate does not vary much by segment or sample size, and stays around or below .05. Table 5shows the mean false positive rate for each segment and also shows that the false positive rate ranged from .04-.06, regardless of sample size or segment.

Figure 2: False Positive rate per sample size

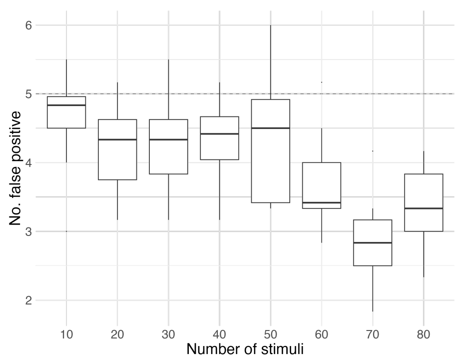

Table 5: Percentage of significant Welch's T-tests per sample size and segment

n	p	t	k	b	d	g
10	0.04	0.05	0.05	0.04	0.06	0.04
20	0.04	0.04	0.05	0.04	0.04	0.05
30	0.05	0.04	0.04	0.05	0.03	0.05
40	0.04	0.04	0.05	0.03	0.05	0.05
50	0.05	0.04	0.05	0.05	0.03	0.04
60	0.04	0.04	0.04	0.04	0.03	0.04
70	0.03	0.03	0.02	0.02	0.03	0.03
80	0.03	0.04	0.03	0.04	0.03	0.02

4. DISCUSSION

The results suggest that, at large, more tokens per segment should be used in speech research, at least for VOT studies in English. These results do not directly take into account various additional factors which have been found to impact VOT, and it is possible that fewer tokens might be needed in the event that there is reason to believe that less variation exists in the underlying distribution from which an experiment is sampling. The means and standard deviation from the present study [4] included each stop 50 times in 10 vocalic contexts, which therefore likely includes a wider range of possible values since the following vowel has been shown to impact VOT [16].

In general, this study serves as an example for speech researchers to consider how repeated measures are justified in their research, and should be considered in tandem with the smallest effect size of interest. For the purpose of this paper, the assumption was made that any sample that fell within a small effect size of the underlying distribution was practically equivalent, but this may not be a consensus for other measures, or even in VOT for different purposes. For example, very small effects might be taken as evidence for subtle changes in speech such as cross-linguistic influence, studies in phonetic drift or phonetic accommodation. In these cases, the quantity of repeated measures may need to be increased even more for better precision.

These results provide additional context in the issues surrounding low statistical power in not linguistic research with and emphasis on speech research. A recent review of linguistic research at large found that studies are typically under powered in regard to number of participants [3]. That is, the low quantity of participants per group or condition increase the probability of a false negative finding and virtually guarantee an over-estimated effect size when an effect is found. The finding, together with the present study, suggest that researchers analyze how many participants are necessary and how many tokens to elicit from each participant per condition prior to carrying out a study. Power analysis, again, provides researchers with a necessary tool for examining these questions.

The present study also includes open materials, including the code used to run all analyses and to produce this manuscript. Additionally, a shiny application was created to easily reproduce this analysis for a single segment given a mean and standard deviation. That is, by using the app and plugging in the mean and standard deviation, an underlying distribution of 1000 points is created, and samples of 10, 20, 30, 40, 50, 60, 70 and 80 are taken, randomly, 100 times each. A plot is then produced which shows the number of times that each of these samples is practically equivalent in a test of equivalence (bounds d = +/- .4). The equivalence bounds and sample sizes can also be adjusted.

5. CONCLUSION

In conclusion, the present study has argued that more attention should be paid to the justification of the quantity of repeated measures in speech research. In specific, it is recommended that 60 tokens per segment be given to a speaker in a specific condition. For example, if one wants to study VOT of a bilingual in two languages, at least 60 tokens per segment per language would be ideal, based on the current data. A companion shiny application was also created together with this paper, and can be used to quickly replicate this analysis during planning of research projects: https://kparrish92.shinyapps.io/repeated_measures_app/.

6. REFERENCES

[1] Antoniou, M., Best, C. T., Tyler, M. D., Kroos, C. 2011. Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. *Journal of phonetics* 39(4), 558–570.

[2] Baayen, R. H., Davidson, D. J., Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language* 59(4), 390–412.

[3] Brysbaert, M. 2021. Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition* 24(5), 813–818.

[4] Chodroff, E., Wilson, C. 2017. Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in american english. *Journal of Phonetics* 61, 30–47.

[5] Cohen, J. 1992. Statistical power analysis. *Current directions in psychological science* 1(3), 98–101.

[6] Fish, M. S., García-Sierra, A., Ramírez-Esparza, N., Kuhl, P. K. 2017. Infant-directed speech in english and spanish: Assessments of monolingual and bilingual caregiver VOT. *Journal of Phonetics* 63, 19–34.

[7] Gabriel, C., Krause, M., Dittmers, T. 2018. VOT production in multilingual learners of french as a foreign language: Cross-linguistic influence from the heritage languages russian and turkish. *Revue française de linguistique appliquée* 23(1), 59–72.

[8] Gandour, J., Dardarananda, R. 1984. Voice onset time in aphasia: Thai ii. production. *Brain and Language* 23(2), 177–205.

[9] Gorba, C., Cebrian, J. 2022. The acquisition of l2 voiced stops by english learners of spanish and spanish learners of english. *Speech Communication*.

[10] Harada, T. 2003. L2 influence on L1 speech in the production of VOT. *Proceedings of the 15th international congress of phonetic sciences*. Causal Productions Barcelona 1085–1088.

[11] Hussain, Q. 2018. A typological study of voice onset time (VOT) in indo-iranian languages. *Journal of Phonetics* 71, 284–305.

[12] Khattab, G. 2002. VOT production in english and arabic bilingual and monolingual children. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4* 1–38.

[13] Kupske, F. F. 2017. A complex approach on integrated late bilinguals' english VOT production: a study on south Brazilian immigrants in London. *Ilha do Desterro* 70, 81–94.

[14] Lakens, D. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science* 8(4), 355–362.

[15] Llama, R., López-Morelos, L. P. 2016. VOT production by spanish heritage speakers in a trilingual context. *International Journal of Multilingualism* 13(4), 444–458.

[16] Nearey, T. M., Rochet, B. L. 1994. Effects of place of articulation and vowel context on VOT production and perception for french and english stops. *Journal of the International Phonetic Association* 24(1), 1–18.

[17] Nielsen, K. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39(2), 132–142.

[18] Olson, D. J. 2013. Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production. *Journal of Phonetics* 41(6), 407–420.

[19] Plonsky, L., Oswald, F. L. 2014. How big is "big"? interpreting effect sizes in l2 research. *Language learning* 64(4), 878–912.

[20] Riney, T. J., Takagi, N., Ota, K., Uchida, Y. 2007. The intermediate degree of VOT in japanese initial voiceless stops. *Journal of Phonetics* 35(3), 439–443.

[21] Ryalls, J., Zipprer, A., Baldauff, P. 1997. A preliminary investigation of the effects of gender and race on voice onset time. *Journal of Speech, Language, and Hearing Research* 40(3), 642–645.

[22] Winter, B. 2015. The other n: The role of repetitions and items in the design of phonetic experiments. *ICPhS*.

[23] Zlatin, M. A., Koenigsknecht, R. A. 1976. Development of the voicing contrast: A comparison of voice onset time in stop perception and production. *Journal of Speech and Hearing Research* 19(1), 93–111.