THE OTHER N: THE ROLE OF REPETITIONS AND ITEMS IN THE DESIGN OF PHONETIC EXPERIMENTS

Bodo Winter
University of California, Merced, Cognitive and Information Sciences

ABSTRACT

Some branches of phonetics prefer experiments that feature a large number of repetitions and only few unique items. This paper discusses this relatively common experiment design choice, arguing that repetitions do not always help for drawing sound conclusions from phonetic data. It is recommended that phonetic experiments should be designed with less repetitions and more distinct items.

Keywords: repetitions; experimental design; statistics; phonetic experiments

1. INTRODUCTION

Speech is inherently variable. One source of variability is differences between speakers. Sometimes inter-speaker variability is of critical interest to phoneticians, such as when doing individual differences studies [e.g., 9]. Most often, however, phoneticians seek generalizations that hold across particular speakers.

Another source of variability is the fact that even within a given speaker, multiple productions of the same utterance can never be exactly the same [see, e.g., 7]. Sometimes, differences between particular productions are of special interest, such as when studying repetition priming or predictability effects [1, 8, 21]. Most often, however, phoneticians seek generalizations across particular production events.

Putting repetitions into phonetic experiments is a frequent strategy to counteract utterance-by-utterance variability, especially in the sub-field of speech production research. In fact, a brief review of all 2014 issues of the *Journal of Phonetics* shows that about half of the experimental (non-corpus, non-modeling) studies featured exact repetitions of the same item. A lot of these studies featured more than five repetitions. This shows that designing experiments with repetitions is relatively common in phonetics.

2. SAMPLING FROM A POPULATION OF PRODUCTION EVENTS?

Putting the preceding discussion into the terms of inferential statistics, phoneticians generally sample from a population of speakers. And when experiments have multiple repetitions, the idea is to sample from a population of possible production events. When using inferential statistics (such as t-tests, ANOVAs, regression, mixed models etc.), we wish to make claims about the population of speakers and the population of production events based on limited information from our samples.

In theory, repetitions could allow getting a more precise estimate of what a speaker would “usually” or “ideally” say. For example, Broad and Clermont [5: 54] mention that in their speech production study, they averaged over five repetitions to assure “statistical stability.”

Such an average over repetitions, however, only works if at least two conditions are met: First, repeated events need to be independent (i.e., one rendering of an utterance is not influenced by another rendering of the same utterance). Second, repeated events need to be approximately normally distributed.

Addressing the first point, it is clearly not the case that repetitions are independent events. How an utterance is being produced depends on whether and how many times it has been produced before [1, 8, 21], for example, repeatedly saying the same word generally leads to reduction. Moreover, an analysis of over 1,000 repetitions of the word “bucket” indicates the presence of long-range correlations in repeated productions of the same word [10]. In the case of inter-dependent rather than independent production events, taking the mean across repetitions will lead to conflating a more precise estimate of a production target with whatever systematic repetition effect there is. This means that the mean becomes a biased estimator of a phonetic target.

Kello’s study on long-range correlation in repetition data [10] furthermore shows that variation across repetitions is not normally distributed, but follows heavy-tailed distributions. It is known that the mean as an estimator of central value is unreliable in such a situation.

Thus, prior research on repetition priming and long-range correlations in repetition data suggests that averaging over repetitions is not likely to lead to a more precise estimate of a production target. When phoneticians do average over repetitions to gain a more precise statistical estimate, they implicitly assume that variation across repetitions is normally distributed and that repetitions are independent events.

3. THE OTHER N: GENERALIZING OVER ITEMS

Besides inter-speaker variation and inter-utterance variation, there is also variation between different linguistic items. Phonetic and phonological phenomena may be present to differing extents for different words or sentences. This is to be expected based on exemplar theoretic accounts and the word-specific phonetics they entail [19]. And variation across words is to be expected because different parts of the lexicon participate to differing extents in sound change [15, 25]. On top of this, there are predictable lexical effects, such as those involving word frequency.

Some phonetic experiments only sample very few items. In the extreme, only one item is analyzed. Take, for example, categorical perception [16]. Many experiments on this topic, including the author's own [27], only use one item pair (e.g., *bear/pear*) or one item pair per condition (e.g., one minimal pair with a bilabial voicing contrast and one minimal pair with an alveolar voicing contrast). Then, an acoustic continuum is generated between the two members of the minimal pair. This results in many tokens that vary in the acoustic dimension of interest, however, there is still only one item *type*, i.e., only one minimal pair.

In such a single item design, generalization over items is impossible. From a strict logical viewpoint, we do not know whether any result obtained with a single item may apply to any other item at all. Although it is reasonable to assume that the case of *bear/pear* would carry over to pairs such as *bay/pay*, only testing one item entails that there is no statistical demonstration that the observed results apply to other items as well. Hence, on the basis of a single item design, we may argue that the results carry over to other items, but we have not quantified the extent to which this actually happens.

For other phonetic phenomena, in particular phenomena that are characterized by small effect sizes or that are currently undergoing change, conclusions based on few items may, however, be much more off from reality than in the case of categorical perception.

In some phonetic studies, "N" is implicitly characterized either as the number of participants, or as the number of data points in total (tokens rather than types). What is oftentimes missing is the idea that one should also aim for a high "N" of items: A strict test of any phonetic hypothesis should demonstrate that the phenomenon in question applies to a sufficiently large number of speakers, as well as a sufficiently large number of items. A similar conceptual shift has been undergoing for a long time in psycholinguistics [6], nowadays often in the form of using mixed models with subjects *and* items as random effects [2, 3]. The necessity of inferential statistics for not only subjects but also items has also been stated for other fields [13, 14].

4. TYPE I ERROR SIMULATION

How is the issue of items connected to the issue of repetitions? In this section, a simulation is presented which shows that studies with few items and many repetitions are more likely going to obtain Type I errors, that is, erroneously significant results.

4.1. Simulating the effects of repetitions

A Type I error simulation was conducted with R [22] to explore the effect of differing item and repetition numbers on common analysis choices observed in the phonetic community. For examples of similar simulations, see [3, 23, 26].

To make things concrete, imagine that a researcher is analyzing the difference between Seoul Korean lax and aspirated stop VOTs. Korean has recently undergone sound change to the extent that the VOT distributions of lax and aspirated stops have effectively merged [12, 24]. Emulating this merger, we simulate a lexicon of 5,000 words with tense and lax stops that *both* have a mean VOT of 60ms with 10ms standard deviation. Even though they come from the very same distribution, half of this lexicon is marked "lax," the other half "aspirated."

This means that there is no significant difference between tense and lax stops in this "population" of 5,000 words (unpaired t-test across items; t(4998)=0.3, p=0.74). However, due to chance sampling, there are always going to be small differences between aspirated and lax stops in any given subset of this lexicon. We simulated 1,000 datasets with 2, 4, 8, or 16 unique items drawn from the lexicon, as well as 2, 4, 8, 16 or 32 repetitions of each item. Each of these artificial "experiments" is conducted with 12 speakers.

With 1,000 datasets randomly drawn from the same distribution, we should expect about 50 significant results for an alpha level of 0.05 (a common significance level in phonetics). Knowing that there is no lax/aspirated contrast in the population, any significant result is *by definition* a Type I error.

To put the simulation as much as possible in favor of experiments that have repetitions, no repetition priming effect was implemented. Instead, there only was random trial-by-trial variation, drawn from a normal distribution with SD=20ms. This represents the optimal situation for using repetitions, where having more repetitions actually yields a more precise estimate of the underlying production target.

Several common analysis choices have been implemented. Some of those include a series of mixed models [20] constructed with the *lme4* package [4] and different model specifications (some of them ignoring repetition as a factor in the experimental design). All models included random intercepts for both subject and items [2], and additionally by-subject and by-item slopes for the effect of "consonant type" (lax vs. aspirated), since mixed models without random slopes are known to be anti-conservative [3, 23].

For the present discussion, the most important analysis choice is the "subjects-analysis," where the researcher averages over items and repetitions so that each subject only has two unique data points, one for lax and one for aspirated. This is a common analysis choice for phonetic data. Averaging makes sure that each subject only provides one data point, as is required for the paired t-test that would be used for the comparison of lax and aspirated stops. More details on different analysis choices can be found in the simulation script, which, in line with standards of reproducible research [17, 18], can be retrieved from the author's webpage.

4.2. Results

Table 1 shows the average Type I error rate for the subjects analysis described above. Ideally, each cell should be around 0.05, the community's accepted alpha level. But, as can be seen, Type I error rates are much higher than that when doing *only* a subjects-analysis (and no items-analysis). Moreover, there is a marked trend with Type I error rates increasing as the simulated experiments contain more repetitions.

Table 1: Type I error rates of subjects-analyses (within subjects, paired t-test) for phonetic experiments with 2, 4, 8, or 16 items and 2, 4, 8, 16 or 32 repetitions. Cells give proportions of significant results (p<0.05) out of 1,000 simulations.

Items	R=2	R=4	R=8	R=16	R=32
2	0.62	0.75	0.81	0.87	0.90
4	0.64	0.72	0.82	0.87	0.89
8	0.63	0.72	0.81	0.84	0.91
16	0.65	0.75	0.82	0.87	0.89

Why is it that having more repetitions increases Type I error rates? And why are Type I error rates so high overall for the subjects-analysis? If one averages over repetitions and items, each subject only contributes one data point per condition and a paired t-test would have the appropriate degrees of freedom, which, in theory, should lead to an acceptable Type I error rate.

To illustrate the cause of the Type I error inflation shown in Table 1, consider a sample of only four items. In this particular sample, there happens to be a 3 ms difference between lax and aspirated stops. For example, it might be that the researcher (inadvertently) selected two somewhat more frequent lax words, leading to shorter VOTs and two somewhat less frequent aspirated words, leading to longer VOTs. In the population, there is no difference, but chance sampling produced a 3 ms "effect."

Remember that in the simulation, there is trial-by-trial noise added to each data point (SD=20ms). Trial-by-trial noise will assure that most of the time, this 3 ms difference will not become significant. Having many repetitions, however, will mean that precision is increased, so that across subjects, the 3 ms difference will surpass the noise of trial-by-trial variation. That is, for every subject, there is going to be a difference of around 3 ms. Thus, when doing a paired t-test with lax versus aspirated (the actual subjects-analysis), the 3 ms difference is consistent enough to reach the threshold of significance. Moreover, to the researcher's eye it will look like a remarkably consistent effect because many subjects show it. But that is only because all subjects were presented with the same unlucky choice of items. Or, in other words: Having repetitions does, in fact, increase statistical stability… however, not of the estimates of the effect (which does not exist in the simulated population), but of the idiosyncratic differences between items.

In this particular case, an items-analysis (see suggestions by Clark [6]) would help, since for only four items, any inferential test with one mean per item is unlikely going to be significant. And indeed, in the simulation, Type I errors for the items-analysis "stay put" at around 0.05, unaffected by the number of repetitions or items, as can be seen in Table 2.

Table 2: Type I error rates of items-analyses (between items, unpaired t-test) for phonetic experiments with 2, 4, 8, or 16 items and 2, 4, 8, 16 or 32 repetitions. Cells give proportions of significant results ($p<0.05$) out of 1,000 simulations. These proportions are averaged over different analysis choices (see main body of text). The first row is by definition zero because no statistical test can be conducted for just two items.

Items	R=2	R=4	R=8	R=16	R=32
2	0.00	0.00	0.00	0.00	0.00
4	0.03	0.03	0.03	0.02	0.04
8	0.04	0.03	0.04	0.04	0.06
16	0.05	0.04	0.04	0.05	0.05

The preceding discussion shows that for an analysis that is technically valid (the subjects-analysis) Type I error rats are higher than we would like them to be. Only performing an items-analysis safeguards the researcher from drawing erroneous conclusions in this situation.

5. A FINAL CONCERN: ECOLOGICAL VALIDITY

A final reason to be concerned about repetitions has to do with ecological validity. There is a strong desire to produce results under controlled laboratory situations that nevertheless carry over to behavior outside the lab [11]. In line with this, it is a desirable goal that phonetic experiments, as much as possible, mirror speech production and perception as it happens in the real world. Although it is a matter of continuing debate as to how much controlled laboratory situations are desirable or harmful for developing phonetic theories (see, e.g., [28]), it is clear that designing phonetic experiments with less repetitions would reduce the gap between laboratory speech and "real" speech: People in the real world do not repeatedly utter the same word over and over again, except in some very special circumstances. On average, an experiment with a large number of repetitions is going to be less reflective of real speech than an experiment with fewer repetitions. Thus, having a disproportionate number of repetitions works against the desire to have phonetic laboratory experiments as ecologically valid as possible.

6. RECOMMENDATIONS

The present paper questioned whether experiments with many repetitions should be a preferred design choice for phonetic experiments. In particular, it was argued that averaging over repetitions implicitly ignores the inter-dependent and non-normal nature of repetitions. Then it was shown that even if repetitions are independent and normally distributed (contra to fact), having many repetitions increases Type I error rates when only a subjects-analysis is conducted. Finally, repetitions tend to decrease the ecological validity of phonetic experiments, since speakers in real communicative contexts rarely repeat the exact same word multiple times in a row. On the other hand, it was pointed out that it is crucial to think about how one can generalize over a set of language items, and not just a set of speakers [cf. 6]. Thus, "the other N," the number of unique items, should perhaps be given more weight at the design stage of phonetic experiments.

Clear recommendations follow from this discussion:

- phonetic experiments should have as few repetitions as possible
- phonetic experiments should have as many distinct items as possible

The argument presented in this paper suggests that following these recommendations, as much as possible given the constraints of a particular study, will lead to more generalizable results and to more confident statistical estimates.

7. REFERENCES

[1] Aylett, M., Turk, A. 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language & Speech* 47, 31–56.

[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, 390–412.

[3] Barr, D. J., Levy, R., Scheepers, C., Tily, H. J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* 68, 255–278.

[4] Bates, D., Maechler, M., Bolker, B., Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7.

[5] Broad, D. J., Clermont, F. 2014. A method for analyzing the coarticulated CV and VC components of vowel-formant trajectories in CVC syllables. *Journal of Phonetics* 47, 47–80.

[6] Clark, H. H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior* 12, 335–359.

[7] Gracco, V. L., Abbs, J. H. 1986. Variant and invariant characteristics of speech movements. *Experimental Brain Research* 65, 156–166.

[8] Gregory, M., Raymond, W., Bell, A., Fosler-Lussier, E., Jurafsky, D. 1999. The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society* 35, 151–166.

[9] Johnson, K., Ladefoged, P., Lindau, M. 1993. Individual differences in vowel production. *Journal of the Acoustical Society of America* 94, 701–714.

[10] Kello, C. T., Anderson, G. G., Holden, J. G., Van Orden, G. C. 2008. The pervasiveness of 1/f scaling in speech reflects the metastable basis of cognition. *Cognitive Science* 32, 1217–1231.

[11] Kingstone, A., Smilek, D. & Eastwood, J. D. 2008. Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology* 99, 317–345

[12] Kirby, J. 2013. The role of probabilistic enhancement in phonologization. In Yu A. (ed.), *Origins of Sound Change: Approaches to Phonologization*. Oxford: Oxford University Press, 228–46.

[13] Kroodsma, D. 1989. Suggested experimental designs for song playbacks. *Animal Behavior* 37, 600–609.

[14] Kroodsma, D. E. 1990. Using appropriate experimental designs for intended hypotheses in 'song' playbacks, with examples for testing effects of song repertoire sizes. *Animal Behavior* 40, 1138–1150.

[15] Labov, W. 1994. *Principles of Linguistic Change, Volume 1: Internal Factors*. Cambridge: Blackwell.

[16] Liberman, A., Franklin, C., Shankweiler, D., Studdert-Kennedy, M. 1967. Perception of speech code. *Psychological Review* 74, 431–461.

[17] Mesirov, J. P. 2010. Computer science. Accessible reproducible research. *Science* 327, 5964.

[18] Peng, R. D. 2011. Reproducible research in computational science. *Science* 334, 1226–1227.

[19] Pierrehumbert, J. 2002. Word-specific phonetics. *Laboratory Phonology VII*, Mouton de Gruyter, Berlin, 101–139.

[20] Pinheiro, J. C., Bates, D. M. 2000. *Mixed-effects models in S and S-PLUS*. New York: Springer.

[21] Pluymaekers, M., Ernestus, M., Baayen, R. H. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62, 146–159.

[22] R Core Team 2014. R: A language and environment for statistical computing (version 3.1.0). R Foundation for Statistical Computing, Vienna, Austria.

[23] Schielzeth, H., Forstmeier, W. 2009. Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology* 20, 416–420.

[24] Silva, D. J. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23, 287–308.

[25] Wang, W. S.-Y. 1969. Competing changes as a cause of residue. *Language* 45, 9–25.

[26] Winter, B. (2011). Pseudoreplication in phonetic research. *Proc. 17th ICPhS* Hong Kong, 203–206.

[27] Lancia, L., Winter, B. 2013. The interaction between competition, learning and habituation dynamics in speech perception. *Laboratory Phonology* 4, 221–257.

[28] Xu, Y. 2010. In defense of lab speech. *Journal of Phonetics* 38, 329–336.