This manuscript was submitted to the *Journal of the Acoustical Society of America* on February 28, 2022 and is currently in a second round of revisions. The published version will have some small changes.

This version reflects changes through October 13, 2022

# Sample Size Matters in Calculating Pillai Scores

Joseph A. Stanley,[1] and Betsy Sneller[2]

*[1] Linguistics Department, Brigham Young University, Provo, Utah 84602, United States*

*[2] Department of Linguistics, Languages, and Cultures, Michigan State University, East Lansing, Michigan 48824, United States*

## Abstract

Since their introduction to sociolinguistics by Hay et al. (2006), Pillai scores have become a standard metric for quantifying vowel overlap. However, there is no established threshold value for determining whether two vowels are merged, leading to conflicting ad hoc measures. Furthermore, as a parametric measure, Pillai scores are sensitive to sample size. In this paper, we use generated data from a simulated pair of underlyingly merged vowels to demonstrate (1) larger sample sizes yield reliably more accurate Pillai scores, (2) unequal group sizes across the two vowel classes is irrelevant in the calculation of Pillai scores, and (3) it takes many more data than many sociolinguistic studies typically analyze to return a reliably low Pillai score for underlyingly merged data. We provide some recommendations for maximizing reliability in the use of Pillai scores, and provide a formula to assist researchers in determining a reasonable threshold to use as an indicator of merged status given their sample size. We demonstrate these recommendations in action with a case study.

# I. INTRODUCTION

Quantifying vowel overlap is an important component of many linguistic studies, ranging from sociolinguistics to laboratory phonology. While various methods of quantifying vowel overlap have been proposed (see, e.g., Nycz & Hall-Lew 2013), Pillai scores have emerged in recent years as the most commonly used method, particularly within sociolinguistics (Hay, Warren & Drager 2006; Nycz & Hall-Lew 2013), because of its ability to measure a distinction in multivariate space while also accounting for fixed effects like phonological context. However, there is no standard value of Pillai score that is broadly accepted to be a threshold for "merged" or "distinct"; indeed, individual studies make this determination primarily by comparison between individuals in a single data set, to show that some speakers are "more merged" while others are "less merged". In this paper, we provide a critical look into how Pillai scores are calculated and reported.

We focus on demonstrating how sample size plays a major role in the resulting Pillai score. We also show that the common approach of reporting Pillai scores alone is incomplete without also reporting sample sizes and *p*-values. Using simulation data drawn from an underlyingly merged data set, we show how larger samples produce lower Pillai scores (in other words, a more "merged" score). We further demonstrate that within Pillai, it is the total *n* across both samples that matters, and provide a formula that researchers can use to determine a threshold for "merged" given their own sample size. We highlight some important takeaways about using Pillai scores to measure vowel overlap and the potential risks for across- and within-study comparisons of speakers. We end with a case study which demonstrates how researchers can implement the recommendations in this paper when analyzing real data from sociolinguistic interviews.

## A. Pillai as a measure of vowel overlap

### *1. Pillai score overview*

A multivariate analysis of variance (MANOVA) is an extension of the (univariate) analysis of variance (ANOVA). The difference is that while an ANOVA evaluates whether the difference between two or more groups in a single numeric variable can be predicted by some number of categorical independent variables (such as the F2 of /u/ across older and younger participants), a MANOVA can evaluate two or more dependent numeric variables simultaneously (such as F1, F2, F3, and duration of /u/ by older and younger participants). So, a researcher analyzing American English vowels (which typically are differentiated primarily on vowel quality in F1-F2 space) could use a MANOVA to see whether a speaker pronounces two historically distinct vowel classes differently, while including the effects of duration and place of articulation as independent variables. In this case, F1 and F2 would be the dependent variables, and vowel class, duration, and place of articulation would be the independent variables.

Simplified somewhat, the null hypothesis of a MANOVA is that category membership (for instance, two historically distinct vowel classes) offers no explanatory power for any of the dependent variables. In the case of a MANOVA that is fit to vowel data, the null hypothesis is that a merger is present. In other words, there would be no way to guess which historic vowel class a particular token came from by its acoustic measurements alone. Typically, the researcher's aim is to find evidence to reject that hypothesis. We note that high *p*-values associated with MANOVAs only indicate a lack of evidence to reject the null hypothesis that the two vowel classes are the same, rather than evidence *for* the null hypothesis. MANOVA cannot prove that two vowels are merged, only that there is little evidence to suggest that they're distinct.

There are four main test statistics associated with a MANOVA to compare what the data shows to the null hypothesis: Wilk's lambda, the Lawley-Hotelling trace, Roy's largest root, and the

Pillai-Bartlett trace. Of these four, the lattermost is the most robust for non-normally distributed data and other violations of the assumptions of a MANOVA for tests that compare more than two groups (Olson 1976). In sociophonetic data, where it is more common to compare only two groups, the Pillai-Bartlett trace has less advantage over the other three in statistical validity, but does benefit from being both easy to run in commonly used statistical environments and relatively easy to interpret. For introductory overviews of these four test statistics, including their mathematical definitions and conceptual explanations, see Bray & Maxwell (1985: 27–29), John & Wichern (2012: 336), Rencher & Christensen (2012: 169–188), and Upton & Cook (2014, "multivariate analysis of variance (MANOVA)").

The Pillai-Bartlett trace, often called the *Pillai score* or occasionally just *Pillai* in linguistics studies (a convention that we adopt here), ultimately comes from Pillai (1955) and Bartlett (1939). It returns a value that ranges between 0 and 1, with smaller values occurring when there is greater overlap between the two groups in multivariate space, and larger numbers for less overlap. In other words, small Pillai scores suggest a vowel merger. In reality, determining whether a merger is present is not quite as simple as merely observing overlap. Two vowels may occupy the same F1-F2 space but the distinction between phonemes may be maintained though some other cue like voice quality (Di Paolo & Faber 1990), duration (Labov & Baranowski 2006), or vowel trajectory (Stanley 2020). We adopt a simplified approach here since our focus is to explore the effects of sample size, but we acknowledge that there is more to merger than overlapping midpoints in the F1-F2 space.

## *2. Meta-analyses of Pillai scores vs. other metrics*

Prior to the introduction of Pillai scores to sociophonetics, perhaps the most common way to assess merger was through auditory coding. In its most basic form, the phonetically-trained researcher would listen and evaluate whether there was a difference between the two sounds. But, as shown below, Pillai scores are most commonly used on low vowels, which are notoriously difficult for fieldworkers to accurately transcribe (Johnson 2010: 28–29). In fact, Moulton (1968: 464) rather strongly states that early fieldworkers for the Linguistic Atlas Projects were "hopelessly and humanly incompetent at transcribing phonetically the low and low back vowels that they heard from their informants." Fortunately, formulaically quantifying overlap provides less subjectivity that earlier researchers lacked.

While Pillai scores are currently the most common metric for quantifying mergers, especially within sociolinguistic work, there are also other approaches available. In addition to the auditory coding mentioned above, much early work used the Euclidean Distance between the point representing the mean F1 and mean F2 of one vowel class and the point representing the mean F1 and mean F2 of a second vowel class as the primary metric for vowel merger (Hay, Warren & Drager 2006; Nycz & Hall-Lew 2013; Han & Kang 2013; Hall-Lew 2013). This early approach is not particularly satisfying, as it fails to take into account the distributional properties of the data, including the degree of overlap and the distribution of tokens within a vowel class (Kelley & Tucker 2020: 137). The Spectral Overlap Assessment Metric (SOAM; Wassink 2006), which calculates overlap between ellipses or ellipsoids fitted to the vowel distribution (see Wassink 2006 for details on the fitting) and calculates the area or volume of their overlap, is one method adopted and recommended in other sociophonetic work (Di Paolo, Yaeger-Dror & Wassink 2011: 103; Kendall & Fridland 2021: 56). We refer interested readers to Nycz and Hall-Lew (2013) and Kelley and Tucker (2020) for in-depth assessments of these measures and several others, compared with Pillai scores. Kelley and Tucker assess four different metrics of vowel overlap, using a Monte Carlo simulation to test accuracy, and find that Pillai scores produced the most accurate and precise values when compared to ground truth values in their simulated data. They also recommend Pillai scores when sample sizes are "small," which they define as 30 observations per group.

For sociophonetic data, and especially for naturalistic sociophonetic data, obtaining more than 30 observations per group is not always feasible, making Pillai an especially valuable tool for sociophonetics. As perhaps foregrounded by the title of this current paper, sample size plays an important role in the resulting Pillai score, which we highlight in detail in Sections 2–3. We note, however, that this is generally true of all measures of vowel overlap compared in the overview papers mentioned above (Nycz and Hall Lew 2013; Kelley and Tucker 2020). Indeed, any measure that uses means, standard deviations, and/or variance as inputs will be impacted by sample size. Likewise, larger sample sizes impact statistical significance, with larger sample sizes leading to smaller *p*-values. Our aim here is not to simply recommend that researchers obtain larger sample sizes, which in many cases is either not possible or may be at odds with other important data collection considerations, but rather to elucidate *how* sample size impacts resulting scores, so that researchers can be best informed when using Pillai as a measure of vowel overlap.

Perhaps an even bigger concern than total sample size with naturalistic data is the near impossibility of obtaining balanced token counts across two categories from naturalistic data. In wordlist data, researchers can carefully craft their wordlist to obtain balanced data: this typically means obtaining the same number of observations in each vowel class and ensuring that the wordlist is balanced for additional factors, such as phonological context. In naturalistic (i.e., conversational) speech, there is no way to ensure that a speaker produces balanced token counts across vowel category and across phonological contexts. As a result, researchers investigating something like vowel overlap need to understand precisely *how* Pillai scores may be affected by unbalanced data.

Finally, while Pillai has emerged as the best option so far for most sociophonetic data, there is one additional concern to address, which is that MANOVAs assume the data within each group follows a multivariate normal distribution (Bray & Maxwell 1985). Formant measurements for a given vowel from a given speaker in naturalistic vowel data, being nonnormally distributed, does not fit this assumption (though see Whalen & Chen 2019 for some evidence that vowel formant data, even with coarticulation, can be normally distributed). Recent work has suggested Bhattacharyya's Affinity[i] as a better measure for non-normally distributed data (Fieberg & Kochanny 2005; Johnson 2015), and has been taken up by some subsequent work (Strelluf 2018; Warren 2018; Jensen & Braber 2021). While the benefits of Bhattacharyya's Affinity being a nonparametric measure make it particularly appealing for the kind of distributions of data found in naturalistic speech, there is not yet a mechanism for easily incorporating fixed effects like phonological context or speaker age.[ii] Furthermore, Bhattacharyya's Affinity also works best on a relatively large data set of over 30 observations per category (Seaman et al. 1999). While future work may make nonparametric methods like Bhattacharyya's Affinity more easily integratable with the kind of statistical models linguists typically use, for now we focus on Pillai score and the considerations necessary to make its use maximally standardized.

## *3. How Pillai scores are used to measure overlap in sociophonetics*

Pillai scores have been used to analyze a variety of phenomena in several languages. Perhaps the most common application of Pillai scores is to measure overlap between /ɑ/ and /ɔ/ in North American English (Hall-Lew 2013; Kendall & Fridland 2017; Havenhill 2015; Stanford et al. 2019 inter alia). In fact, using Pillai scores to measure this merger was explicitly recommended in Becker (2019), a volume of different studies all analyzing the spread of the /ɑ-ɔ/ merger in North American English. Many conditioned mergers in English have been quantified with Pillai scores as well (Schmidt, Diskin-Holdaway & Loakes 2021; Austin 2020; Freeman 2021; Newbert 2021 inter alia). To a lesser extent, Pillai scores have been used in other languages when analyzing vowels that are marginally contrastive (Galician: Amengual & Chamorro 2015; Italian: Nadeu & Renwick 2016; Bangla: Islam & Ahmed 2020; Hawai'ian: Kettig 2021; Swiss German: Joo, Schwarz & Page 2018; Austrian German:

Sloos 2013). Some more innovative uses of Pillai scores include using them to analyze tones in varieties of Cantonese (Fung & Lee 2019; Tse 2018) and in Spanish fricative mergers (Regan 2020).

Pillai scores have also been used to quantify splits, chiefly among phonological low vowels. Fisher et al. (2015) assess the degree to which the Philadelphia short-*a* split is found in their sample of Philadelphians. Relatedly, Hall-Lew et al. (2017) look at the *bath-trap* split in Scottish Parliament data. Brozovsky (2020) uses Pillai scores to measure the raising (and separation) of prenasal /æ/, using Pillai scores to measure the overlap between prenasal /æ/ and preobstruent /æ/ in Taiwanese Texans. In a study looking at the possible effect of salience on a given lexical item compared to the rest of its canonical vowel class, Bray (2021) analyzes the lowered realization of the vowel in *hockey* compared to other relatively more raised /ɑ/ tokens in professional American hockey players.

As highlighted by the selection of citations above, since being introduced to the field, Pillai scores have become widespread in sociophonetic studies. With the support of meta-analyses that compare other competing measures, Pillai has become a useful go-to for measuring both the overlap and the distinction of speakers' pronunciation of two phonological categories, especially in sociolinguistic studies that need to compare across individual speakers. While Pillai scores are clearly a valuable tool in measuring vowel overlap, there remain some outstanding issues with using it. In the following sections, we highlight some of these issues.

## B. Issues with Pillai scores

### *4. What is considered merged?*

As useful as Pillai scores are for quantifying the degree of overlap, they do not necessarily answer researchers' underlying question of whether two vowels are merged. Pillai scores range from 0 to 1, but there is no agreed-upon cutoff value or threshold for determining whether the two groups are merged or not. As a result, many studies rely on an ad-hoc threshold to interpret the merged status of their speakers. Some work has suggested specific thresholds for mergers. For instance, Jibson (2021) suggested a Pillai threshold of 0.3 as an indicator of "merged" status, after a shuffling procedure identified 0.3 as the 95th percentile of "merged" between 20 tokens of two vowel classes from his speakers. Wassink (2006) likewise suggests some provisional thresholds for SOAM, where 0-20% overlap represents "distinct", 20-40% represents "partially merged", and >40% represents "merged". Relying on provisional or ad-hoc thresholds, however, is risky because sample size matters and it is likely not comparable across studies or even between speakers.

One solution for determining whether a given Pillai score should be interpreted as an indication of "merger" is to examine the *p*-values that are associated with the MANOVA model from which the Pillai scores are generated. The model assumes that the vowel variable contributes no information to differentiating between two groups. In other words, it assumes the two vowels are underlyingly merged. A small *p*-value associated with the vowel variable would provide evidence against that null hypothesis, allowing the researcher to conclude that the difference between the two groups is likely true (i.e., that the speaker is not merged). We note that Pillai scores and *p*-values are inversely correlated: lower Pillai scores typically accompany higher *p*-values. In fact, since Pillai scores are just test statistics, they and *p*-values are functions of each other. Nevertheless, *p*-values are not typically reported in sociophonetic studies that use Pillai (some exceptions include Wong & Hall-Lew 2014; Nadeu & Renwick 2016; Amengual & Chamorro 2015; Berry 2018; Sloos 2013).

There are a few points of caution to make about using and interpreting *p*-values, as we discuss throughout this paper. One of these concerns the potential distinction between a statistically significant difference (as defined by the model) and a ground truth difference for speakers. For a speech community that has a ground truth merger in two vowel categories, there will be no difference in their perception of these two vowel categories. However, as sample size increases, so does the likelihood of a model returning a *p*-value below a given significance threshold (typically 0.05); it is possible that even for pairs of sounds that are truly merged, a sufficiently large dataset can interpret random variance in the data as a meaningful difference, as shown in the experiments in Section 3. An additional caution to make regarding interpreting *p*-values alone as an indicator of merger is that a statistically significant difference in formant values may not map onto a perceptible difference for the human auditory system (see, e.g., Kewley-Port and Watson, 1994). *P*-values alone can likewise be misleading in the opposite direction: previous work has shown that speakers and listeners can produce and perceive reliable but small differences, including sub-phonemic differences, in what may otherwise appear to be merged sounds (for instance, with cases of incomplete neutralization; Warner et al. 2004; Pfiffner 2021). Vowel distinctions that are maintained by small effect sizes, or by sub-phonemic distinctions not captured by the measurements in the model, may appear artificially to be merged according to a *p*-value because it takes more data for smaller differences to be detected by the statistical model. For these reasons, additional information such as Pillai scores can aid in the interpretation of *p*-values, and vice versa.

## *5. Sample size*

As suggested by the parametric nature of Pillai, and the discussion of ad-hoc thresholds above, a major component of deciding which threshold should be used to determine merger status is the number of tokens being analyzed. Previous work on Pillai scores and sample size have expressed concern over too-small sample sizes (Gorman & Johnson 2013), and over unbalanced sample sizes across the two vowel classes being analyzed (Nycz & Hall-Lew 2013; Johnson 2015).

Despite sample size having a major impact on Pillai scores, most studies in sociophonetics that use Pillai as a measure of vowel overlap do not also clearly report sample size (exceptions include Wong & Hall-Lew 2014; Holland & Brandenburg 2017; Berry 2018; Berry & Ernestus 2018). This in turn makes it difficult both to assess the findings in an individual paper and to compare speakers within and across studies. In the simulation experiments below, we show just how important sample size is to the resulting Pillai score, and provide a formula to calculate a recommended Pillai score threshold for merger status, given a particular sample size.

## II. METHODS

In this section, we present the results of Monte Carlo simulation experiments designed to test the effect of different sample sizes on resulting Pillai score. In these simulations, we create two vowel classes that are perfectly merged underlyingly, and alter (1) the sample sizes between the two vowel classes to test the effect of unbalanced samples across vowel classes and (2) the overall sample size, considering both vowel classes together, to test the effect of unbalanced total sample size across speakers.

### A. Data generation

We generated data using a Monte Carlo simulation (Metropolis & Ulam 1949). This is a procedure where random draws are taken from an underlying probability distribution or existing dataset and analyzed. This process is repeated independently many times and the information about each iteration is aggregated. To begin the simulation, a bivariate normal distribution was generated in R to simulate a single theoretical underlying vowel in the F1-F2 dimension for a single theoretical speaker. For the sake of simplicity, the mean for F1 and F2 were both set to zero and the standard deviation was 1. The correlation coefficient between the formants was zero, producing a circular (rather than elliptical) distribution. Because the generated data was uncorrelated, the distribution is most easily generated by combining two independent univariate normal distributions because the product of their probability densities is equal to their joint probability density (Johnson & Wichern 2012 chapter 4 page 151). Specifically, in R, we generated F1 by running `rnorm(x, mean = 0, sd = 1)`, where `x` is the number of tokens generated. F2 was generated using the same code, and the two sets of numbers were combined to create the bivariate normal distribution.

We acknowledge that this generated data deviates from real vowel data in several ways. First, the units here were standardized rather than simulating formant frequencies, though this approach can also be thought of as simulating a *z*-scored (Lobanov) normalized vowel space. Vowel data is often not normally distributed (though, see Whalen & Chen 2019), so this distribution is not necessarily representative of actual acoustic data. Even for vowel data that does appear be multivariate normally distributed, there is often some degree of correlation in F1 and F2 (seen as elliptical distributions in the F1-F2 space). However, since the focus of this study is on how sample size affects Pillai, rather than another variable like skewness, we wanted to keep the distribution as simple as possible. In the future we hope to replicate this study with an underlying distribution that is based on formant measurements that reflect naturalistic speech.

To simulate vowel data, random draws were taken from that single bivariate normal distribution, and assigned to one of two "vowel class" labels. We note that it is somewhat nonsensical to refer to these generated numbers as "vowels", especially since Pillai scores can be calculated on non-vowel data. However, since the majority of Pillai scores in sociophonetic analyses are based on vowels, for clarity, we will refer to this simulated data points as "vowels" and their arbitrary groups as "vowel classes." These random draws represent a linguist sampling data from our theoretical speaker. For the sake of illustration, we will say 30 such observations were generated. These 30 observations were treated as tokens from a single underlying vowel class.[iii] Another 30 random draws were then taken from the same bivariate normal distribution. These 30 observations were treated as tokens from a different underlying vowel class. Generating two groups from the same underlying distribution therefore creates a simulated pair of merged vowels. In theory, the two simulated vowel classes should not be statistically different from each other in any way because they were drawn from the same underlying distribution.

### B. Two experiments

For this study, we ran two experiments. In Experiment 1, the two simulated vowel classes for each "speaker" were of equal size. We began with a sample size of 5 observations per vowel class. We then moved on to 6 observations per vowel class, and so on, until we reached 100 observations per vowel class. For each of these 95 sample sizes, we repeated the simulation 1000 times, each representing a different instance of a linguist sampling data from that one underlyingly merged speaker. This produced 96,000 pairs of simulated vowel data, where each pair consisted of equal-sized vowel classes, enabling us to test the effect of overall sample size on resulting Pillai score.

In Experiment 2, we varied the sample size between the two vowel classes for each "speaker". We began with 5 tokens from one vowel and 6 from another. We then took 5 tokens of one and 7 from the other. We increased the size of the second group by steps of 1 until it contained 100 tokens. We then repeated this process with the first group having 6 tokens, and increased the second group from 5 to 100 in steps of 1. We iterated over these steps, increasing the sample size of the first group up to 100, thereby generating pairs of vowel data where every combination of sample sizes from 5 to 100 was represented. We repeated this simulation 100 times per combination. This produced 921,600 pairs of simulated vowel data, where each pair consisted of different-sized vowel classes, enabling us to test the effect of unbalanced vowel class size on resulting Pillai score.

Across both experiments, Pillai scores were calculated for each pair of simulated vowel data. Pillai scores were calculated by fitting a MANOVA model to the data using the `manova()` function in R. The simulated F1-F2 measurements were the dependent variables and the vowel class was the only independent variable. While it would be possible to incorporate additional simulated independent variables such as place of articulation and duration into the MANOVA, we consider this to be beyond the scope of the current paper, which focuses on the effect of sample size on resulting Pillai scores. We therefore include only historical vowel class in our models, and leave more complex MANOVA models to future work. The Pillai scores and *p*-values associated with the vowel class variable were then extracted from that MANOVA model. To reiterate, the Pillai scores for all of these distribution should be very close to zero (indicating complete overlap) because every data point was generated from the same underlying bivariate normal distribution. Because the data is randomly generated, some Pillai scores will be higher than others, but by rerunning the simulation many times per sample size, we can begin to see patterns that may emerge at a given sample size.

The coding for this study was done in the R programming language (R Core Team 2021) with the help of the `tidyverse` suite of packages (Wickham et al. 2019) and `joeyr` (Stanley 2021). Visualizations were generated using `ggplot2` (Wickham 2015) and `see` (Lüdecke, Patil, et al. 2021) with color palettes from `ggthemes` (Arnold 2018) and `scico` (Pedersen & Crameri 2020).

## III. RESULTS

### A. Experiment 1: Equal sample sizes

To address how sample size affects Pillai scores, we first present the results from Experiment 1, where the two simulated vowel classes were the same size. Before inspecting the results of all sample sizes though, it is important to understand how the 1000 Pillai scores were distributed within a given sample size. Figure 1 shows two different views of the distribution of Pillai scores when the sample size for both groups was 10. We see that the distribution of points representing resulting Pillai scores is rather wide, a consequence of using such a small sample size for inferential statistics, ranging from less than 0.001 to 0.568. Much of the data is clustered near the bottom of the plot but there is a long "tail" extending upwards. This is not a haphazard pattern, but rather follows a distribution that can be easily transformed into an *F* distribution and reflects the underlying mathematical properties of how Pillai scores are calculated (cf Rencher & Christensen 2012: 182). For this particular sample size, the mean Pillai score was 0.104, the median was 0.077, and the 95th percentile was 0.294. As seen below, these numbers change depending on the sample size, but the underlying distribution of the Pillai scores is consistent across sample sizes. We show this distribution to dispel any misconceptions that Pillai scores are uniformly distributed within a particular range, and to highlight that generally they fall near the lower end of the distribution.

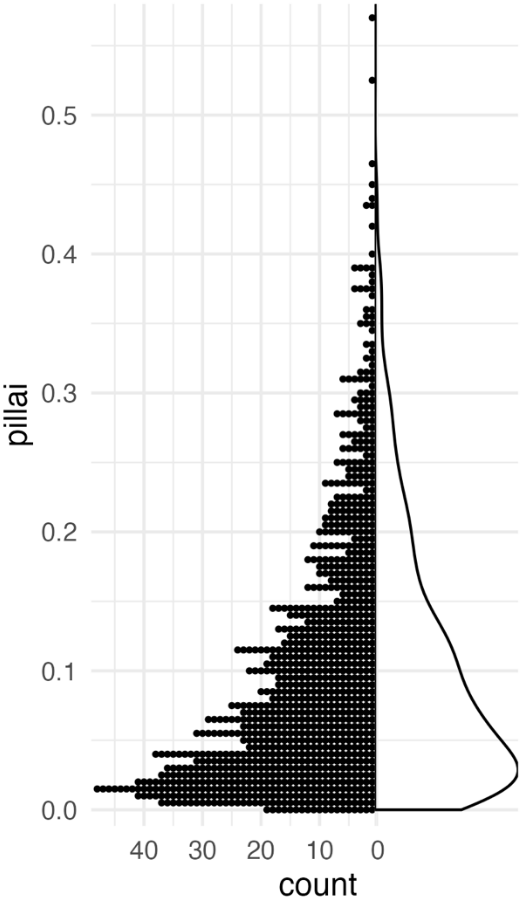

Figure 1: Distribution of Pillai scores on 1000 pairs of simulated groups, each with a size of 10.

With that distribution in mind, we can now zoom out to view all samples at once. Figure 2 shows all 96,000 Pillai scores by their sample size. Though present at all sample sizes, the "bottom-heavy" distribution shown in in Figure 1 is not displayed in Figure 2, in order to make the general trend across samples easier to see. It is immediately apparent that larger groups more consistently produced lower Pillai scores. With very small sample sizes (groups of fewer than 10 observations per vowel class), Pillai scores were quite high. For these small sample sizes, Pillai scores were sometimes closer to 1 than 0, even for these underlyingly merged vowel classes which should, in principle, return a Pillai score of 0. In other words, these small sample sizes sometimes resulted in very misleading Pillai scores that may cause a researcher to interpret two vowel classes as distinct even if the true underlying distribution was perfectly merged. As the sample size increases, Pillai scores were more consistently low, as we would expect for underlyingly merged vowels.

The black line overlayed on Figure 2 indicates the 95th percentile for each sample size. This line also very closely corresponds to the threshold for vowel class being statistically significant in the MANOVA models: almost all points above that line had *p*-values less than 0.05 while almost all points below it had greater *p*-values. Because we are modeling the null hypothesis (*i.e.* underlyingly merged vowel classes), the distribution of *p*-values is uniform. It therefore is unsurprising, and in fact expected, that the highest 5% of Pillai scores within a given sample size also return *p*-values less than 0.05. This line shows that if there are just 10 observations per vowel class, 95% of the Pillai scores were under 0.3. However, as is evident in Figure 2, this threshold is only applicable to two groups of 10 since Pillai scores decrease with larger samples. For example, the 95th percentile of returned Pillai scores does not drop to 0.1 until there are 30 observations per group.

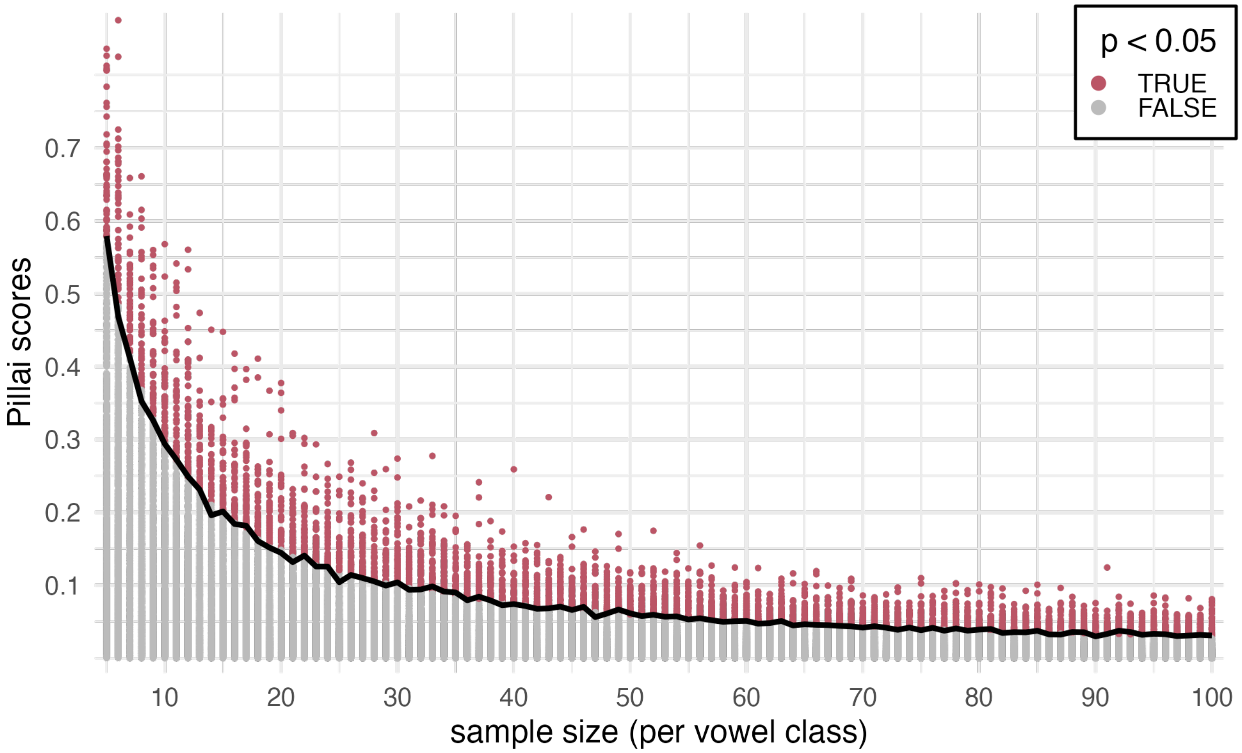

Figure 2 (color online): Pillai scores for all simulations of equally-sized groups simulations, by sample size.

### B. Experiment 2: Unequal sample sizes

While the previous section found that sample size affects Pillai scores in a predictable way, with larger samples returning more reliable Pillai scores, in this section we conduct further simulations to explore what effect, if any, an unbalanced sample has on Pillai scores. Unless data collection is carefully controlled to include a fixed number of tokens per vowel class, Pillai scores are run on vowel classes that are not comprised of the same number of tokens. Here we ask: what effect do grossly unbalanced groups have on Pillai scores?

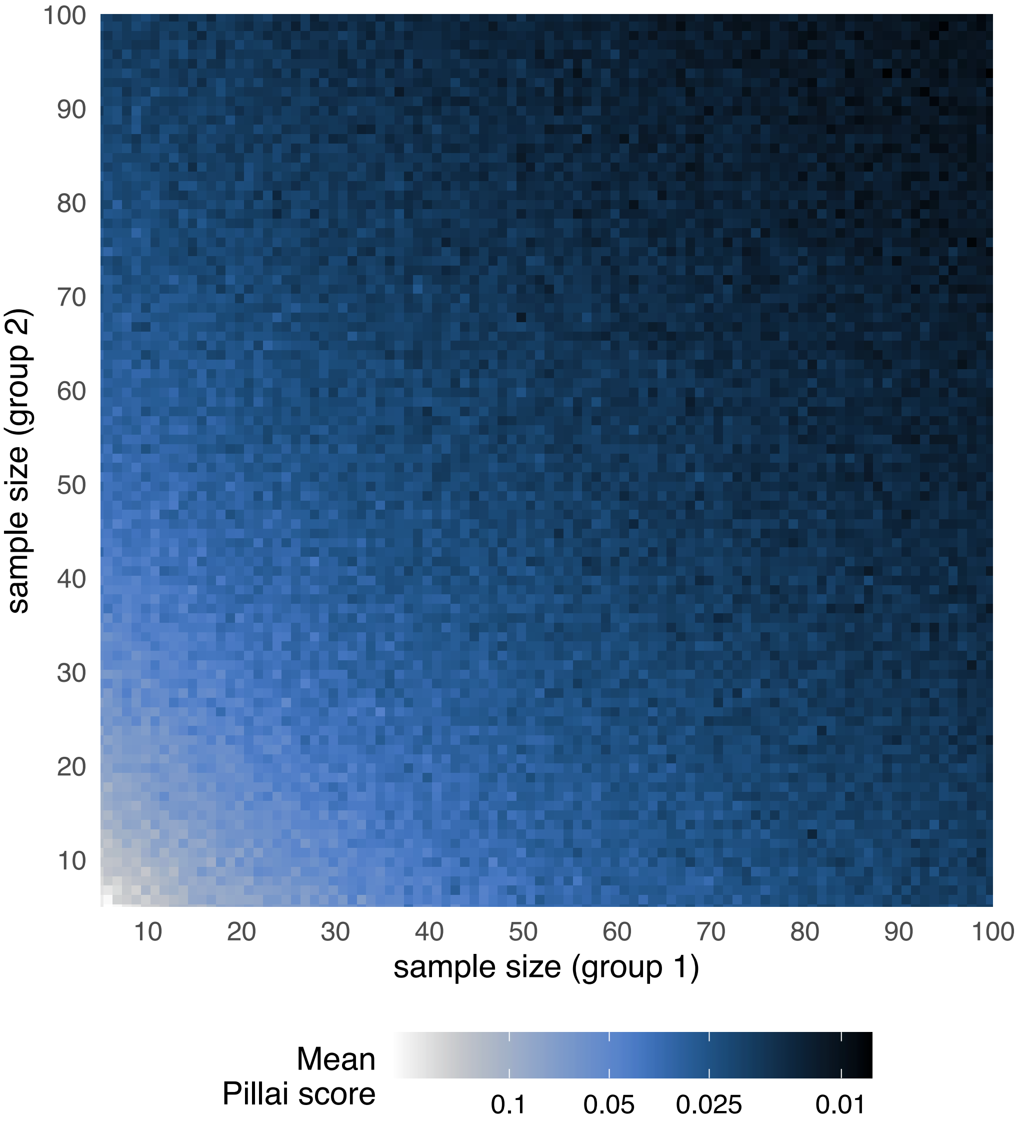

Figure 3 (color online): Mean Pillai scores for all simulations.

Figure 3 shows the mean Pillai scores from the 100 simulations for each combination of sample sizes between the two vowel classes. Going from the bottom left corner (two small sample sizes from each vowel class) to the top right corner (two large sample sizes from each vowel class), we see a general trend of decreasing Pillai scores as sample sizes increase. Reflecting Figure 2, we see these decreasing Pillai scores dropping more sharply as sample sizes are small (under around 30 tokens per vowel class).

A surprising pattern emerges when we look closely at the resulting Pillai scores from unequal sample sizes: namely, that unequal samples across the two vowel classes do not impact Pillai score—it is only the *total* sample size taking both vowel classes together that matters. For example, the mean Pillai score for a pair of 10 tokens and 50 tokens drawn from this underlyingly merged distribution is 0.0358 and the mean Pillai score for a pair of 20 and 40 tokens is 0.0355. A pair of 30 tokens and 30 tokens drawn from an underlyingly merged distribution is nearly identical: 0.0381. We see this pattern visually reflected along the diagonal between the top left corner and the bottom right corner of Figure 3, which shows a symmetrical resulting Pillai score for all unequal pairs of samples that sum to the same total. In other words, we find that neither the existence of an unequal sample size between vowel classes nor the degree of unequalness impact Pillai score.

These findings bring us to recommend simply using as many tokens as researchers have available for an individual speaker, to bring the *total* sample size as high as possible. When comparing across speakers, it is important to recall that total sample size impacts the resulting Pillai score, and we recommend normalizing and reporting total sample sizes across speakers in a study to make resulting Pillai scores maximally comparable.

## IV. IMPLICATIONS

### C. Choosing a threshold based on sample size

As a result of these simulations, we are in a position to recommend a formula that researchers can use as a guide to determine whether a given resulting Pillai score indicates a merger. We sought a formula that was a function of sample size and could provide the 95th percentile for Pillai scores given that two groups come from the same underlying distribution. In other words, we wanted to find the function for the black line in Figure 2. We chose the 95th percentile for this distribution to align with the common benchmark of $p < 0.05$.

The formula is based on the observation that by taking the natural log of the Pillai score and the natural log of half of the total sample size of both groups (essentially log-transforming the axes in Figure 2), the 95th percentile per sample size followed a straight line with an intercept of 1 and a slope of –1. The formula therefore begins as *pillai* = 1 – *n*. But because this straight line only makes sense when both *x* and *y* are log-transformed, it must be modified to be log(*pillai*) = 1 – log(*n*). At this point, we solve for *pillai*, using $e^x$ as the antilog function, yielding $p_{95} = e^{1-log\,(\frac{n}{2})}$. Reducing this equation[iv] produces the formula that we recommend for determining a Pillai score cut off,

$$p_{95} = \frac{2e}{n} \tag{1}$$

where $p_{95}$ is the 95th percentile of Pillai scores given a total sample size *n*. The curve that is generated by this formula very nearly follows the black line in Figure 2, which represents the 95th percentile of Pillai scores given the *total* sample size (summing the samples from each group together). To ensure the validity of the formula, we reran the entire simulation 100 more times and found that in every case the formula is a very good approximation of the 95th percentile for Pillai scores. While we do not have the mathematical expertise to prove whether Equation 1 represents the true underlying distribution from which the 95th percentiles are drawn, we believe it is close enough to be fruitfully used to determine a reasonable cutoff for “merged” status for a given total sample size.

Using Equation 1, we see that, for example, a sample of 20 total observations would produce a Pillai score less than or equal to 0.2718 95% of the time if the two vowel classes were underlyingly merged (we note that this is a close approximation to the cutoff of 0.3 that Jibson (2021) chose for his total sample size of 20). For illustration, Table 1 presents threshold suggestions drawn from this formula for different total sample sizes, highlighting that it takes a great deal of data to reliably return

Pillai scores that are close to zero (recall that Pillai scores closer to zero reflect a more "merged" production). We recommend that researchers use Equation 1 in conjunction with *p*-values to make a more informed decision about the merged status of two vowel classes rather than an ad hoc or arbitrary cutoff.

Table 1: Pillai thresholds at various sample sizes based on Equation 1.

| Total sample size | Pillai threshold for "merged" |
|---|---|
| 20 | 0.2718 |
| 40 | 0.1359 |
| 60 | 0.0906 |
| 80 | 0.0680 |
| 100 | 0.0544 |
| 120 | 0.0453 |

As pointed out to us by a reviewer, one potential caveat for Equation 1 is that it is based on uncorrelated data, as we mention in Section II.A above. Recall that the formants were generated independently of each other and the correlation coefficient between them was zero, producing a circular rather than elliptical distribution. Since vowel formant data is typically correlated, the utility of this function is limited if it is not applicable to real data. To evaluate whether the correlation of the vowel formants affects Pillai scores, we generated yet another set of simulated data. We followed the methods described in Experiment 2 sample sizes ranging from 5 to 100 tokens per vowel, though for the sake of creating a manageable amount of data, each vowel's sample size was limited to multiple of 5. This time though, we used the `mvrnorm()` function in the MASS package (Venables, Ripley &

Venables 2002) to generate correlated bivariate normal data, with correlation coefficients ranging from 0 to 0.9, in intervals of 0.1. We generated 100 vowel pairs per combination of sample sizes and correlation coefficients, resulting in 361,000 new sets of simulated vowel data. We again calculated the Pillai score for each vowel pair and then ran a linear regression model on these Pillai scores, with the log-transformed Pillai score as the dependent variable and the log-transformed total sample size and the correlation as predictors. We find that correlation was not a significant predictor in these 361,000 simulations, meaning there was no significant difference between the Pillai scores of uncorrelated data and correlated data, given a particular total sample size. We also find that unbalanced data did not affect Pillai scores, even in these correlated datasets. Thus we can be confident that Equation 1 is applicable even to correlated, unbalanced data like what is typically found in real vowel formant measurements.

### D. Sample size matters across speakers but not across vowels

One important takeaway from these simulations is that it takes a relatively large amount of data to reliably (meaning 95% of the time) return a low Pillai score such as 0.1 from two underlyingly merged vowels. In an analysis of English /ɑ/ and /ɔ/, for example, conversational data can typically provide a sufficient number of observations for a robust analysis of overlap. However, few studies that analyze wordlist data contain many more than 10 tokens of these two vowels. And even within long-form conversational data, if the research question focuses on an infrequent phonological variable, the total sample size quickly drops. Because total sample size has a major impact on the resulting Pillai score, we recommend researchers choose a relevant threshold for "merger" status, based on their total sample size, and use it in conjunction with the resulting *p*-value to make a determination about merger status for their data.

Perhaps the most surprising takeaway from these simulations, for both authors, was that although Pillai is not a nonparametric test, it does not actually matter if the token counts across the two categories being investigated are unequal. Instead, the most critical consideration is the total number of tokens, summed across both categories. This is particularly important for naturalistic sociolinguistic work, which relies on casual conversation rather than carefully constructed word lists for data, meaning that is it often not possible to obtain balanced token counts across categories. Following the results of Experiment 2, we can reassure researchers that unbalanced tokens across vowel classes will not impact the resulting Pillai score. Instead, and following the results of Experiment 1, we recommend using as many total tokens possible for an analysis of a single speaker, regardless of unbalanced samples across vowel classes for that speaker.

At the same time, any study aiming to compare the "merger" status of multiple individual speakers should take into account the total sample sizes for each speaker, and especially consider the fact that sample sizes (and therefore, the interpretation of resulting Pillai scores) may be different across speakers. One of the primary goals for a robust measure of vowel overlap in sociolinguistics has been to track the development of vowel mergers and splits across speakers in a given corpus (Nycz & Hall-Lew 2013; Johnson 2010; Strelluf 2018; Labov et al. 2016) to analyze the trajectory of large-scale language change over time. Because distributions with lower token counts produce inflated Pillai scores, this means speakers who are less talkative will artificially appear to have more distinct vowels than speakers who are very talkative. We recommend using one of two ways to account for unequal sample sizes across speakers.

One option is for researchers to conduct an analysis of individual speakers, incorporating all the relevant pieces of evidence (the recommended Pillai threshold given an individual speaker's sample size, Pillai scores, *p*-values, and visualizations), to make a determination about the merged status of each speaker. Section V.A presents an example of this method in action, where we demonstrate how we leveraged all these pieces of evidence together to try to understand the merged status of each speaker. This method allows researchers to obtain a fairly robust understanding of individuals as they compare to each other and across styles. However, since this method requires researchers to synthesize a number of gradient measures into discrete categories (such as “merged”, “ummerged”, and perhaps “partially merged”), it makes it more difficult to track fine grained changes in the development of a merger across a large speech community over time .

For researchers aiming to analyze fine grained differences over real or apparent time across many speakers, however, it may be more beneficial to keep Pillai as a gradient measure. To analyze Pillai across many speakers at once, we recommend a second approach to account for sample size, which we present two options for. The basic goal for both options is to make Pillai scores comparable across speakers. This can be done either by accounting for sample size in the model (see our example in Section V.B.3), or by analyzing the same number of tokens per speaker (see our example in Section V.B.4). Our recommendations for how best to account for sample size in the model and how best to analyze the same number of tokens per speaker are discussed in detail in Section V.B.3-4, and an example of the R code used to apply these recommendations is provided in the supplementary file.

A similar issue arises when comparing merger status from a single individual speaker but across multiple speech styles. In particular, sociophonetic work often compares casual speech styles like conversation to more formal speech styles, such as a minimal pair list or a reading task. Critically for Pillai scores, since lower token counts will artificially appear more distinct, speech styles such as word lists and reading tasks with lower token counts will likewise artificially appear more distinct than speech styles with higher token counts (such as casual speech). In sum, there is a strong risk that a Pillai score difference between speech styles may be interpreted as a stylistic difference (whereby speakers appear to be maintaining a distinction in wordlists that they do not maintain in casual speech), when the effect is entirely driven by differences in total sample sizes across the two styles. It’s not uncommon, for instance, to hear of speakers apparently undoing mergers in read speech in comparison to their casual speech (cf. Labov 1994: 80; Berry 2018; Berry & Ernestus 2018); while Labov (1994) correctly points out that apparent unmergers in careful speech may be speakers hypercorrecting in response to orthographic differences across historically distinct vowel classes, we can add that distinctions measured by Pillai score will additionally be impacted by sample size. We urge researchers who use Pillai as a metric of overlap to pay close attention to differences in total sample sizes across speech style,[v] and either control for total sample size across styles or adjust their threshold of "merger" in each style accordingly, following the same recommendations provided above and elaborated on in Section V.

### E. What to report when using Pillai scores

Finally, we end with some general recommendations for what researchers should include in their results when using Pillai scores as a measure of overlap. First, because total sample size impacts Pillai score so strongly, we recommend always reporting total sample size per speaker (or per style, for studies that compare merger across speech styles). This practice will have the added benefit of enabling better comparisons across sociolinguistic studies, in turn enabling researchers to gain a clearer picture of the large-scale spread of some ongoing mergers such as the /ɑ/-/ɔ/ merger.

Second, in addition to reporting and controlling for total sample sizes, we recommend reporting *p*-values where possible. In other words, where studies report individual speakers' (or styles') Pillai scores, those should always include the total sample size and the *p*-value as well. We are not necessarily advocating for an increased reliance upon *p*-values as a binary meaningful threshold, particularly since some statisticians are urging quantitative researchers to abandon declarations of "statistical significance" and to instead understand *p*-values as one gradient measure in concert with additional evidence (Wasserstein, Schirm & Lazar 2019). However, reporting a Pillai score without its accompanying *p*-value is akin to reporting a regression estimate without a *p*-value, a *p*-value without an effect size, or a mean without a standard deviation. A Pillai score is a test statistic and should be reported as such. These two numbers in conjunction paint a better picture of the merged status of a pair of vowels than either one in isolation.

## V. A CASE STUDY

The following is a case study to illustrate how one might conduct an analysis of vowel merger using the recommendations in this paper. As illustrative data, we draw from sociolinguistic interviews conducted and analyzed by the first author with residents of southwest Washington State, a region where the low back vowels (/ɑ-ɔ/) are often merged, though with some indication of separation in some speakers (see Stanley 2020 for more details). On average across the 52 participants analyzed here, interviews lasted 46 minutes and yielded 143 tokens containing either /ɑ/ or /ɔ/ in preobstruent position. After the interviews, 30 of those participants then read a wordlist containing another 20 tokens in a more careful style.

### A. Analysis of individual speakers

We performed a MANOVA on each speaker's F1 and F2 measurements, separately for each style, with historic vowel class as the only predictor. Given the number of observations produced by each speaker, Equation 1 was implemented to establish a potential cutoff value for each style. The *p*-values and the Pillai scores were extracted from the MANOVA model and the latter were compared to the cutoff values.

In this sample, there are some cases where the evidence overwhelmingly points towards a merger. For example, 48-year-old Donna produced 179 low back vowels in the conversational portion of her interview. With that many tokens, Equation 1 suggests that if her vowels were underlyingly merged, her Pillai score should be less than 0.0304 about 95% of the time. In other words, anything less than 0.0304 would be evidence for a merger. As it turns out, the MANOVA model performed on her data yielded a Pillai score of 0.0289 with a *p*-value of 0.0756. The fact that her Pillai score is less than the threshold for her token count and that the *p*-value above 0.05 means that historical vowel class category does not predict Donna's acoustic realization, suggesting that Donna's two vowel classes are likely underlyingly merged. In the wordlist data, there were only 20 tokens, so the suggested cutoff value determined by Equation 1 is much higher at 0.2718 because of the smaller sample size. The MANOVA performed on Donna's wordlist data produced a Pillai score of 0.1648 ($p$ = 0.216). We reiterate that sample sizes also impact *p*-values, with smaller sample sizes producing higher *p*-values. Taken all together, this data makes a strong case that /ɑ/ and /ɔ/ are underlyingly merged in Donna's speech in both speech styles.

On the other hand, some speakers' data are indicative of a distinction. Kim, another 48-year-old woman, produced 137 low back tokens in the conversational portion of her interview. Equation 1 suggests that a Pillai score less than 0.0397 would be evidence for a merger, but the MANOVA performed on her data returned a Pillai score of 0.0658 ($p$ = 0.010). Though the Pillai score is relatively close to zero, we do not interpret her data as underlying merged, since with that many observations from a truly merged distribution we would expect an even lower Pillai score (less than 0.0397). Based on the 20 tokens from her wordlist, the cutoff would be 0.2718 but the MANOVA on those 20 tokens yielded a Pillai score of 0.3700 ($p$ = 0.020), further suggesting a distinction. Because Kim's Pillai scores were higher than the thresholds and were accompanied by low *p*-values for both speech styles, we conclude that Kim's low back vowels, while close in acoustic space, are not fully merged.

However, even when considering *p*-values alongside recommended thresholds, not all cases are as straightforwardly interpretable as Donna's and Kim's. Scott is a 28-year-old man whose interview contained 195 low back tokens. The Pillai score based on his data was 0.1975, far higher than the threshold (0.0279) produced by Equation 1, and was accompanied by a low *p*-value ($p$ <

0.001), suggesting two distinct underlying vowel distributions. However, the Pillai score based on the 20 tokens he produced in his wordlist (0.0397) was much lower than the threshold (again, 0.2718 ), and had a high *p*-value (*p* = 0.7090). When the Pillai score is lower than the threshold and is accompanied by a high *p*-value his wordlist data, it is tempting to interpret Scott as being a rare case of producing a merger in the wordlist that he does not produce in the conversational portion of the interview. However, because the sample size is so small in the wordlist, the Pillai score (and indeed, any result of an inferential statistical test) should be taken with a large grain of salt. We include Scott's data here in part to demonstrate that even when leveraging Pillai scores alongside a recommended threshold and a *p*-value, data with low token counts can still be difficult to interpret.

We can add one additional tool to the suite of evidence we consider when diagnosing individual speakers: a visual inspection of a plot. Figure 4 shows the distributions of /ɑ/ and /ɔ/ tokens in F1-F2 space for conversational data (left) and wordlist data (right) for both Donna (top), Kim (middle), and Scott (bottom). Ellipses represent one standard deviation for each vowel class. A visual inspection of these plots shows that both vowel classes exhibit a fair amount of overlap in both styles, with what appears to be more overlap in Donna's, as her /ɔ/ vowel class actually encompasses /ɑ/ in both speech styles (a distributional property that indicates merger). Adding the measures of Pillai scores, using the recommended thresholds for "merged" given the specific token counts, along with *p*-values for these speakers in these two styles, enables us to more confidently state that Donna's two vowels are underlyingly merged, while Kim's are distinct in both styles and Scott's are distinct at least in the conversational style. For both Donna and Kim, the Pillai scores were higher in the wordlist style compared to the conversation. Seeing these differences in Pillai scores alone, a researcher may be tempted to conclude that there is style shifting occurring for both speakers, such that the merger undoes itself in more careful speech. Leveraging all of the evidence together – Pillai scores alongside the recommended threshold for a given sample size, as well as *p*-values and a visual inspection of the data – allows us to reject this interpretation and instead see important differences between speakers in the sample.

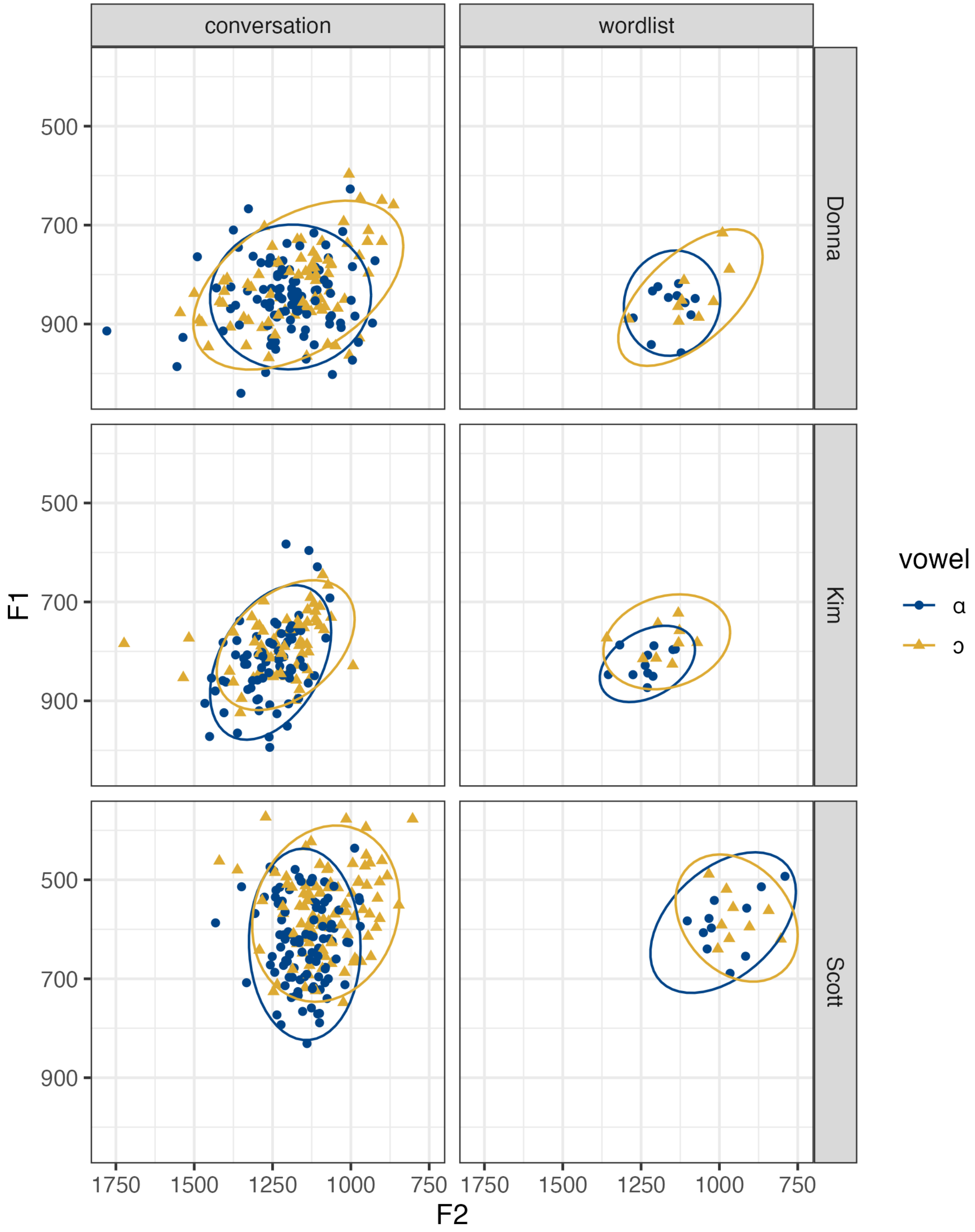

Figure 4 (color online): Low back vowels from three speakers across two styles.

We note, in fact, that interpreting Pillai score alone without the additional evidence of threshold and *p*-value provides a misleading interpretation of the entire dataset. Across the 30 speakers in the sample who completed both tasks, a one-sided paired *t*-test comparing Pillai scores in conversational data to Pillai scores in wordlist data suggests that there is a statistically significant trend towards higher Pillai scores (i.e., a "less merged" pronunciation) in the wordlist data ($t$ = -3.3296, df = 29, $p$ = 0.001). This pattern obtains across most speakers, suggesting on the surface that almost every speaker in the sample "unmerges" their vowels in wordlist style data or that they only merge the vowels in casual conversation. Given that the low-back merger is a change in progress (cf. Labov, Yaeger & Steiner 1972 for other examples), this pattern obtaining across this many speakers of all ages is suspiciously regular – much more regular than we would expect given patterns of variation and change in sociophonetic work. In fact, it was this apparently regular unmerging found in these participants' wordlists that led us to investigate the effect of sample size in the first place. Incorporating all of the relevant pieces of evidence (recommended threshold given sample size, along with Pillai scores and *p*-values) allows us to understand the suspiciously regular finding as an artefact of wordlists having far smaller sample sizes than conversational speech. Likewise, leveraging all of the evidence, including sample size differences across styles, allows a clearer understanding of the individual speakers in the sample and whether they are likely to be truly merged or not.

### B. Analysis of many speakers

While the level of scrutiny in the previous section is appropriate and encouraged for analyzing individual speakers, we acknowledge that researchers may have a different goal in mind. For instance, researchers tracking the development of a merger as it becomes closer together in phonetic space (and before a categorical merging) will want to understand how the two vowel classes become less distinct across generations – even before a categorical merger has taken place. Analyzing data speaker-by-speaker requires us to take gradient measures (Pillai scores, threshold, and *p*-value) and interpret them into discrete categories for each speaker and style ("merged", "not merged" or "partially merged", depending on the researcher's interpretation); this discrete interpretation in turn makes it difficult to track how large-scale change proceeds across generations. In this section we explore how Pillai scores have changed over time in this sample by fitting linear models to the data. The first and more straightforward approach simply accounts for sample size in the modeling; the second and more robust method implements a bootstrapping procedure to remove difference between sample sizes across speakers.

### *6. Incorporating sample size in a regression model*

We next wish to identify whether there were patterns across genders or time, so we would like to fit the Pillai scores to a linear model. However, since Pillai scores are typically non-normally distributed, the results from a regression model fit to them may be unreliable. However, as mentioned above, Pillai scores and *p*-values are functions of each other, and as seen in Figure 2, if the Pillai score for a given sample size is higher than the 95th percentile threshold, the *p*-value will be less than 0.05. Pillai scores and *p*-values essentially say the same thing. We can therefore fit a model to the *p*-values instead of Pillai scores without loss of interpretation. As mentioned above, *p*-values are uniformly distributed under the null hypothesis of underlyingly merged vowel classes, which is what we model here. And a uniform distribution can be easily transformed into a normal distribution using the `qnorm()` function in R. We can therefore can fit a linear regression model to the transformed *p*-values to study change over time. In this case, modeling on normalized *p*-values resulted in a better model fit than modeling raw *p*-values or on raw or transformed Pillai scores.[vi] As independent variables, we included gender, birth year (scaled and centered around zero using `scale()`), and their interaction. Since each speaker's Pillai scores were calculated on different amounts of data, we incorporated speaker sample size into the model as well, after transforming it using Equation 1. In R, the formula for this model was

```
lm(qnorm_p ~ gender * scale(yob) + equation1(n), data = df)
```

where `qnorm_p` represents the *p*-values normalized with the `qnorm()` function, `scale(yob)` is the scaled birth years, `equation1(n)` is the application of Equation 1 on each speaker's sample size (`n`), and `df` is the data frame of containing one row for each of the 52 speakers, and columns for metadata, sample size, and *p*-values. We direct readers to the supplementary file for an example of the R code used to implement this model.

As seen in Figure 5, the model suggests a significant interaction between gender and birth year. Men's *p*-values were relatively low compared to women's, which correspond to higher Pillai scores and greater separation between /ɑ/ and /ɔ/. We interpret this as suggesting that women are, overall, more merged than men in this sample – in other words, women are leading the change towards a merger in Washington. In fact, 19 of the 24 men had Pillai scores above their respective thresholds, suggesting a distinction in their low vowels, while 18 of the 28 women had Pillai scores lower than their thresholds, suggesting a merger. The significantly positive interaction term suggests difference between genders shrinks among younger people.

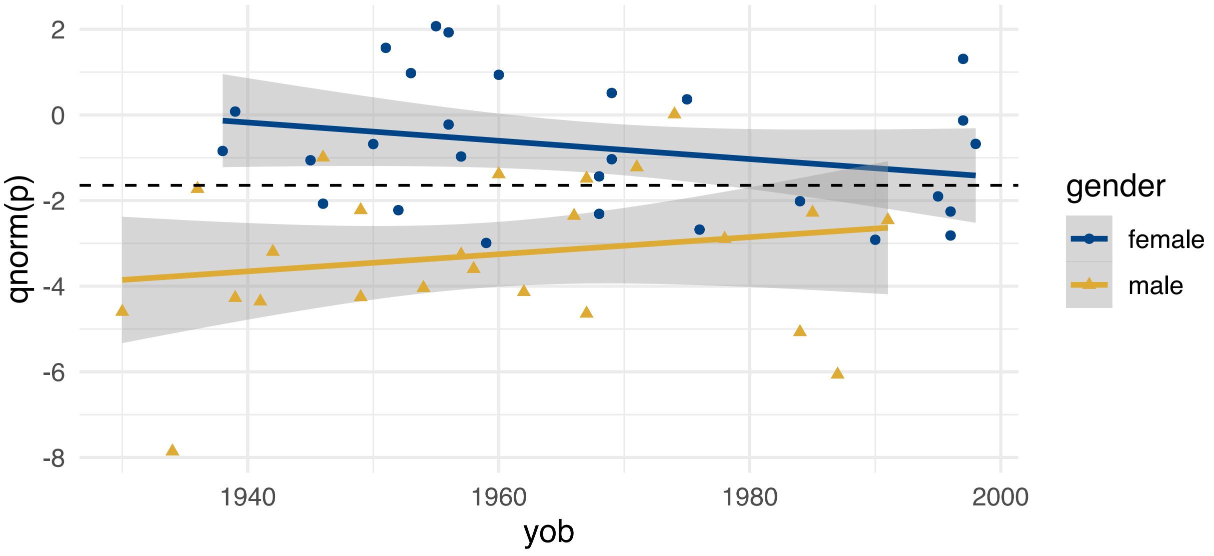

Figure 5 (color online): Normalized *p*-values by gender and birth year with predicted regression lines. The horizontal dotted line crosses the *y*-axis at -1.67, corresponding to *p*-values of 0.05; speakers below that line have *p*-values less than 0.05 and Pillai scores higher than the threshold, given their sample size.

We note that threshold calculated using Equation 1, which serves as a proxy for sample size, was a significant predictor in this model ($p$ = 0.0485). It's positive coefficient suggests that as the threshold increases (corresponding to *smaller* sample sizes), the normalized *p*-value increases (corresponding to an increase in raw *p*-values as well). We interpret this to mean that, even after incorporating sample size in the model, people with less data had larger Pillai scores. For datasets that include a wider range of sample sizes, particularly where each person contributed fewer tokens than those in dataset analyzed here, we would expect to find a larger effect of sample size on *p*-values, and consequently Pillai scores.

### 7. *Downsampling and bootstrapping*

As mentioned in Section IV, the other method that we recommend for comparing Pillai scores across a large number of speakers is to reduce the size of each speaker's dataset down to the sample size of the speaker who contributed the least amount of data. Downsampling in this way will allow us to obtain Pillai scores that are comparable across all speakers, and therefore allow us to track fine-grained changes in merger status across apparent time.

To begin our downsampling example, we first restrict our analysis to the more prolific conversational portion of the interview, to maximize the possible total sample size per speaker. The least talkative speaker in this sample produced only 82 tokens in this style. So, even though some speakers produced many more (as many as 327 in one case), we took a random sample (with replacement[vii]) of just 82 tokens from each speaker. However, we found that Pillai scores from a random sample of one speaker's data varied considerably from a different random sample from that same speaker. So, we implemented a bootstrapping procedure and took 1000 random samples (with replacement) of 82 tokens from each speaker's dataset, calculated the Pillai scores and other summary statistics from each sample, and aggregated them by speaker. We find that these aggregated values were very similar to values taken from the full dataset: the correlation between speakers' Pillai scores on the full dataset and speakers' mean Pillai scores across the 1000 samples was 0.9984, suggesting that the aggregated bootstrapped values are a very good approximation of the full dataset's values. The difference is that they are based on equal sample sizes rather than different sample sizes, which means that the resulting Pillai scores will be comparable across speakers.

To identify whether there were patterns across genders or time, we fit the normalized *p*-values to a linear model with gender and birth year as predictors. Since the data was downsampled, sample size does not need to be incorporated into the model; Pillai scores can be directly compared and the influence of particularly talkative or reticent speakers is minimized. We compared two models, one where the dependent variable was the raw *p*-value and the other where the dependent variable was normalized (again using the `qnorm()` function in R). In this case, the model on the normalized *p*-values was a better fit, based on the adjusted $R^2$ and model assumption checks (see also footnote 5). Because the values from this bootstrapped method were similar to the values obtained from the full dataset, it is unsurprising that the model suggests a similar trend: men's Pillai scores start higher and decrease across time while women's Pillai scores are lower and increase very slightly over time, with the two converging among the youngest speakers.

In this dataset, the two approaches to comparing Pillai scores across speakers (incorporating sample size in the model and downsampling with bootstrapping) yielded very similar results. We again attribute this to the fact that the sample sizes were all somewhat large and relatively homogeneous. For datasets where the number of observations per speaker is roughly similar, the downsampling and bootstrapping approach may be more appropriate since it will result in little loss of data. Similarly, for datasets where the total sample size per speaker is large, the downsampling and bootstrapping approach may also be more appropriate since downsampling to a large sample size from a very large sample size will result in little change. The difference between 30 and 60 total tokens has a larger impact on resulting Pillai score than the difference between 300 and 600. On the other hand, for samples with more variation in how many tokens were extracted from each speaker, the approach that incorporates sample size into the model may be more appropriate since the effect of sample size will be more apparent. And for samples where the number of tokens is small, incorporating the sample size into the model may also be more appropriate so that none of the already scant data is lost. In the end, we recommend researchers use the approach that makes the most sense for their data.

## VI. CONCLUSION

In this paper, we present a close view into the effect of sample size on resulting Pillai scores, a common measure for quantifying vowel overlap. We use a series of simulation experiments drawing from an underlyingly merged pair of vowels to demonstrate (1) that larger sample sizes yield reliably lower Pillai scores, (2) that unequal group sizes across the two vowel classes is irrelevant in the calculation of Pillai scores, and (3) that it takes more data than many sociolinguistic studies collect to reliably return a low Pillai score (e.g., under 0.1) even for underlyingly merged data. These results have implications for how Pillai scores are compared across studies and between speakers or speech styles within the same study. We provide some recommendations for maximizing reliability in the use of Pillai scores, and provide a formula to assist researchers in determining a reasonable Pillai score threshold to use as an indicator of merged status given their sample size. We recommend the use of Equation 1 a in conjunction with the Pillai scores' accompanying *p*-values to make informed decisions about the merged status of two vowels in a given speaker. By properly using and reporting aspects of Pillai score, researchers can come closer to accurately quantifying vowel overlap, identifying vowel mergers, and ultimately understanding broad patterns of variation and change in vowel merger.

## ACKNOWLEDGMENTS

Thanks to the Linguistics Discussion Group at BYU and the audience at ASA2021 in Seattle for their useful comments. We thank William Christensen for his help in understanding underlying distributions and transformations. We also thank the two anonymous reviewers for their thoughtful and encouraging feedback, without which this paper would have been substantially less clear and less impactful. Any remaining faults are our own. The first author also graciously acknowledges the University of Georgia Graduate School Dean's Award for funding the fieldwork that produced the data used in the case study.

See supplementary material at [URL will be inserted by AIP] for a brief tutorial on how to implement these recommendations in R.

## REFERENCES

Amengual, Mark & Pilar Chamorro. 2015. The Effects of Language Dominance in the Perception and Production of the Galician Mid Vowel Contrasts. *Phonetica* 72(4). 207–236. https://doi.org/10.1159/000439406.

Arnold, Jeffrey B. 2018. ggthemes: Extra Themes, Scales and Geoms for “ggplot2.” https://CRAN.R-project.org/package=ggthemes.

Austin, Martha. 2020. Mismatches between Linguistic and Sociolinguistic Perception. Presented at the The 94th Annual Meeting of the Linguistic Society of America, New Orleans, LA.

Bartlett, M. S. 1939. A note on tests of significance in multivariate analysis. *Mathematical Proceedings of the Cambridge Philosophical Society* 35(2). 180–185. https://doi.org/10.1017/S0305004100020880.

Becker, Kara. 2019. Introduction. In Kara Becker (ed.), *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and short front vowel shifts across North America* (Publication of the American Dialect Society 104). Durham, NC: Duke University Press.

Berry, Grant M & Mirjam Ernestus. 2018. Phonetic alignment in English as a *lingua franca* : Coming together while splitting apart. *Second Language Research* 34(3). 343–370. https://doi.org/10.1177/0267658317737348.

Berry, Grant Michael. 2018. *Liminal voices, central constraints: Minority adoption of majority sound change*. The Pennsylvania State University Ph.D. Dissertation.

Bray, Andrew. 2021. [ˈhɑ̝ki]: An Emerging Third-Order Index of a Hockey-Based Persona. Presented at the New Ways of Analyzing Variation 49, Austin, Texas.

Bray, James H. & Scott E. Maxwell. 1985. *Multivariate Analysis of Variance* (Quantitative Applications in the Social Sciences). Vol. 07–054. Newbury Park, CA: Sage Publications.

Brozovsky, Erica Sharon. 2020. *Taiwanese Texans: A Sociolinguistic Study of Language and Cultural Identity*. Austin, TX: University of Texas at Austin Ph.D. Dissertation.

Di Paolo, Marianna & Alice Faber. 1990. Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change* 2(02). 155–204. https://doi.org/10.1017/S0954394500000326.

Di Paolo, Marianna, Malcah Yaeger-Dror & Alicia Beckford Wassink. 2011. Analyzing Vowels. In Marianna Di Paolo & Malcah Yaeger-Dror (eds.), *Sociophonetics: A Student's Guide*, 87–106. 1st edn. London: Routledge.

Fieberg, John & Christopher O. Kochanny. 2005. Quantifying Home-Range Overlap: The Importance of the Utilization Distribution. (Ed.) Lanham. *Journal of Wildlife*

*Management* 69(4). 1346–1359. https://doi.org/10.2193/0022-541X(2005)69[1346:QHOTIO]2.0.CO;2.

Fisher, Sabriya, Hilary Prichard & Betsy Sneller. 2015. The Apple Doesn't Fall Far From the Tree: Incremental Change in Philadelphia Families. *University of Pennsylvania Working Papers in Linguistics* 21(2). http://repository.upenn.edu/pwpl/vol21/iss2/7.

Freeman, Valerie. 2021. Vague eggs and tags: Prevelar merger in Seattle. *Language Variation and change* 1–24. https://doi.org/doi:10.1017/S0954394521000028.

Fung, Roxana S. Y. & Chris K. C. Lee. 2019. Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception. *The Journal of the Acoustical Society of America* 146(5). EL424–EL430. https://doi.org/10.1121/1.5133661.

Gorman, Kyle & Daniel Ezra Johnson. 2013. Quantitative Analysis. In Robert Bayley, Richard Cameron & Ceil Lucas (eds.), *The Oxford Handbook of Sociolinguistics*, vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199744084.013.0011.

Hall-Lew, Lauren. 2013. 'Flip-flop' and mergers-in-progress. *English Language and Linguistics* 17(02). 359–390.

Hall-Lew, Lauren, Ruth Friskney & James M. Scobbie. 2017. Accommodation or political identity: Scottish members of the UK Parliament. *Language Variation and Change* 29(3). 341–363. https://doi.org/10.1017/S0954394517000175.

Han, Jeong-Im & Hyunsook Kang. 2013. Cross-generational Change of /o/ and /u/ in Seoul Korean I: Proximity in Vowel Space. *Phonetics and Speech Sciences* 5(2). 25–31. https://doi.org/10.13064/KSSS.2013.5.2.025.

Havenhill, Jonathan. 2015. Maintenance of the COT-CAUGHT Contrast Among Metro Detroit Speakers: A Multimodal Articulatory Analysis. *University of Pennsylvania Working Papers in Linguistics* 21(2).

Hay, Jennifer, Paul Warren & Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* (Modelling Sociophonetic Variation) 34(4). 458–484. https://doi.org/10.1016/j.wocn.2005.10.001.

Holland, Cory & Tara Brandenburg. 2017. Beyond the Front Range: The Coloradan Vowel Space. In Valerie Fridland, Alicia Beckford Wassink, Tyler Kendall & Besty E. Evans (eds.), *Speech in the Western States, Volume 2: The Mountain West* (Publication of the American Dialect Society 102), 9–30. Durham, NC: Duke University Press. DOI: 10.1215/00031283-4295277.

Islam, Md Jahurul & Iftakhar Ahmed. 2020. Mid-front and back vowel mergers in Mymensingh Bangla: An acoustic investigation. *Linguistics Journal* 14(1). 206–232.

Ismay, Chester & Albert Y. Kim. 2020. *Statistical Inference via Data Science: A ModernDive into R and the Tidyverse* (The R Series). Boca Raton, FL: Taylor & Francis Group. https://moderndive.com/index.html.

Jensen, Sandra & Natalie Braber. 2021. The BATH-TRAP split in the East Midlands. Poster presented at the New Ways of Analyzing Variation 49, Austin, Texas.

Jibson, Jonathan. 2021. Merged status thresholds for Pillai scores. Poster presented at the New Ways of Analyzing Variation 49, Austin, Texas.

Johnson, Daniel Ezra. 2010. *Stability and Change Along a Dialect Boundary: The Low Vowels of Southeastern New England* (Publication of the American Dialect Society 95). Durham, NC: Duke University Press.

Johnson, Daniel Ezra. 2015. Quantifying vowel overlap with Bhattacharyya’s affinity. Presented at the New Ways of Analyzing Variation (NWAV44), Toronto.

Johnson, Richard A. & Dean W. Wichern. 2012. *Applied Multivariate Statistical Analysis*. Phi Learning Private Limited.

Joo, Hyoun-A, Lara Schwarz & B. Richard Page. 2018. Nonconvergence and Divergence in Bilingual Phonological and Phonetic Systems: Low Back Vowels in Moundridge Schweitzer German and English. *Journal of Language Contact* 11(2). 304–323. https://doi.org/10.1163/19552629-01102006.

Kelley, Matthew C. & Benjamin V. Tucker. 2020. A comparison of four vowel overlap measures. *The Journal of the Acoustical Society of America* 147(1). 137–145. https://doi.org/10.1121/10.0000494.

Kendall, Tyler & Valerie Fridland. 2017. Regional relationships among the low vowels of U.S. English: Evidence from production and perception. *Language Variation and Change* 29(2). 245–271. https://doi.org/10.1017/S0954394517000084.

Kendall, Tyler & Valerie Fridland. 2021. *Sociophonetics* (Key Topics in Sociolinguistics). Cambridge: Cambridge University Press.

Kettig, Thomas T. 2021. *Ha’ina ’ia Mai Ana Ka Puana: The Vowels of ’Ōlelo Hawai’i*. Mānoa, Hawai’i: University of Hawai’i at Mānoa Ph.D. Dissertation.

Labov, William. 1994. *Principles of linguistic change. Vol. 1: Internal features* (Language in Society). Oxford: Wiley-Blackwell.

Labov, William & Maciej Baranowski. 2006. 50 msec. *Language Variation and Change* 18(03). https://doi.org/10.1017/S095439450606011X.

Labov, William, Sabriya Fisher, Duna Gylfadottír, Anita Henderson & Betsy Sneller. 2016. Competing systems in Philadelphia phonology. *Language Variation and Change* 28(3). 273–305. https://doi.org/10.1017/S0954394516000132.

Labov, William, Malcah Yaeger & Richard Steiner. 1972. *A quantitative study of sound change in progress: Volume 1*. Philadelphia, PA: US Regional Survey.

Lüdecke, Daniel, Mattan Ben-Shachar, Indrajeet Patil, Philip Waggoner & Dominique Makowski. 2021. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software* 6(60). 3139. https://doi.org/10.21105/joss.03139.

Lüdecke, Daniel, Indrajeet Patil, Mattan Ben-Shachar, Brenton Wiernik, Philip Waggoner & Dominique Makowski. 2021. see: An R Package for Visualizing Statistical Models. *Journal of Open Source Software* 6(64). 3393. https://doi.org/10.21105/joss.03393.

Metropolis, Nicholas & S. Ulam. 1949. The Monte Carlo Method. *Journal of the American Statistical Association* 44(247). 335–341. https://doi.org/10.1080/01621459.1949.10483310.

Moulton, William G. 1968. Structural Dialectology. *Language* 44(3). 17.

Nadeu, Marianna & Margaret E.L. Renwick. 2016. Variation in the lexical distribution and implementation of phonetically similar phonemes in Catalan. *Journal of Phonetics* 58. 22–47. https://doi.org/10.1016/j.wocn.2016.05.003.

Newbert, Cornelia. 2021. *Language variation in South Africa: A sociophonetic study of the vowel system of Black South African English*. Regensburg, Germany: University of Regensburg Ph.D. Dissertation.

Nycz, Jennifer & Lauren Hall-Lew. 2013. Best practices in measuring vowel merger. *Proceedings of Meetings on Acoustics* 20(1). 060008. https://doi.org/10.1121/1.4894063.

Olson, Chester L. 1976. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin* 83(4). 579–586. https://doi.org/10.1037/0033-2909.83.4.579.

Pedersen, Thomas Lin & Fabio Crameri. 2020. scico: Colour Palettes Based on the Scientific Colour-Maps. https://CRAN.R-project.org/package=scico.

Pfiffner, Alexandra M. 2021. *Cue-Based Features: Modeling change and varitaion in the voicing contrasts of Minnesotan English, Afrikaans, and Dutch*. Washington, DC: Georgetown University Ph.D. Dissertation.

Pillai, K. C. S. 1955. Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics* 26(1). 117–121. https://doi.org/doi:10.1214/aoms/1177728599.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Regan, Brendan. 2020. Extending Pillai Scores to Fricative Mergers: Advancing a Gradient Analysis of a Split-in-Progress in Andalusian Spanish. *University of Pennsylvania Working Papers in Linguistics* 26(2).

Rencher, Alvin C. & William F. Christensen. 2012. *Methods of multivariate analysis* (Wiley Series in Probability and Statistics). Third Edition. Hoboken, New Jersey: Wiley.

Schmidt, Penelope, Chloé Diskin-Holdaway & Debbie Loakes. 2021. New insights into /el/-/æl/ merging in Australian English. *Australian Journal of Linguistics* 41(1). 66–95. https://doi.org/10.1080/07268602.2021.1905607.

Seaman, D. Erran, Joshua J. Millspaugh, Brian J. Kernohan, Gary C. Brundige, Kenneth J. Raedeke & Robert A. Gitzen. 1999. Effects of Sample Size on Kernel Home Range Estimates. *The Journal of Wildlife Management* 63(2). 739. https://doi.org/10.2307/3802664.

Sloos, Marjoleine. 2013. The reversal of the BÄREN-BEEREN merger in Austrian Standard German. *The Mental Lexicon* 8(3). 353–371. https://doi.org/10.1075/ml.8.3.05slo.

Sneller, Betsy. 2018. *Mechanisms Of Phonological Change*. Philadelphia: University of Pennsylvania Ph.D. Dissertation.

Stanford, James N., Monica Nesbitt, James King & Sebastian Turner. 2019. Pioneering a dialect shift in the Pioneer Valley: Evidence for the Low-Back-Merger Shift in Western Massachusetts. Presented at the New Ways of Analyzing Variation 48, Eugene, Oregon.

Stanley, Joseph A. 2020. *Vowel dynamics of the Elsewhere Shift: A sociophonetic analysis of English in Cowlitz County, Washington*. Athens, Georgia: University of Georgia Ph.D. Dissertation.

Stanley, Joseph A. 2021. joeyr: Functions for Vowel Data. https://joeystanley.github.io/joeyr/.

Strelluf, Christopher. 2018. *Speaking from the heartland: The midland vowel system of Kansas City* (Publication of the American Dialect Society 103). Durham, NC: Duke University Press.

Tse, Holman. 2018. *Beyond the Monolingual Core and out into the Wild: A Variationist Study of Early Bilingualism and Sound Change in Toronto Heritage Cantonese*. Pittsburgh, PA: University of Pittsburgh Ph.D. Dissertation.

Upton, Graham & Ian Cook. 2014. *A Dictionary of Statistics*. Oxford University Press. https://doi.org/10.1093/acref/9780199679188.001.0001.

Venables, W. N., Brian D. Ripley & W. N. Venables. 2002. *Modern applied statistics with S* (Statistics and Computing). 4th ed. New York: Springer.

Warner, Natasha, Allard Jongman, Joan Sereno & Rachèl Kemps. 2004. Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics* 32(2). 251–276. https://doi.org/10.1016/S0095-4470(03)00032-9.

Warren, Paul. 2018. Quality and quantity in New Zealand English vowel contrasts. *Journal of the International Phonetic Association* 48(3). 305–330. https://doi.org/10.1017/S0025100317000329.

Wasserstein, Ronald L., Allen L. Schirm & Nicole A. Lazar. 2019. Moving to a World Beyond "p < 0.05." *The American Statistician* 73(sup1). 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Wassink, Alicia Beckford. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties. *The Journal of the Acoustical Society of America* 119(4). 2334–2350.

Whalen, D. H. & Wei-Rong Chen. 2019. Variability and Central Tendencies in Speech Production. *Frontiers in Communication* 4. 49. https://doi.org/10.3389/fcomm.2019.00049.

Wickham, Hadley. 2015. *ggplot2: Elegant Graphics for Data Analysis* (Use R!). 2nd edn. New York: Springer.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* 4(43). 1686. https://doi.org/10.21105/joss.01686.

Wong, Amy Wing-mei & Lauren Hall-Lew. 2014. Regional variability and ethnic identity: Chinese Americans in New York City and San Francisco. *Language & Communication* 35. 27–42. https://doi.org/10.1016/j.langcom.2013.11.003.

[i] See also Kelley & Tucker (2020) for details on the overlapping coefficient, another nonparametric method for calculating overlap.

[ii] Preliminary work for Sneller (2018) attempted, with mixed results, a modified model approach of Bhattaharyya's Affinity by extracting model values for fixed effects and manually adjusting the data. However, the onerousness of this approach makes it not widely implementable in comparison with Pillai, which has the advantage of being able to easily integrate commonly used mixed-effects models.

[iii] In this paper, we focus on the effect of sample size for underlyingly merged speakers. Future work may fruitfully investigate how sample size interacts with Pillai score for underlyingly unmerged speakers as well.

[iv] Let $p_{95} = e^{1-log\,(\frac{n}{2})}$ . Since $e^{x-y} = e^{x+(-y)} = e^{x}e^{-y} = e^{x}\frac{1}{e^{y}} = \frac{e^{x}}{e^{y}}$, and since $e^{1} = e$, then $p_{95} = \frac{e}{e^{\ln\,(\frac{n}{2})}}$. Since $e^{\ln\,(x)} = x$, then $p_{95} = \frac{e}{\frac{n}{2}}$. Multiplying the numerator and the denominator by 2 produces the final equation, $\frac{2e}{n}$. (We are grateful to the anonymous review who pointed out these simplification steps for this formula.) This is most easily implemented in R as `2*exp(1)/n`.

[v] One reviewer asked why we did not recommend accounting for the difference in sample sizes across speech styles by simply adding a term for style in the linear model. The reasoning is fairly straightforward: researchers cannot control how many tokens of interest are produced in the conversational portion of the interview, allowing the difference in token count across style to vary wildly by speaker. Since sample sizes influence Pillai score in a predictable way but style influences Pillai score in an unpredictable way (as it's dependent on the differences in sample sizes), it is inadvisable to use style as the predictor rather than a more straightforward way to control for or incorporate sample size.

[vi] We determined this using the `compare_performance()` and `check_model()` functions in the `performance` package in R (Lüdecke, Ben-Shachar, et al. 2021). The model fit to the raw *p*-values had a lower adjusted $R^2$ than the one fit to the transformed data (0.2169 vs. 0.4048). It also satisfied assumptions of linearity, homogeneity of variance, and normality of residuals, unlike the

model fit to raw *p*-values. Models fit to raw or log-transformed Pillai scores had lower $R^2$ values and failed to meet satisfy of the model assumptions.

[vii] Note that we sample with replacement `(replace=TRUE)`, in order to make our resampling comparable across all speakers including our least talkative speaker.

If we were to resample *without* replacement, this introduces a confound related to sample size: the amount of error introduced per speaker is proportional to their sample size. Resampling without replacement 1000 times would produce 1000 identical distributions for the least talkative speaker, but 1000 different distributions for every other speaker. Resampling *with* replacement allows us to obtain a standard deviation for each speaker with similar confidence, and introduces a similar amount of uncertainty across speakers in the means of their distributions. For more information on bootstrapping with replacement, we refer the reader to Ismay & Kim (2020).