Laboratory Phonology
Journal of the Association for Laboratory Phonology

Cohen Priva, U., & Sanker, C. 2019 Limitations of difference-in-difference for measuring convergence. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1): 15, pp. 1–29. DOI: https://doi.org/10.5334/labphon.200

JOURNAL ARTICLE

Limitations of difference-in-difference for measuring convergence

Uriel Cohen Priva and Chelsea Sanker
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, US
Corresponding author: Uriel Cohen Priva (uriel_cohen_priva@brown.edu)

Linguistic convergence is the phenomenon in which interlocutors' speech characteristics become more similar to each other's. One of the methods frequently used to measure convergence is the difference-in-difference (DID) approach, comparing change in absolute distance between a subject and an interlocutor or model talker. We show that this approach is not a reliable measure of convergence when the starting values of the subject and the interlocutor or model talker are close, which can result in the measurement of apparent divergence, while extreme starting points can result in overestimation of convergence. These biases are of particular concern in studies that look for individual differences in convergence. We propose an alternative approach, linear combination, which does not have the same biases, and demonstrate the advantages of this method using data from convergence studies of four linguistic characteristics and simulated data.

Keywords: Convergence; Difference-in-difference; individual differences; methodology

1. Introduction

Convergence is the phenomenon in which speakers become more similar to their interlocutors, which has been observed in many characteristics, both linguistic and non-linguistic. Many studies find variation in performance across participants, which is used as evidence for individual differences or population differences (e.g., Natale, 1975; Yu, Abrego-Collier, & Sonderegger, 2013). However, examining differences across participants can be particularly sensitive to biases in the way that convergence is measured. Convergence is often measured based on change in absolute distance between each participant and the model talker or interlocutor, in the difference-in-difference (DID) method.

Using data from four convergence studies, we demonstrate that DID is not a suitable measure of convergence in the following ways: (1) DID underestimates convergence and can even produce apparent divergence when the subject's baseline performance is close to the reference value of the interlocutor or model talker, and (2) DID interprets regression to the mean as convergence. For these reasons, DID measures of convergence for individual talkers are unreliable. Our proposed alternative, linear combination, while fully capable of measuring convergence (Cohen Priva, Edelist, & Gleason, 2017), is not subject to these issues and thus provides more reliable estimates of the convergence exhibited by each individual participant.

1.1. Convergence tasks and measurements

Convergence to an interlocutor within a conversation or to a model talker during shadowing or other exposure is often measured by the change in difference between the two speakers. In shadowing tasks, subjects are exposed to recordings of a model talker, which they repeat after, and the comparison is made of their speech before and after exposure. Conversational interactions present additional complications in defining reference points for the speakers, because both of the interlocutors are potentially changing.

In shadowing tasks, recordings are often naturally produced (e.g., Pardo, Urmanche, Wilman, & Wiener, 2017; Babel, 2012), but are sometimes manipulated to create extreme values in a particular measure (e.g., Nielsen, 2011; Yu et al., 2013), which ensures that the reference values which subjects are exposed to will be relatively far from all of the subjects' starting values and differ in the same direction, eliminating possible effects of variation in absolute starting distance or the direction of the difference. However, it is possible that presenting reference values outside the range of normal human performance could undermine the ecological validity of such studies.

In interactional studies, it is difficult to reliably expose subjects to consistent extreme productions; while confederate interlocutors can be trained to produce certain behavioral patterns such as face rubbing and foot shaking (Chartrand & Bargh, 1999) and scripting can control syntactic and lexical choices (e.g., Branigan, Pickering, & Cleland, 2000), confederates cannot precisely control the phonetic details of their speech. Gijssels, Casasanto, Jasmin, Hagoort, and Casasanto (2016) resolved this issue of phonetic control in conversation by using a virtual interlocutor whose speech was entirely controlled and varied only in F0, the variable of interest. Felker, Tronsco-Ruiz, Ernestus, and Broersma (2018) offered a different resolution that allowed control of the productions while still presenting natural speech, by using a ventriloquist setup to present pre-recorded speech as if it were being produced live. While such a setup might allow acoustic manipulations in conversational interactions, the vast majority of current interactional studies of convergence do not control the interlocutors' speech, so biases created by variation in starting distance can create large problems in analysis of by-speaker patterns.

Even when stimuli are not manipulated, the choice of model talker or the particular assigned interlocutor can influence how distant subjects are from these other speakers. Some studies use multiple model talkers (e.g., Pardo et al., 2017; Babel, McGuire, Walters, & Nicholls, 2014), and often find differences depending on the talker. However, many studies have a single model talker (e.g., Babel, 2010; Dias & Rosenblum, 2016; Mitterer & Ernestus, 2008) or have each participant converse with only a single interlocutor (e.g., Pardo, 2006; Abel & Babel, 2017), which could obscure possible differences in convergent behavior due to particular interlocutors or model talkers or the distance between them and the subjects. Studies always have multiple subjects, which reduces some of the potential noise due to variation in individual starting distance.

1.2. Modelling convergence

In shadowing tasks, subjects' productions are measured before exposure to the model talker (S_b) and after exposure (S_R), and compared to the recordings of the model talker (R). Convergence in this system would be any change from S_b to S_R that makes S_R more R-like. In interactional tasks, it is similarly possible to compare productions from before the conversation or in conversations with other partners to productions within the shared conversation. When multiple interlocutors are available, subjects' performance can be approximated with a linear combination of *consistency* (*self-correlation*), measured in β_{S_b}, and *convergence*, measured in β_R, as well as noise, as given in (1).[1]

(1) $$S_R \approx \beta_0 + \beta_{S_b} S_b + \beta_R R + \epsilon$$

[1] Some work uses a similar model, but includes a coefficient only for the reference value of the model talker or interlocutor, and model $\beta_{S_b} S_b$ as a random intercept (e.g., Schweitzer & Lewandowski, 2013; Schweitzer & Walsh, 2016).

However, many studies do not model both the effects of the subject and the model talker, instead looking at the change in similarity of the subject and the model talker. Some studies measure similarity subjectively with AXB designs (e.g., Goldinger, 1998; Pardo, Gibbons, Suppes, & Krauss, 2012); these tasks yield holistic judgements about whether the subject's productions before or after exposure are more similar to the recordings of the model talker. Other studies measure similarity in particular acoustic characteristics as the absolute difference between the subject and the model talker (e.g., Babel, 2012; Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013), with convergence as change in that difference (difference-in-difference, DID), as in (2).

(2) $$\text{DID} := |S_R - R| - |S_b - R|$$

DID is sometimes used to quantify convergence in conversational tasks, looking at the change from the two speakers' starting distance and ending distance (e.g., Pardo, Cajori Jay, & Krauss, 2010), or their distance earlier and later in the conversation (e.g., Levitan & Hirschberg, 2011; Abel & Babel, 2017). When comparing interlocutors' productions within a shared conversation, conversations may appear to be convergent due to participants independently being similarly influenced by the task, e.g., speaking more quickly as they become familiar with the task or their interlocutor; some studies compare distance between interlocutors to distance between individuals who were not interacting with each other, to control for such effects (e.g., Sanker, 2015; Oben & Brône, 2016), though many do not. In AXB or similar perceptual testing of distance, the X reference point for the interlocutor is sometimes taken from the middle of the conversation, compared to the other speaker's productions before and after the conversation or before and during the conversation (e.g., Pardo, 2006). Many conversational convergence studies compare synchronous variations in both participants (e.g., Levitan & Hirschberg, 2011; Schweitzer & Lewandowski, 2013), either in addition to or instead of examining overall convergence across the conversation.

Some studies, rather than measuring change in distance between two interlocutors, look at correlations between partners' productions or compare speakers' productions in different conditions. Comparing participants' productions under two manipulation conditions is a particularly common method in syntactic priming paradigms, e.g., comparing participants' use of dative indirect objects and double accusative constructions when they had been exposed to descriptions using one or the other (Bock, 1986; Branigan et al., 2000). It is also used in some phonetic studies, e.g., comparing a condition of exposure to lengthened VOTs to a condition with shortened VOTs (Nielsen, 2011). Some work also compares correlations between conversational partners to correlations between speakers who were not conversing with each other (e.g., Gregory & Webster, 1996), or models predictors of each speaker's productions, including the interlocutor's productions as a predictor (Cohen Priva & Sanker, 2018; Schweitzer & Lewandowski, 2013), which both examine how conversational partners' speech patterns are related to each other, rather than measuring distance per se.

1.3. Possible issues with DID

We argue that the DID approach has biases that make it unreliable for investigations of individual differences, except when the reference value is outside the range of normal productions. DID can still capture convergence broadly when aggregated across participants, but it introduces a degree of noise that could obscure convergence, particularly when compared across groups of participants.

The first issue is that DID does not distinguish between different trajectories producing the final distance; distance due to lack of convergence is treated the same way as distance due to speakers over-converging. For example, if $S_b = 5$ and $R = 4$, then $S_R = 5$ and $S_R = 3$ would be equally non-convergent (DID = 0), though the former reflects a lack of change and the latter reflects over-convergence.

The second issue is that eliminating the term for speakers' consistency (β_{S_b} above) and the error term means that DID will underestimate convergence when the speaker and the interlocutor or model talker had a small initial distance, because individual variability from a nearly shared starting point is more likely to appear divergent than variability when the speaker and interlocutor or model talker had a larger starting distance. When the starting distance is small, there is little room for convergence, so noise from random variability is more likely to overshadow actual convergent shifts. This bias may produce the appearance of individual differences in convergence, even when the actual variation is simply reflecting differences in starting distance.

Some work includes a comparison of distance between subjects and their actual interlocutors or model talkers versus distance between 'pseudo-pairs' of subjects and speakers or model talkers they did not interact with (e.g., Levitan & Hirschberg, 2011; Miller, Sanchez, & Rosenblum, 2014; Sanker, 2015). Such methods are generally aimed at reducing measurement of convergence in shifts that are actually due to task-based effects. They additionally may reduce possible artifacts due to how convergence is measured, correlating with individuals' starting distance from the interlocutor or starting distance from the population mean; such artifacts would be present both for real pairs and the pseudo-pairs, and be factored out.

The final issue is that difference-in-difference is susceptible to effects of *regression to the mean*, because it does not have an error term to control for noise. Effects of novelty and repetition may produce non-representative measured baselines even beyond effects of noise. In comparisons of pre-task and post-task productions of a word list (e.g., Pardo et al., 2010), participants may produce atypical utterances during initial pre-task productions based on lack of familiarity with the task or the particular words. While comparison of speech from early and late in the conversation (e.g., Levitan & Hirschberg, 2011; Abel & Babel, 2017) avoids some of these risks, the data contains additional noise due to less strict control over the items produced and the environments in which they were said. If a speaker's baseline is not estimated correctly and an extreme variant is taken as that speaker's baseline, reversion to less extreme values in subsequent performance after exposure or during an interaction could be interpreted as convergence with the interlocutor or model talker, whose baseline value is likely to be less extreme. Such an effect would make speakers whose baseline performance was measured as being closer to the mean seem less convergent than speakers whose baseline performance was further than the mean.

Differences in starting distances between subjects and interlocutors or model talkers and distance between subjects and the population mean may account for some of the variation in convergence across measures. In the same task, there can be differences in overall convergence in different measures (e.g., Babel, 2012; Sanker, 2015) and also differences in effects of conditioning factors (e.g., Bilous & Krauss, 1988; Pardo et al., 2017). Differences in production variability by measure might result in apparent differences in convergence, and obscure potential parallels across measures. Moreover, such measure-specific patterns mean that the results of a study might seem to be very different depending on which measure is used.

In this paper, we use data from four interactional studies of convergence (from Cohen Priva & Sanker, 2018) and simulated data, defined to lack any individual differences in convergence, to demonstrate limitations of the DID approach, and offer an alternative analytical method, linear combination, that does not suffer from these limitations. Though existing theories of convergence remain broadly consistent with the data and our results do not motivate a paradigm shift in analyzing the mechanism of convergence, the methodological issues that we present have implications for the role of individual differences in convergence, as we demonstrate that much of the appearance of individual differences is likely an artifact of how convergence is measured.

2. Materials

2.1. Underlying studies

We use data from the four convergence studies used by Cohen Priva and Sanker (2018). All studies are based on the Switchboard corpus (Godfrey & Holliman, 1997), a large collection of telephone conversations. In this corpus, each speaker was randomly paired with another speaker and given a conversation topic, participating in several such conversations. Most conversations involved interlocutors who participated in at least one other conversation, making it possible to compare their performance across multiple interactions. However, a speaker did not interact with the same interlocutor more than once. This provides a large corpus of natural speech for many speakers conversing with several different partners. The conversation sides are recorded separately, facilitating reliable measurements for each speaker.

Each conversation is annotated for clarity. To ensure reliable acoustic measurements (F0 median and variance), calls with high levels of background noise, echoing, or other issues were omitted. This left 464 speakers used for acoustic measures. For the other measures (*uh*:*um* ratio and speech rate), no conversations were omitted, comprising 518 speakers. Conversations averaged 6:20 minutes. The word-level annotations produced at MS State (Harkins, Feinstein, Lindsey, Martin, & Winter, 2003) were used to measure word duration.

We follow Cohen Priva et al. (2017) and Cohen Priva and Sanker (2018) in treating each subject's performance as the value to be predicted (S_R), the average performance of that subject in other conversations as their baseline (S_b), and the average performance of their interlocutor in other conversations as the reference value (R). We did not use the interlocutor's performance in the same conversation as the reference value because (a) it could be affected by the subject's performance and (b) the subject and interlocutor could co-vary without converging, due to factors influencing both of them, such as the conversation topic, or the amicability of their interaction. We do not expect this choice to affect the arguments made in this paper; our goal is to compare different methods of measuring convergence, rather than to compare effects of using different reference values within a given method.

We use the four measures presented by Cohen Priva and Sanker (2018):

F0 median F0 is the measure of the frequency of wave cycles produced by the vibration of the vocal folds. The measurement of frequency was converted into the mel scale, which provides a better approximation of human perception than Hz (Stevens, Volkmann, & Newman, 1937).

F0 range F0 range was measured as the log of the ratio of the 75th percentile to 25th percentile of F0 measurements in mels.

Speech rate Speech rate in a conversation was measured as the mean log speech rate of individual utterances. Following Cohen Priva et al. (2017), point-wise speech rate was measured as the actual utterance duration (including pauses) divided by the expected utterance duration. Expected utterance duration was calculated as the sum of the predicted durations of words in the utterance, each calculated as the predicted value of a linear regression using the median duration of that word in the entire corpus, the length of the utterance, and the distance from the end of the utterance. Unlike F0 measurements, speech rate was calculated based on hand-corrected values.

***uh*:*um* ratio** This measure was calculated as the log odds of *uh* versus *um*, two frequently-used filled pauses in English. The use of one or the other has been attributed to processing factors (e.g., Clark & Fox Tree, 2002), but is also influenced by other factors such as gender (Acton, 2011). Log odds were calculated as the predicted values plus the residuals of a logistic regression between the number of *uh* uses and *um* uses in each conversation side, which could be evaluated even when a subject never used one or the other.

2.2. Convergence models

We compare two approaches to modelling convergence: (1) difference in difference (DID) and (2) mixed-effects linear regression using the subject's and interlocutor's baselines as predictors of each subject's productions. All the measures were standardized.

DID follows the calculation in (2). Convergence is measured as the difference between two values: (a) The absolute difference between a subject's performance without the exposure (in other conversations, in this case) and their interlocutor's baseline, and (b) the absolute difference between the subject's performance during the shared conversation and the interlocutor's baseline. Positive numbers indicate convergence: The two speakers have more similar productions during the shared conversation than they do when not speaking to each other. Negative values indicate divergence: Speakers are more distinct from each other during their shared conversation than they were when not speaking to each other. These DID measures can then be used as an input in more complex models: grouped by subject to yield each subject's tendency to converge, or compared across conditions to detect whether those conditions influence degree of convergence.

While DID is generally not examined with regression models, we use these models to more closely parallel the alternative linear combination model that we propose, which can only be done with regression. The limitations of DID are based on comparing distance without reference to the raw values for each speaker, rather than being based on how those differences are modelled, so the regression structure should not change the behavior of DID results.

When multiple subjects and interlocutors are present, as in Cohen Priva et al. (2017), (2) would take the form of (3), given in lme4 syntax. `DID` stands for the difference-in-difference. The random intercept `(1 | subject)` can capture some subjects' tendency to have higher or lower difference-in-difference values, the random intercept `(1 | interlocutor)` can capture some interlocutors eliciting higher or lower difference-in-difference values, and the random intercept `(1 | conversation)` can capture particular conversations having higher or lower difference-in-difference values.

(3)
```
DID ~ 1 +
    (1 | subject) + (1 | interlocutor) + (1 | conversation)
```

The intercept would be significantly positive if overall convergence is detected. The random intercepts for subject, interlocutor, and conversation would model individual differences in convergent behaviors by individual, by interlocutor, or by conversation.

Linear combination follows the calculation in (1). Convergence is measured in a mixed-effects linear regression in which subjects' performance is regarded as a linear combination of their baselines and their interlocutors' baselines. Random intercepts are minimally used for subject, interlocutor, and conversation. This yields a formula of the form (4), given in lme4 syntax, in which `subject.value` represents the speakers' performance in conversation, `subject.baseline` represents the subject's performance outside the conversation, and `interlocutor.baseline` represents the interlocutor's performance outside the conversation. The random intercepts `(1 | subject)`, `(1 | interlocutor)`, and `(1 | conversation)` account for additional per-subject, per-interlocutor, and per-conversation variability, respectively.

(4)
```
subject.value ~ 1 + subject.baseline +
interlocutor.baseline + (1 | subject) +
(1 | interlocutor) + (1 | conversation)
```

The intercept in this case is expected to be zero if the predictors are standardized, as they are in the studies presented here. The subjects' baseline models *consistency*, the extent to which subjects' speech patterns are consistent across conversations, and the interlocutors' baseline would model *convergence*, the extent to which subjects are affected by their interlocutors' performance.

The by-subject intercept is expected to explain little variance, as the subject's baseline is explicitly provided as a fixed predictor, and is included only for completeness. The by-interlocutor intercept is somewhat more necessary, as it can capture variance not due to convergence, such as modifying one's speech when speaking with figures of authority or older people. The by-conversation intercept could capture interactional effects that are not convergence per se, which influence both speakers' performance in the same direction, rather than toward one another.

Additional fixed and random effects could be added to address particular research questions. For instance, adding a by-subject random slope for the interlocutor's baseline could be used to model individual variation in convergence, yielding a formula of the form (5). The random slope would be non-zero if there is variance among speakers with respect to the reliance on the interlocutors' baseline, thereby capturing individual differences in convergence. The formula explicitly requests that the variance-covariance matrix between the random intercept and the random slope is not evaluated because the random intercept is expected to be zero (so that the model would converge). This is done by replacing the single pipe | with a double pipe || in the expression specifying the random effects structure per subject.[2]

[2] In lme4, the expression `(1 + slope || group)` is equivalent to the expression `(1 | group) + (0 + slope | group)`.

(5)
```
subject.value ~ 1 + subject.baseline +
interlocutor.baseline +
(1 + interlocutor.baseline || subject) +
(1 | interlocutor) + (1 | conversation)
```

Similarly, if two convergence conditions are compared, a fixed effect per condition could be added, and the particular effect of the condition on convergence would be the coefficient for the interaction term between the condition and the interlocutor's baseline, as in (6).

(6)
```
subject.value ~ 1 + subject.baseline +
interlocutor.baseline * condition + (1 | subject) +
(1 | interlocutor) + (1 | conversation)
```

Crucially, measuring convergence would always involve a manipulation of the coefficient of the interlocutor's baseline in the mixed effects model. The use of this method is currently rather limited, but has been established as effective for measuring convergence. Cohen Priva et al. (2017) use this method, adding demographic fixed predictors, for convergence in speech rate. Cohen Priva and Sanker (2018) use the same method, with by-subject random slopes for the interlocutor's baseline to measure individual differences in convergence, and find convergence in each of the four datasets (also used here). Cohen Priva and Sanker (n.d.) add two additional datasets, for which the linear combination method also proves powerful enough to capture convergence. Schweitzer and Lewandowski (2013) and Schweitzer and Walsh (2016) use a similar model, though they omit the subject baseline. Omitting the subject baseline as a fixed effect makes the per-subject random intercept serve as the speaker's baseline. This should be very similar to using the explicit baseline, but the range of baseline values would be assumed to be normally-distributed, which may not be the case, as in the case in bimodal distributions (e.g., F0 values, see **Figure 2**).

3. Studies

3.1. Study 1: Proximity leading to divergence in DID models

3.1.1. Introduction

Current accounts of convergence do not predict that subjects who start out with productions close to their interlocutors' will diverge, and it is not a pattern that seems motivated by social or phonological factors. However, we show here that one of the major shortcomings of the DID approach is that it is prone to overestimate divergence when a subject's baseline performance is close to the reference value.

To test the relationship between starting distance and convergence, we fit DID models that include the absolute difference between the speaker's and interlocutor's baseline as a predictor. We contrast these models with linear combination models, which do not create an artificial relationship between starting distance and convergence; without this artifact, no relationship is expected.

3.1.2. Methods and materials

For the DID models, we extend the general model described in (3) by adding an additional predictor, the absolute difference between the subject's and interlocutor's baseline, yielding the formula in (7). The coefficient of this predictor will be positive if our claim holds, signifying that a smaller initial difference is more likely to result in lower DID values. We fit this model to each of the four measures.

(7)
```
DID ~ 1 + abs(subject.baseline - interlocutor.baseline) +
(1 | subject) + (1 | interlocutor) + (1 | conversation)
```

We also trained the linear combination equivalent model. We extend the formula in (4) by adding the interaction term between the interlocutor's baseline (which measures convergence) and the absolute difference between the subject and the interlocutor. This yields the formula in (8), in which the interaction term on the second line is the variable of interest. In both of these models, each data point represents one conversation side.

(8)
```
subject.value ~ 1 + subject.baseline + interlocutor.baseline +
interlocutor.baseline:abs(subject.baseline - interlocutor.
baseline) + (1 | subject) + (1 | interlocutor) +
(1 | conversation)
```

In interest of deviating the least from the DID formula, we specified the interaction term in (8) using : rather than *. The implications are that we do not estimate how the absolute difference in baselines affects the speakers' performance, only how the absolute difference affects the speakers' convergence, as in the DID models. An effect on the speakers' performance, rather than their convergence, would be expected if speakers e.g., speak faster when their interlocutors' performance is far from their own, regardless of whether their speech rate is fast or slow. There is no parallel measure in the DID models, and it is not expected to relate to convergence. In this model and in subsequent models in which we used the same approach, we verified that the results for convergence in minimally different models in which the term is not excluded (in which * is used to introduce the interaction) do not differ in their statistical significance from the convergence results in the reported models.

3.1.3. Results and discussion

All models and data are available in the supplementary materials, and are summarized below.

In all of the four measures, the absolute distance between the subject's baseline performance and the interlocutor's baseline performance was positively correlated with higher DID values, as shown in (9). That is, the DID models were more likely to find convergence for subjects whose baselines were further from their interlocutors' baselines. This is consistent with our predictions, and is not expected to hold in other methods of measuring convergence.

(9) Study 1: Regression results for the absolute distance between the subject's baseline performance and the interlocutor's baseline performance across four measures, for DID models.

	β	*SE*	*df*	*t*	*p*
F0 median	0.16	0.02	3600	8.2	< 0.0001
F0 variance	0.54	0.02	3368	29.9	< 0.0001
Speech rate	0.45	0.02	3604	29.1	< 0.0001
uh:um ratio	0.43	0.02	4071	25.8	< 0.0001

The estimate for the intercepts was negative in all measures, as shown in (10). This indicates that for subjects whose baselines differed very little from their interlocutors, the DID value was negative, not just a smaller positive value, and would thus appear to be divergence.

(10) Study 1: Regression results for the DID models' intercepts, across four measures.

	β	*SE*	*df*	*t*	*p*
F0 median	–0.17	0.03	1430	–6.3	< 0.0001
F0 variance	–0.60	0.03	1117	–22.7	< 0.0001
Speech rate	–0.51	0.02	991	–21.8	< 0.0001
uh:um ratio	–0.48	0.03	1301	–19.3	< 0.0001

In contrast, the linear combination models found no significant effect for the interaction between absolute distance and the interlocutors' baseline in any of the measures, as shown in (11). These results suggest that there is no special status for the initial distance between the interlocutor and the subject in influencing how much they actually converge, and the significant interactions found in the DID models were indeed only artifacts of how convergence was defined. **Figure 1** illustrates the differences between the coefficients for the absolute distance between the interlocutors' baseline and the speakers' baseline. The DID model coefficients are large for all measures, while the coefficients for linear combination models are all negligible.

(11) Study 1: Regression results for the interaction between the interlocutor's baseline performance and the distance between the subject's baseline and the interlocutor's baseline across four measures, for linear combination models.

	β	*SE*	*df*	*t*	*p*
F0 median	–0.0061	0.007	2882	–0.9	0.39
F0 variance	–0.0108	0.013	1215	–0.8	0.42
Speech rate	0.0018	0.008	780	0.2	0.82
uh:um ratio	–0.0138	0.011	1988	–1.2	0.21

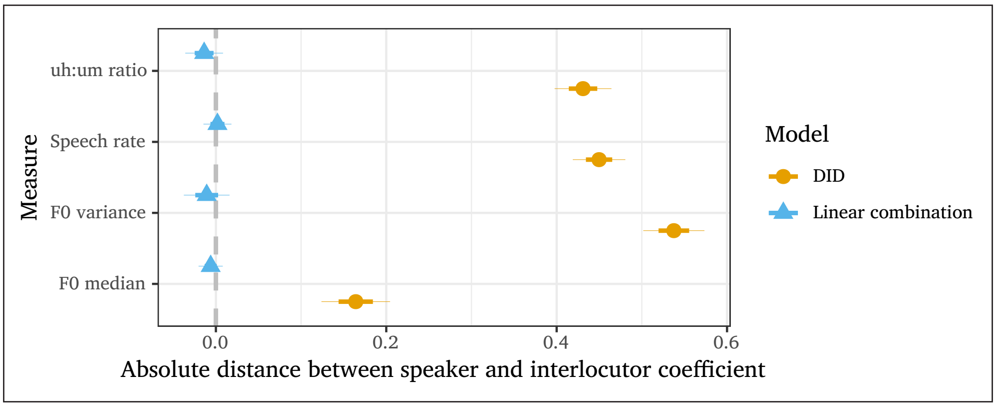

Figure 1: A comparison between the coefficients of DID and linear combination models for the four measures in Study 1. Each point is the estimate for the measure in that model, the thick lines are one standard error in each direction, and the thin lines are two standard errors in each direction. The two types of models are distinguished by color and shape. A dashed line marks zero; the linear combination model coefficients are all close to zero, having no effect, while the DID model coefficients are much larger.

3.2. Study 2: Extreme baseline values appearing as convergence in DID models

3.2.1. Introduction

Another shortcoming of the DID approach follows from the likely interpretation of regression to the mean as convergence, which the linear combination approach does not do. As shown in this study, this effect would make it likely that speakers whose initial measurements are extreme would appear to converge, even though in many cases they are simply reverting to less extreme values.

The prediction that DID would be susceptible to regression to the mean relies on two components. First, regression to the mean predicts that extreme values are less likely to be repeated, meaning that if the measured baseline value for a subject is extreme, the subject's actual performance is likely to be closer to the population mean. Second, given that interlocutors are more likely to be close to the mean in a unimodal distribution, initial extreme values for a subject are likely to be further away from the interlocutor's baseline than the subject's actual performance is. In a DID model, a shift to more typical productions would be interpreted as convergence. In contrast, a linear combination model measures convergence as the amount of variance in speakers' behavior that is predicted by the interlocutors' behavior; in these models, variance does not have to be attributed to the interlocutors' baseline, which better captures the range of factors that cause speakers to vary.

The overestimation of convergence is more likely to be an issue if subjects' measured baselines are noisy, and thus more likely to differ from their actual baselines. Measured baselines that are not representative are a particular risk when they are established based on a small set of items. If measurement of each subject's baseline is based on a large amount of data, this bias can be reduced, but it will not be fully eliminated as long as speakers produce variation not driven by their interlocutors.

It is not clear whether a real relationship between convergence and distance between speakers should be expected. It is possible that subjects with more extreme baselines are more likely to converge than subjects with more central baselines, though this might depend on the measure or be directional; for example, fast speakers might slow down, but slow speakers might be more limited in how much they can speed up. Some studies have noted that there is more convergence between individuals who start with more distinct productions based on speaking different dialects, suggesting that this is because they have more room for convergence (e.g., Babel, 2010; Walker & Campbell-Kibler, 2015). In contrast, Kim, Horton, and Bradlow (2011) found that speakers converge less when their speech differs more from their interlocutors', based on speaking different dialects or having different native languages; however, their XAB measures of similarity differed from most other studies in that the items being compared were intonational phrases rather than words and the phrases were not identical, which may be responsible for the different results.

To test this prediction, we fit DID models that include the absolute distance from the mean of the distribution as a predictor. We contrast these models with the parallel linear combination models, in which the distance from the mean is not expected to be a predictor.

3.2.2. Methods and materials

For the DID models, we extend the models from Study 1 by adding an additional predictor, the absolute distance between the subject's baseline and the mean of the distribution. Since all the variables were already standardized, the mean was zero, so the absolute distance between the subject's baseline and the mean was equivalent to the absolute value of the subject's baseline. This resulted in the formula in (12). The coefficient of this predictor will be positive if our claim holds, signifying that more extreme initial values are more likely to result in higher DID values. We fit this model for each of the four measures.

(12)
```
DID ~ 1 + abs(subject.baseline) +
abs(subject.baseline - interlocutor.baseline) +
(1 | subject) + (1 | interlocutor) + (1 | conversation)
```

We also trained the linear combination equivalent model for each measure. We extend the linear combination models in Study 1 by adding the interaction term between the interlocutor's baseline (which measures convergence) and the absolute difference between the subject's baseline and the mean. Since all the variables were already standardized, the mean was zero, so this difference was equivalent to the absolute value of the subject's baseline. This resulted in the formula in (13), in which the interaction terms on the second and third lines are the variables of interest.

(13)
```
subject.value ~ 1 + subject.baseline + interlocutor.baseline +
interlocutor.baseline:abs(subject.baseline) +
interlocutor.baseline:abs(subject.baseline - interlocutor.
baseline) + (1 | subject) + (1 | interlocutor) +
(1 | conversation)
```

In both sets of models, each data point represents one conversation side.

3.2.3. Results and discussion

All models and data are available in the supplementary materials, and summarized below.

In three of the four measures (excluding F0 median), the absolute distance between the subject's baseline performance and the mean of the distribution was positively associated with higher DID values, as shown in (14). This means that these models were more likely to find convergence for subjects whose initial values were more extreme, consistent with our predictions.

(14) Study 2: Regression results for the absolute distance between the subject's baseline performance and the mean of the distribution for DID models, across four measures.

	β	*SE*	*df*	*t*	*p*
F0 median	0.034	0.04	3683	0.8	0.4
F0 variance	0.187	0.03	495	6.8	< 0.0001
Speech rate	0.147	0.02	4645	5.9	< 0.0001
uh:um ratio	0.119	0.03	4663	4.5	< 0.0001

The distance between the subjects' and interlocutors' baselines (the focus of Study 1) was still positively correlated with higher DID values, as shown in (15), which suggests that these are two distinct effects. As before, the estimate for the intercepts was negative in all measures, signifying that DID was predicted to be negative for small differences in baselines (16).

(15) Study 2: Regression results for the absolute distance between the subject's baseline and the interlocutor's baseline performance for DID models, across four measures, based on the revised formula in (12).

	β	*SE*	*df*	*t*	*p*
F0 median	0.16	0.02	3562	7.8	< 0.0001
F0 variance	0.48	0.02	3068	23.9	< 0.0001
Speech rate	0.40	0.02	2844	22.3	< 0.0001
uh:um ratio	0.39	0.02	3584	21.0	< 0.0001

(16) Study 2: Regression results for the DID models' intercepts, across four measures, based on the revised formula in (12).

	β	*SE*	*df*	*t*	*p*
F0 median	–0.20	0.04	2992	–4.7	< 0.0001
F0 variance	–0.68	0.03	452	–23.7	< 0.0001
Speech rate	–0.56	0.02	1557	–22.6	< 0.0001
uh:um ratio	–0.53	0.03	1996	–19.5	< 0.0001

There was no robust effect of distance from the median as a predictor of convergence for F0 median, in contrast to the DID models for the other three measures. This lack of effect is likely a result of F0 median having two distinct modes, as shown in **Figure 2**. Male and female speakers have little overlap in their F0 median values, so the mean of the joint male and female distribution is not a meaningful reference point. Instead, we may expect regression to the mean to be reflected by shifts of male and female speakers towards the respective means of the distribution of their own group. To test this hypothesis, we retrained the models, replacing absolute distance from the mean with absolute distance from the nearest mode (the nearest peak of the distribution). Indeed, in this post-hoc model, all four measures exhibited a significant correlation between high DID values and distance from the nearest mode (17).

(17) Study 2: Regression results for the absolute distance between the subject's baseline performance and the nearest mode of the distribution for DID models, across four measures.

	β	*SE*	*df*	*t*	*p*
F0 median	0.166	0.06	3674	2.6	0.00938
F0 variance	0.151	0.04	3660	3.9	0.00012
Speech rate	0.205	0.04	4720	5.1	< 0.0001
uh:um ratio	0.094	0.03	4667	3.2	0.00120

In contrast, the linear combination models found no significant effect for the interaction between absolute distance from the mean and the interlocutors' baseline; that is, there was no interaction between convergence and subjects' baseline distance from the mean, as shown in (18). This suggests that there is no actual effect of the distance between subjects'

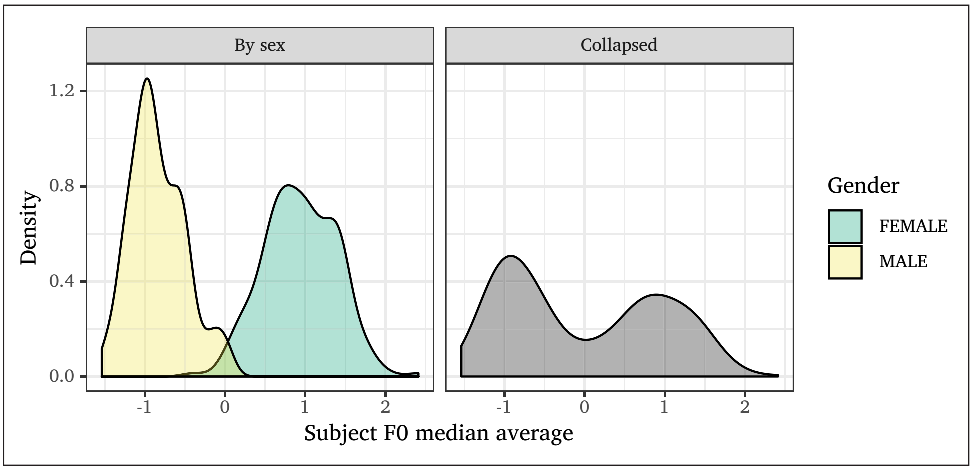

Figure 2: A density plot of F0 values, averaged for each subject. The values are split by sex on the left panel, and collapsed on the right panel.

baseline and the mean on convergence. As in Study 1, the initial distance between the subject and the interlocutor was not significant (19). **Figure 3** illustrates the differences between the coefficients in the DID and linear combination models. The DID model coefficients are visibly much larger than the linear combination model coefficients, which are consistently close to zero.

(18) Study 2: Regression results for the interaction between the absolute distance between the speaker's baseline and the mean, and the interlocutor's baseline performance, across four measures, for linear combination models.

	β	*SE*	*df*	*t*	*p*
F0 median	–0.0085	0.01	266	–0.8	0.41
F0 variance	0.0078	0.02	371	0.4	0.72
Speech rate	0.0122	0.01	286	0.8	0.40
uh:um ratio	0.0113	0.02	375	0.6	0.52

(19) Study 2: Regression results for the interaction between the interlocutor's baseline, and the absolute distance between the speaker's baseline and the interlocutor's baseline, across four measures, for linear combination models, based on the revised formula in (13).

	β	*SE*	*df*	*t*	*p*
F0 median	–0.0045	0.007	3312	–0.6	0.54
F0 variance	–0.0123	0.014	1333	–0.9	0.38
Speech rate	–0.0014	0.009	1118	–0.2	0.88
uh:um ratio	–0.0157	0.011	2093	–1.4	0.17

These results suggest that the DID approach is prone to two distinct types of artifacts: Extreme baseline values for a subject can result in the overestimation of convergence, and small initial difference between a subject and interlocutor can result in the overestimation

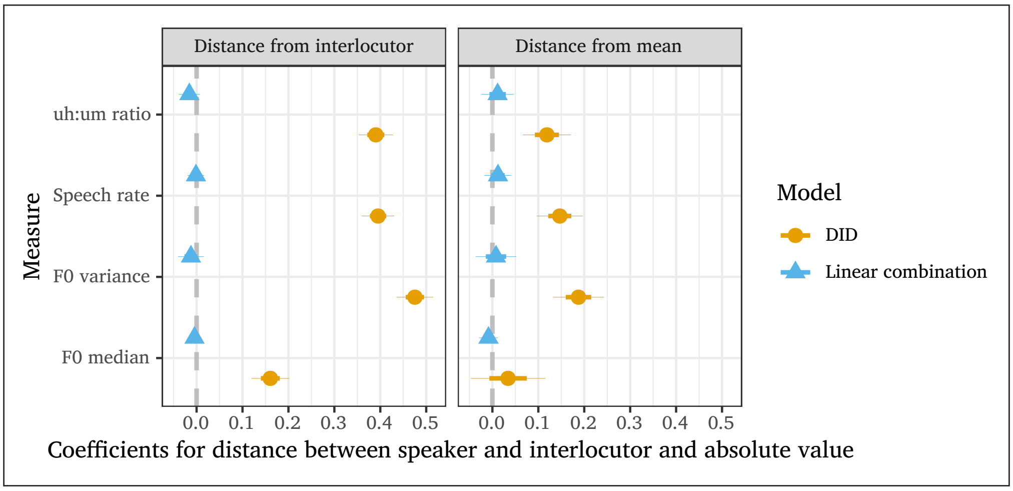

Figure 3: A comparison between the coefficients of DID and linear combination models for the four measures in Study 2 (before the post-hoc adjustment). Each point is the estimate for the measure in that model, the thick lines are one standard error in each direction, and the thin lines are two standard errors in each direction. The two types of models are distinguished by color and shape. A dashed line marks zero; the linear combination model coefficients are all close to zero, having no effect, while the DID model coefficients are much larger.

of divergence. Study 3 shows how these effects can lead to the overestimation of individual differences in convergence among subjects.

3.3. Study 3: Individual differences

3.3.1. Introduction

The goal of this study is to show that it is likely that the effect of absolute initial differences between interlocutors (as shown in Study 1) and regression to the mean (as shown in Study 2) could inflate the appearance of individual differences in convergence, regardless of whether or not such differences exist in the underlying data. Using the same dataset we are examining here, Cohen Priva and Sanker (2018) did not find any individual differences with the linear combination approach.

Recent years have seen rising interest in identifying predictors of individual differences in convergence (e.g., Lewandowski, 2012; Lev-Ari, 2018; Weatherholtz, Campbell-Kibler, & Jaeger, 2014; Yu et al., 2013); studies vary in what linguistic characteristic they measure, but studies of phonetic characteristics often use DID. Some work has found consistency in individuals' convergence across replication of the same task or similar tasks (e.g., Sanker, 2015; Tamminga, Wade, & Lai, 2018), but many of these studies do not re-test individuals to establish that their tendencies in convergence are consistent. Other studies have found much less evidence for individual consistency in phonetic convergence; these results might reflect the way convergence is measured or be due to larger differences across tasks. Pardo et al. (2018) found weak individual tendencies in convergence across a shadowing task and a conversation, using AXB to measure convergence. Cohen Priva and Sanker (2018) and Cohen Priva and Sanker (n.d.) found even weaker individual tendencies in convergence, across conversations with different partners and conversational topics, using linear combination models to measure convergence.

The results of Study 1 and Study 2 suggest that the DID approach might inflate or artificically produce individual differences in convergence. Study 1 found that in every measure, subjects were more likely to appear as divergent if their baseline values were close to their interlocutors' baseline values. Study 2 found that extreme subject baseline values relative to the mean of the distribution are likely to appear as convergent, and that these effects exist even when controlling for the findings of Study 1. These two effects can influence our estimation of individual differences in convergence as measured by DID. If the appearance of individual differences in convergence is largely an artifact of how convergence is measured, rather than reflecting actual behavior differences, this might explain why studies looking for individual tendencies in convergence across different characteristics have found no such tendencies (e.g., Bilous & Krauss, 1988; Cohen Priva & Sanker, n.d.; Pardo et al., 2012; Weise & Levitan, 2018), aside from those due to relationships between the measures (e.g., F0 mean and F0 variability, Cohen Priva & Sanker, 2018), and the occasional significant correlation in studies making a large number of comparisons (e.g., Sanker, 2015).

Subjects whose baseline values are close to the distribution's mode are more likely than others to be close to their interlocutors' baselines, if the interlocutors come from the same distribution. Therefore, they are more likely than others to have negative (divergent) DID values, as Study 1 shows. Subjects whose baseline performance is distant from the mode are more likely to appear convergent, as Study 2 shows. We therefore predict that distance from a mode of the distributions would affect the detection of individual differences. A correlation between distance from the mode and convergence is not expected when convergence is measured in other means, namely using linear combination models, and thus individual variation in distance from the mode should not produce apparent individual differences in convergence in such models.

3.3.2. Materials and methods

In contrast to Study 1 and and Study 2, the focus of the investigation here is the individual. For DID-based models, we estimate individual differences in convergence as the mean per-conversation DID values for each subject. For linear combination models, individual differences were measured as the per-subject random slope for the interlocutor's baseline, as in (5). Distance from the mode was calculated as the absolute distance of the subject's mean performance from the closest mode.

Since the model has no repeated values per subject, we used a simple linear regression to test for a possible correlation between distance from a mode of the distribution and individual differences in convergence. We repeated this analysis for each of the four measures, for both the DID models and the linear combination models.

3.3.3. Results

The results for the two models are summarized in (20). For DID models, there was a positive correlation between mean DID and mean performance for all measures; that is, the subject's distance from the nearest mode was related to measured convergence. For the linear combination models, there was a much smaller relationship between per-individual slopes and mean performance, which did not consistently reach significance. **Figure 4** shows the correlation between per-subject mean DID estimates and mean subject performance, for the DID models. **Figure 5** shows the relationship between subjects' mean performance and their convergence slopes for the linear combination models; no significant trend is found.

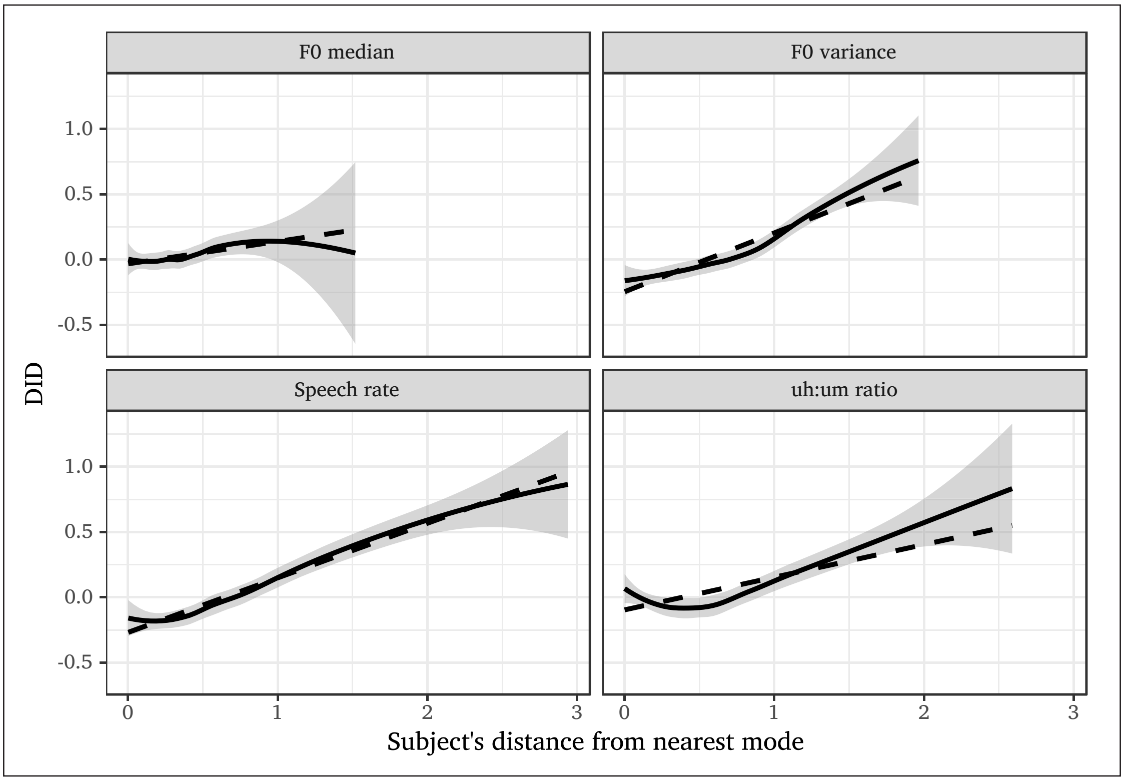

Figure 4: The relationships between proximity to a mode and DID values. The solid line shows the relationship between the variable using local polynomial regression (loess). The dashed line is the linear relationship. The X-axis is the absolute distance between the subject and the nearest mode. The Y-axis is the DID value. Each panel shows the relationship for a different measure. The results show a clear relationship between proximity to a mode and high DID, though it is much weaker for F0 median than for other characteristics.

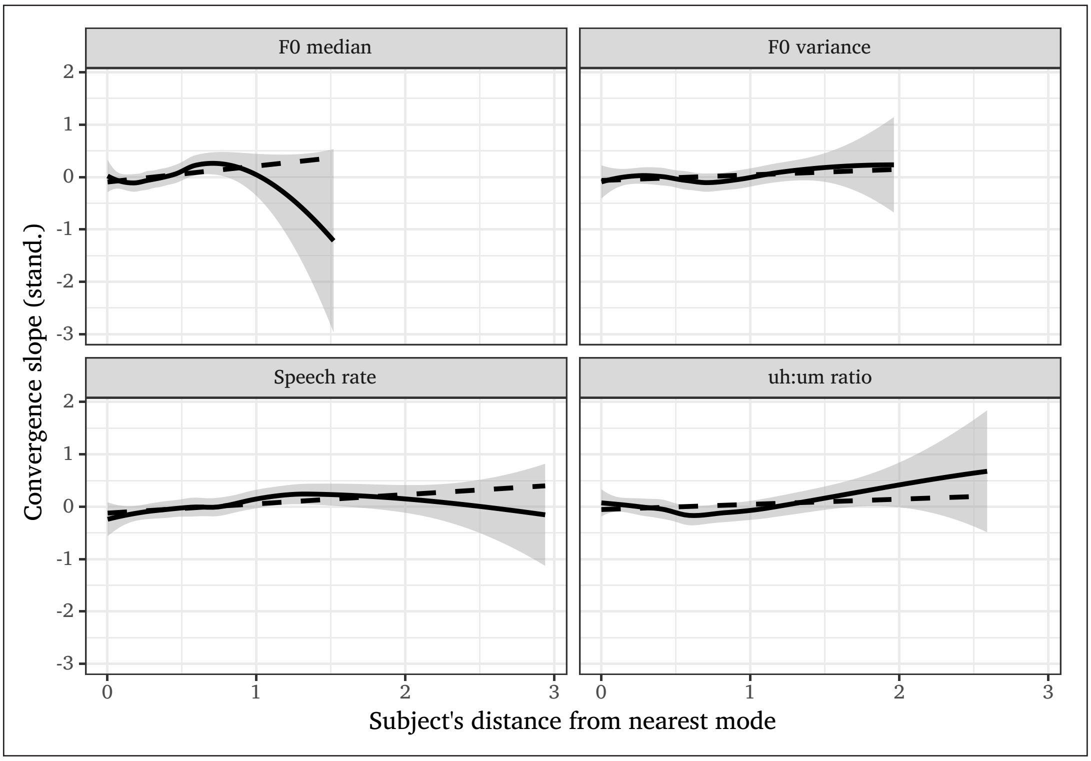

Figure 5: The relationships between proximity to a mode and individual differences in convergence, as measured using a random slope for interlocutors' baseline in a linear combination mixed effects model. The solid line shows the relationship between the variable using local polynomial regression (loess). The dashed line is the linear relationship. The X-axis is the absolute distance between the subject and the nearest mode. The Y-axis is the standardized per-subject convergence slope value. Each panel shows the relationship for a different measure. The results do not show a clear relationship between proximity to the median and convergence.

(20) Comparison of the correlations between absolute distance from the nearest mode and the individual differences in the DID and mixed effects slope methods. In every measure, the relationship between the individual differences measure and absolute distance from the nearby mode was higher for DID than for the linear combination method.

	DID				Slope (Linear Combination)			
	β	*SE*	*t*	*p*	β	*SE*	*t*	*p*
F0 median	0.062	0.03	2.2	0.03	0.017	0.01	1.6	0.12
F0 variance	0.450	0.04	10.8	<0.0001	0.020	0.02	1.0	0.33
Speech rate	0.553	0.05	12.0	<0.0001	0.055	0.03	2.2	0.03
uh:um ratio	0.299	0.05	6.2	<0.0001	0.024	0.02	1.1	0.29

These results demonstrate major differences between the two methods of measuring convergence. The DID models produced a consistently significant correlation between the subject's distance from the nearest mode and the convergence measured for that subject, across all measures. The random slopes measured in the linear combination models did not produce significant correlations between individual convergence and distance from a mode in most measures. The key difference between this method and DID is that it does not assign value to the initial distance between the subject and the interlocutor, and leaves room for the estimation of noise. This allowance for noise and decreased reliance on starting distance make the linear combination method better suited to measuring convergence for particular individuals, without inflating convergence for subjects whose baselines were far from their interlocutors' and understimating convergence for subjects with values close to a mode. The one measure in which the linear combination method found a statistically significant correlation between a subject's convergence and baseline distance from the closest mode was speech rate, but as **Figure 5** shows, that trend was minimal, was not present in the loess trend (as opposed to DID trends for the same data), and did not remain when *p*-values were corrected for multiple comparisons.

3.4. Study 4: Sampling in a simulated convergence dataset

3.4.1. Introduction

The results in Studies 1–3 demonstrate clear differences in the extent to which DID and linear combination models find individual differences in convergence. We argue that these apparent effects found by DID models are purely mathematical artifacts of how convergence is measured, rather than reflecting actual relationships between convergence and distance from the population mean and from the interlocutor. However, these measurements use existing studies of convergence (Cohen Priva & Sanker, 2018), so we cannot rule out a priori the possibility that the effects found by the DID approach really do exist within the data. Study 4 therefore uses simulated data (cf. Cohen Priva & Jaeger, 2018), in which assumptions about convergence are minimized. Within the generated dataset, the DID approach still finds spurious convergence effects.

The parameters were defined to mirror a typical design of a convergence study on a single phonetic measure, in which each participant participated in a single interaction, with testing *before* and *after* the interaction. To avoid built-in assumptions about how convergence should be defined, which could create a bias towards one model or another, the dataset was defined to lack convergence. Because there is no convergence, there are also no predictors of convergence.

3.4.2. Methods and materials

Data was generated with 50 interacting pairs of speakers. The true baseline of each participant was sampled from a normal distribution of the population, with a mean of 0 and a standard deviation of 1. The mean of zero parallels the normalized data used in the preceding studies.

The *before* and *after* performances for each participant were calculated from the participant's baseline value with normally distributed noise with a mean of 0 and a standard deviation of 0.5. This yielded a Pearson correlation of about 0.8 between *before* and *after* values, which is similar to the correlation between baseline performance and actual performance for speech rate and *uh*:*um* ratio in Cohen Priva and Sanker (2018), which was based on the same dataset used for Studies 1–3.[3] Self-correlation in the actual data varies by measure; some characteristics have a correlation higher than 0.8 and some have a lower correlation. In this simulated data, *after* values were sampled without reference to the interlocutor's productions; there was no convergence.

We calculated DID values for each participant, and performed a linear regression model similar to the mixed effects models for DID in Studies 1–3. For this model, we tested whether the starting distance between the speaker and the mode of the distribution (assumed to be zero) or the starting distance between the speaker and interlocutor would be correlated with DID values. The formula is provided in (21), in which `DID` is the difference-in-difference, `abs(subject.before)` is the absolute distance of the *before* measurement

[3] See also the supplementary materials for this paper, under Study 3.

from the mean (zero, in this case), and `abs(subject.before - interlocutor.before)` is the difference between the speakers' baseline and their interlocutors' baseline. In a real experiment, there would be a random intercept for the conversation, but we did not include conversation-based effects, and replacing simple linear regression with mixed effects regression would have greatly decreased the number of samples we could have included (as mixed effects models take much longer to converge), so we omitted that term.

(21)
```
DID ~ 1 + abs(subject.before) +
abs(subject.before - interlocutor.before)
```

We also performed the equivalent linear combination model, following the modeling approach for the linear combination model in Studies 1–3. In this model, the convergence parameters are the interaction terms between the interlocutors' baseline, and both the absolute distance of the speaker from the mode as well the absolute distance between the speaker and their interlocutor (22), in which `subject.after` is the subjects' performance after the interaction, `subject.before` is their baseline (*before* the interaction) `interlocutor.before` is their interlocutors' baseline, and the interaction terms correspond to the DID variable of interests. The first (on the second line of the formula) aims to capture the effect absolute distance from the mean would have on convergence, and the second (on the third line of the formula) aims to capture the effect that an initial distance between the speaker and the interlocutor may have on convergence. The last three terms are expected to be non-significant, because the model is defined to not have convergence. We omitted the conversation random intercept that a real model would have included, in order to make it possible to draw more samples.

(22)
```
subject.after ~ 1 + subject.before + interlocutor.before +
interlocutor.before:abs(subject.before) +
interlocutor.before:abs(subject.before - interlocutor.before)
```

We repeated this sampling procedure ten thousand times, and provide summary results for the models below.

3.4.3. Results and discussion

For the ten thousand samples, the median correlation between each subject's *before* and *after* values was 0.802.

In DID models, the absolute distance of the subjects' distance from the mean resulted in statistically significant ($p < 0.05$) positive coefficients 24.7% of the time, which is greater than what is predicted by chance. The coefficient for the absolute distance between the speakers and their interlocutors was statistically significant and positive 70.8% of the time.

In the linear combination models, in contrast, the absolute distance of the subject from the mean resulted in statistically significant positive coefficients 2.4% of the time, and the coefficient for the absolute distance between the speakers and their interlocutors was statistically significant and positive 2.5% of the time. Both of the results for the linear combination models are within what would be expected by chance. **Figure 6** shows the density plots of the coefficients in DID and linear combinations. It is evident that the distribution of t-values in the linear combination models is centered around zero, and though for a small fraction of the models a positive correlation is found, negative correlations are similarly likely to be found. In contrast, in the DID models the coefficients

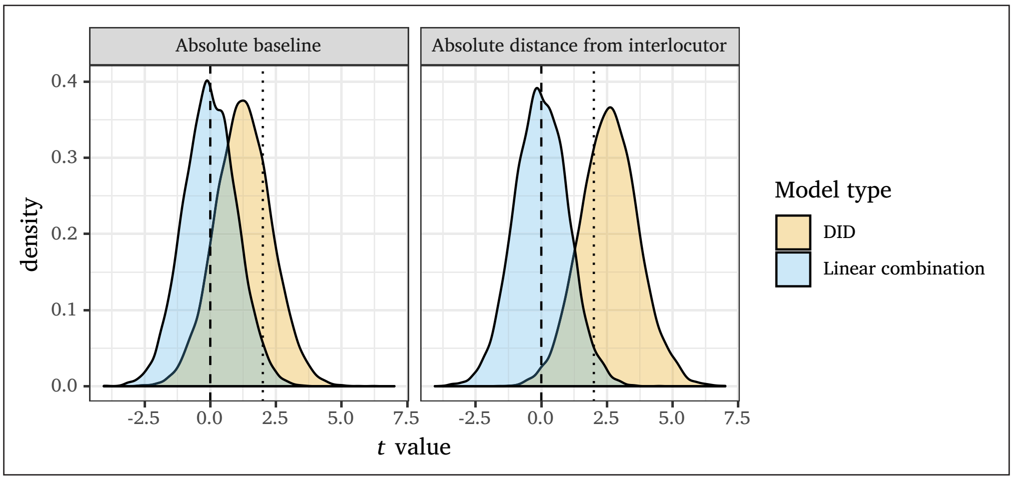

Figure 6: Density plots of *t*-values for the two coefficients in the 10,000 samples taken in Study 4, for both the DID models and in the linear combination models. The dashed line marks 0, where the non-existent effect is supposed to be. The dotted line marks 2, which is roughly the point at which the *t*-value woulds appear to be significant. The distributions of the linear combination models are centered around zero, while the DID *t*-values are shifted such that false positive correlations are more likely to be found.

are systematically biased to be positive, making spurious positive values more likely to be found than chance alone would predict.

The results clearly indicate that even in the complete absence of underlying relationships between convergence and the participant's distance from the mode or the starting distance between the participant and the interlocutor, such effects often spuriously emerge in DID models. When studies then look for individual differences in convergence as measured this way, these effects are likely to be interpreted as evidence for differences in individual tendencies in convergence, even though the true variation across individuals is just in their measured baselines. Because distance from the mode correlates with convergence, speakers whose mean performance is exceptional will seem to converge more than other speakers do, while speakers whose mean performance is close to their interlocutors will seem less convergent or even divergent.

We argue that the spurious effects in the DID model are due to mishandling the noise that exists between measurements of an individual. This predicts that spurious effects would be more likely in more noisy measurements than in less noisy ones. We therefore replicated the results presented above with varying amounts of noise, between 0.1 and 2 standard deviations, as sampled from a uniform distribution (that parameter was fixed at 0.5, as discussed above). Indeed, less noise translates to lower *t*-values for both variables in DID models, and more noise translates to higher *t*-values for both variables. The values seem to peak at high degrees of noise, but that happens after self-consistency drops below Pearson $r = 0.5$, which is not typical of phonetic variables (all the variables in our dataset have higher consistency). In contrast, the *t*-values for the two corresponding variables in linear combination models are not affected by the noise in the sample. **Figure 7** presents these results.

As with the rest of the analysis used in this paper, the code for the sampling procedures is available in the supplementary materials.

4. General discussion

Our results demonstrate that DID is not a suitable measure of convergence because it interprets regression to the mean as convergence and underestimates convergence or even finds divergence when the subject's baseline performance is close to the reference value

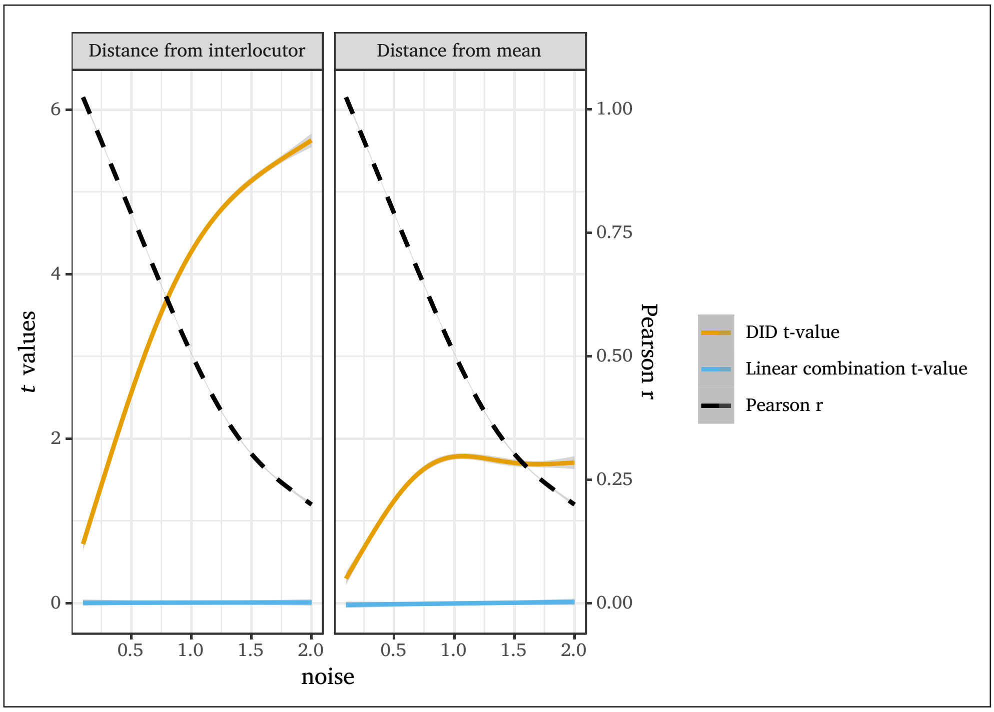

Figure 7: The relationship between speaker self-consistency, as measured by noise SD, and the *t*-values of the DID and linear combination coefficients in Study 4. The x-axis is the degree of noise, in standard deviations (0.5 in the results reported above). The y-axis (left) is the *t*-value of the coefficient. The solid lines represent the trend of the *t*-values of the two variables, one in each panel. The Pearson *r* values associated with each noise value are provided by a black dashed line (marked on the right y-axis). All lines were smoothed by the gam method, using a 4-dimensional spline function for greater smoothing. It is evident that finding convergence is not predicted by the amount of noise present in the data when using linear combination models, but amount of noise does affect DID measures of convergence.

of the interlocutor or model talker. These biases pose a particular problem in estimation of individual differences in convergence. Our proposed alternative, linear combination, is not subject to any of these issues and thus provides more reliable estimates of the convergence exhibited by each individual participant.

Measuring convergence as change in distance can create biases due to the starting distance between the speakers. The natural variability of speakers can make subjects appear to diverge from interlocutors whose baselines are close to their own, while variation at greater starting distances is more likely to appear convergent. In Study 1, we demonstrate that the subject's distance from the interlocutor is a predictor of convergence measured within DID models; greater baseline distances from the interlocutor produce higher measurments of convergence, and small baseline distances can create the appearance of divergence. In contrast, there is no relationship within linear combination models. The effects are parallel across all four linguistic characteristics tested (F0 median, F0 variability, speech rate, and uh:um ratio). We argue that the apparent relationship within the DID models is a purely mathematical artifact of how convergence is measured, not based on any actual behavioral pattern. In Study 4, we demonstrate that this apparent relationship within DID models also arises in simulated datasets which have been defined to lack such a relationship.

Consistent with a lack of actual relationship between starting distance and convergence, previous work does not clearly predict behavioral differences based on starting distance. Some studies have found that there is more convergence between individuals whose starting distance is greater, when the distance is across dialects (e.g., Babel, 2010; Walker

& Campbell-Kibler, 2015). However, they do not argue that greater distance makes speakers more inclined to converge; their interpretation is that greater distance makes convergence possible, while there is simply less possibility for convergence for speakers who already have close baselines. Furthermore, they make no predictions that baseline similarity should result in divergence. It is additionally unclear whether convergence across dialects and specifically convergence to characteristics that differ across those dialects reflects the same process as convergence as it might relate to distance within dialects. The Switchboard data used in our studies includes speakers from a range of American English dialects.

In contrast, Kim et al. (2011) found that speakers converge less when their interlocutors have more distinct speech characteristics based on being from different speech communities; however, the perceptual evaluations of similarity in their study were based on non-identical intonational phrase groups, rather than identical single words, so the results may not be comparable to results of related work. Enzinna (2018) suggests that convergence is more modulated by social effects of how representative the model talker is of the majority population in the subject's speech community than by effects of distance per se between the subject and the model talker, based on finding more convergence to speakers representative of the majority, both in English monolingual majority populations with long VOT and English-Spanish bilingual majority populations with shorter VOT. Babel et al. (2014) similarly found more convergence to voices that were rated as more typical of the gender of the speaker. Gradience of the differences may also play a role; in a shadowing task, Mitterer and Ernestus (2008) found that most Dutch speakers who had an alveolar /r/ never converged to uvular place of articulation.

While some work has addressed potential effects of the prototypicality of the model talker or interlocutor in eliciting convergence, the prototypicality of each subject has been largely overlooked. Looking at convergence patterns produced by L2 speakers, Lewandowski (2012) found that more proficient L2 speakers converged more in that L2 than less proficient L2 speakers, which suggests that at least across native languages, greater distance does not facilitate greater convergence. Enzinna (2018), testing subjects who were prototypical of their speech communities or not in being monolingual English speakers or English-Spanish bilinguals, did not find a strong effect of the subject's prototypicality on convergence, though she did find an effect of the model talker's prototypicality. The individual contributions of each speaker to convergence in dyadic conversations are obscured in DID measurements, so the possible separate effects of prototypicality of the subject and the interlocutor will not be apparent in conversational tasks. It is not clear what might drive greater convergence for subjects who are more distant from the typical productions of the population. In Study 4, we demonstrate that DID models will find a relationship between participants' distance from the mode and their degree of convergence even in simulated data that has been defined to lack such relationships, confirming that the effect is an artifact rather than an actual behavioral pattern.

Convergence depends on establishing clear baselines for the subjects. When the baselines could be extreme due to noise, returning to true baselines is likely to appear convergent, as long as the subject and the interlocutor come from the same population. In Study 2, we demonstrate that greater distance of the subject from the median or the nearest mode also produces higher measured convergence in DID models. In bimodal distributions, as for F0, the distance from the median is not a significant predictor, but distance from the nearest mode is consistently a predictor across the four measures. In contrast, there was no relationship within the linear combination models. The effect of distance from the population mode seems to be a different effect than distance from the interlocutor, because both are significant within the same models.

If starting distance from the interlocutor or the population mode produces biases in the measurement of convergence for each individual, this can produce apparent individual differences in convergence, even though the actual differences are simply in the baselines. In Study 3, we demonstrate that DID models consistently find significant individual differences, across all four measures. In contrast, there was no relationship in three of the linear combination models; the significance of the apparent weak relationship within speech rate disappeared when correcting for multiple comparisons. This result throws into question work that looks for individual tendencies in convergence. Within phonetic convergence, many studies of individual differences use DID, and they often do find individual differences (e.g., Lewandowski, 2012; Yu et al., 2013), which may even be consistent when individuals repeat a task, either with exactly the same parameters or with similar parameters (e.g., Sanker, 2015; Tamminga et al., 2018). Studies not using DID have found less evidence for individual tendencies in convergence, e.g., using AXB perception (Pardo et al., 2018) or linear combination models (Cohen Priva & Sanker, 2018). Some of the differences between studies may relate to how similar the two convergence tasks for each participant were; Pardo et al. (2018) compared the same individuals in shadowing and in conversation, and Cohen Priva and Sanker (2018) used conversations with different partners and different conversational topics.

Shadowing studies have also found differences in convergence across model talkers (e.g., Pardo et al., 2017; Babel et al., 2014); in conversational tasks, it is often difficult or impossible to separate effects of the speaker and the interlocutor, so possible interlocutor effects are not often tested. Paralleling measurements of individual tendencies to converge, results for differences across interlocutors in how much convergence they elicit may simply reflect differences in starting distance. However, some effects of the interlocutor may exist independently of starting distance; Cohen Priva and Sanker (n.d.), using linear combination models with a large number of speakers and interlocutors, found significant per-interlocutor differences in convergence.

Some of the measurement biases produced by DID can be reduced by measuring distance or change in distance between subjects and their actual interlocutors or model talkers as compared to distance between subjects and speakers or model talkers they did not interact with (e.g., Levitan & Hirschberg, 2011; Miller et al., 2014; Sanker, 2015). The same artifacts in measured convergence based on individuals' starting distance from the interlocutor or starting distance from the population mean will be present for the real pairs and the pseudo-pairs. To the extent that it is possible to match starting distances between the real pairs and the pseudo-pairs, this comparison may compensate for the biases produced by difference in difference.

Biases due to starting distance are also likely to create different patterns based on the characteristic in which convergence is measured; different measures have different distributions by speaker, so the biases could create more issues for some measures than others. As demonstrated in Study 4, greater variability increases convergence found by DID, even in data defined to lack convergence. If apparent individual differences in convergence are due to how convergence has been measured, rather than reflecting actual individual tendencies in convergent behavior, this could explain the lack of evidence for individual tendencies in convergence across measures (e.g., Bilous & Krauss, 1988; Pardo et al., 2012; Weise & Levitan, 2018). Across studies looking for such tendencies, the only significant effects can be attributed to physical relationships between the measures (e.g., F0 mean and F0 variability, Cohen Priva & Sanker, 2018), or chance correlations in studies with a large number of comparisons (e.g., Sanker, 2015). If individual tendencies in convergence do exist, it is unclear why they would be specific to particular measures, as broad differences in characteristics such as attention to detail or social engagement should be reflected similarly in convergence across different characteristics.

Effects of starting distance from the interlocutor and distance to a mode on measurement of convergence could also create the appearance of different convergent behaviors across subgroups of the population, particularly if the variability of the particular measure is different within the two groups. Many studies have examined gender as a predictor of convergence; there are sometimes significant differences in convergence based on the gender of the subject or the model talker or interlocutor, or an interaction between the genders of the two speakers, either in overall convergence or in interactions with other factors (e.g., Bilous & Krauss, 1988; Namy, Nygaard, & Sauerteig, 2002; Pardo, 2006; Pardo et al., 2018). The results are inconsistent across studies; for example, Namy et al. (2002) found more convergence among women, while Pardo (2006) found more convergence among men. Other studies have found no difference between men and women (e.g., Pardo et al., 2010, 2017). However, not all of these studies used DID, so the complicated interactions between gender and convergence may go beyond effects of choice of measure and effects of how convergence is measured.

It is also possible that effects of the subject's starting distance relative to the interlocutor or to the population mean could create apparent effects across words. Studies have often found more convergence in lower frequency words than in higher frequency words (e.g., Goldinger, 1998; Dias & Rosenblum, 2016). Low frequency words are particularly prone to hyperarticulation at first exposure; this could result in the measured baselines for low frequency words being particularly poor representatives of the subjects' true baselines, and particularly if the recordings of the model talker are not taken from their first production, this could increase apparent convergence when subjects' subsequent productions are produced more naturally. Nielsen (2011) used the second production of each word as the baseline instead of the first, after a 'warm-up' block of reading each test item, and found a smaller effect of lexical frequency on convergence than other studies have, though it was still significant. If the effect of lexical frequency on convergence is purely a result of hyperarticulation in elicitations of lower frequency words, it is possible that taking a baseline after a larger number of repetitions would eliminate the effect entirely.

In addition to DID measurements of convergence, many studies use AXB tasks to obtain holistic judgments of similarity. Given that AXB is based on testing change in perceived distance, it is possible that it would be subject to some of the same issues as DID. However, as long as listeners in the AXB task are making decisions based on a range of features, as is demonstrated by Pardo et al. (2017), most of the biases should be eliminated, because speakers will generally not consistently have the same baseline difference from the model talker across different measures. There is nonetheless some possibility for artifacts arising when distance in multiple characteristics aligns, as will be the case for F0 and formant measurements across genders.

Another possible issue in the measurement of convergence is normalization by speakers, which might reduce the variation in starting distances between subjects and interlocutors or model talkers. However, it is unclear whether normalization is a better or worse representation of how listeners perceive phonetic input that varies across speakers than raw acoustic measurements. Some studies of convergence measure speakers' production as raw values, while others normalize by speaker. F0 is more often treated as a raw value (e.g., Pardo et al., 2017; Babel & Bulatov, 2011), though some work has normalized by gender (Weise & Levitan, 2018). Measurement of formants is more often normalized by speaker (e.g., Babel, 2010; Pardo et al., 2017), though some work does not normalize (e.g., Delvaux & Soquet, 2007). Some models of perception include perceptual normalization to map phonological categories across speakers (e.g., Joos, 1948; Nordström & Lindblom, 1975), given that listeners are accurate at identifying phonological category membership despite wide variation across speakers. However, listeners' ability to process different speakers' vowels as representatives of the same category does not require normalization per se that transforms all speakers' vowel spaces to align with each other (Johnson, 1997). No convergence work has tested whether normalized or non-normalized data provides a better measure of how listeners are influenced by the voices they hear.

5. Conclusion

We demonstrate several issues with the commonly used difference-in-difference (DID) method for measuring convergence as change in absolute distance between a subject and an interlocutor or model talker. Close baselines for the subject and interlocutor produce underestimation of convergence or apparent divergence, and greater distance from the mode(s) of the population produces overprediction of convergence.

These biases in measurement of convergence can produce the appearance of individual differences in convergence, which can have consequences in motivating directions of future work in individual variation and consistency. Because individual variation in production varies by measure, these biases due to starting values can also make the measurement of convergence more sensitive to the particular characteristic examined.

The alternative method that we propose for measuring convergence, using linear combination models, does not exhibit the same biases, and thus provides a more reliable measure of convergence, particularly for comparing that convergence across individuals and across speech characteristics.

Additional File

The additional file for this article can be found as follows:

- A zip file containing an Rmarkdown file that presents all the code used to create the models in this paper (`limitations-of-DID-code.Rmd`), a compiled pdf version of that Rmarkdown file (`limitations-of-DID-code.pdf`), and a compressed tab-separated file containing the data used in the paper (`limitations-of-DID-data.tsv.bz2`). DOI: https://doi.org/10.5334/labphon.200.s1

Acknowledgements

The authors would like to thank Bodo Winter, an anonymous reviewer, as well as Elinor Amit, Scott AnderBois, Roman Feiman, and Jim Morgan for their helpful feedback on this paper.

Competing Interests

The authors have no competing interests to declare.

Author Contribution

Uriel Cohen Priva and Chelsea Sanker contributed equally to this manuscript.

References

Abel, J., & **Babel, M.** (2017). Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech*, *60*(3), 479–502. DOI: https://doi.org/10.1177/0023830916665652

Acton, E. K. (2011). On gender differences in the distribution of um and uh. In *University of Pennsylvania working papers in linguistics* (Vol. 17).

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, *39*, 437–456. DOI: https://doi.org/10.1017/S0047404510000400

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189. DOI: https://doi.org/10.1016/j.wocn.2011.09.001

Babel, M., & **Bulatov, D.** (2011). The role of fundamental frequency in phonetic accommodation. *Language and Speech*, *55*(2), 231–248. DOI: https://doi.org/10.1177/0023830911417695

Babel, M., McGuire, G., Walters, S., & **Nicholls, A.** (2014). Novelty and social preference in phonetic accommodation. *Laboratory Phonology*, *5*(1), 123–150. DOI: https://doi.org/10.1515/lp-2014-0006

Bilous, F. R., & **Krauss, R. M.** (1988). Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. *Language & Communication*, *8*(3), 183–194. DOI: https://doi.org/10.1016/0271-5309(88)90016-X

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387. DOI: https://doi.org/10.1016/0010-0285(86)90004-6

Branigan, H. P., Pickering, M. J., & **Cleland, A. A.** (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*, B13–25. DOI: https://doi.org/10.1016/S0010-0277(99)00081-5

Chartrand, T. L., & **Bargh, J. A.** (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893. DOI: https://doi.org/10.1037/0022-3514.76.6.893

Clark, H. H., & **Fox Tree, J. E.** (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. DOI: https://doi.org/10.1016/S0010-0277(02)00017-3

Cohen Priva, U., Edelist, L., & **Gleason, E.** (2017). Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *Journal of the Acoustical Society of America*, *141*(5), 2989–2996. DOI: https://doi.org/10.1121/1.4982199

Cohen Priva, U., & **Jaeger, T. F.** (2018). The interdependence of frequency, predictability, and informativity in the segmental domain. *Linguistics Vanguard*, *4*(S2). DOI: https://doi.org/10.1515/lingvan-2017-0028

Cohen Priva, U., & **Sanker, C.** (2018). Distinct behaviors in convergence across measures. In *Proceedings of the 40th annual meeting of the Cognitive Science Society* (pp. 1515–1520).

Cohen Priva, U., & **Sanker, C.** (n.d.). Convergence is predicted by particular interlocutors, not speakers. Brown University manuscript https://urielcpublic.s3.amazonaws.com/CohenPriva_Sanker_Convergence-submitted.pdf.

Delvaux, V., & **Soquet, A.** (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, *64*(2–3), 145–173. DOI: https://doi.org/10.1159/000107914

Dias, J. W., & **Rosenblum, L. D.** (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, & Psychophysics*, *78*(1), 317–333. DOI: https://doi.org/10.3758/s13414-015-0982-6

Enzinna, N. R. (2018). *Automatic and social effects on accommodation in monolingual and bilingual speech* (PhD thesis). Cornell University.

Felker, E., Tronsco-Ruiz, A., Ernestus, M., & **Broersma, M.** (2018). The ventriloquist paradigm: Studying speech processing in conversation with experimental control over phonetic input. *Journal of the Acoustical Society of America*, *144*(4), EL304–EL309. DOI: https://doi.org/10.1121/1.5063809

Gijssels, T., Casasanto, L. S., Jasmin, K., Hagoort, P., & **Casasanto, D.** (2016). Speech accommodation without priming: The case of pitch. *Discourse Processes*, *53*(4), 233–251. DOI: https://doi.org/10.1080/0163853X.2015.1023965

Godfrey, J. J., & **Holliman, E.** (1997). Switchboard-1 release 2.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. DOI: https://doi.org/10.1037/0033-295X.105.2.251

Gregory, S. W. J., & **Webster, S.** (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology*, *70*(6), 1231–1240. DOI: https://doi.org/10.1037/0022-3514.70.6.1231

Harkins, D., Feinstein, D., Lindsey, T., Martin, S., & **Winter, G.** (2003). Switchboard MS State manually corrected word alignments. https://www .isip .piconepress .com/projects/switchboard/.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 1–136. DOI: https://doi.org/10.2307/522229

Kim, M., Horton, W. S., & **Bradlow, A. R.** (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, *2*(1), 125–156. DOI: https://doi.org/10.1515/labphon.2011.004

Lev-Ari, S. (2018). Social network size can influence linguistic malleability and the propagation of linguistic change. *Cognition*, *176*, 31–39. DOI: https://doi.org/10.1016/j.cognition.2018.03.003

Levitan, R., & **Hirschberg, J. B.** (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of interspeech*. Brisbane: International Speech Communications Association.

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence* (PhD thesis). Universität Stuttgart.

Miller, R. M., Sanchez, K., & **Rosenblum, L. D.** (2014). Is speech alignment to talkers or tasks? *Attention, Perception, & Psychophysics*, *75*(8), 1817–1826. DOI: https://doi.org/10.3758/s13414-013-0517-y

Mitterer, H., & **Ernestus, M.** (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, *109*(1), 168–173. DOI: https://doi.org/10.1016/j.cognition.2008.08.002

Namy, L. L., Nygaard, L. C., & **Sauerteig, D.** (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422–432. DOI: https://doi.org/10.1177/026192702237958

Natale, M. (1975). Social desirability as related to convergence of temporal speech patterns. *Perceptual and Motor Skills*, *40*, 827–830. DOI: https://doi.org/10.2466/pms.1975.40.3.827

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132–142. DOI: https://doi.org/10.1016/j.wocn.2010.12.007

Nordström, P.-E., & **Lindblom, B.** (1975). A normalization procedure for vowel formant data. In *Proceedings of the 8th international congress of phonetic sciences*.

Oben, B., & **Brône, G.** (2016). Explaining interactive alignment: A multimodal and multifactorial account. *Journal of Pragmatics*, *104*, 32–51. DOI: https://doi.org/10.1016/j.pragma.2016.07.002

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*(4), 2382–2393. DOI: https://doi.org/10.1121/1.2178720

Pardo, J. S., Cajori Jay, I., & **Krauss, R. M.** (2010). Conversational role influences speech imitation. *Attention, Perception, and Psychophysics*, *72*(8), 2254–2264. DOI: https://doi.org/10.3758/BF03196699

Pardo, J. S., Gibbons, R., Suppes, A., & **Krauss, R. M.** (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*, 190–197. DOI: https://doi.org/10.1016/j.wocn.2011.10.001

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & **Lewandowski, E.** (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, *69*, 183–195. DOI: https://doi.org/10.1016/j.jml.2013.06.002

Pardo, J. S., Urmanche, A., Wilman, S., & **Wiener, J.** (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*, 637–659. DOI: https://doi.org/10.3758/s13414-016-1226-0

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & **Ward, M.** (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, *69*, 1–11. DOI: https://doi.org/10.1016/j.wocn.2018.04.001

Sanker, C. (2015). Comparison of phonetic convergence in multiple measures. In *Cornell working papers in phonetics and phonology 2015* (pp. 60–75).

Schweitzer, A., & **Lewandowski, N.** (2013). Convergence of articulation rate in spontaneous speech. In *Proceedings of interspeech* (pp. 525–529).

Schweitzer, A., & **Walsh, M.** (2016). Exemplar dynamics in phonetic convergence of speech rate. In *Proceedings of interspeech* (pp. 2100–2104). DOI: https://doi.org/10.21437/Interspeech.2016-373

Stevens, S., Volkmann, J., & **Newman, E.** (1937). A scale for the measurement of psychological magnitude pitch. *Journal of the Acoustical Society of America*, *8*, 185–190. DOI: https://doi.org/10.1121/1.1915893

Tamminga, M., Wade, L. A., & **Lai, W.** (2018). Stability and variability in phonetic flexibility. Salt Lake City, Utah.

Walker, A., & **Campbell-Kibler, K.** (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, *6*, Article 546. DOI: https://doi.org/10.3389/fpsyg.2015.00546

Weatherholtz, K., Campbell-Kibler, K., & **Jaeger, T. F.** (2014). Socially mediated syntactic alignment. *Language Variation and Change*, *26*, 387–420. DOI: https://doi.org/10.1017/S0954394514000155

Weise, A., & **Levitan, R.** (2018). Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (Vol. 2, pp. 297–302). DOI: https://doi.org/10.18653/v1/N18-2048

Yu, A., Abrego-Collier, C., & **Sonderegger, M.** (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and 'autistic' traits. *PloS ONE*, *8*(9), e74746. DOI: https://doi.org/10.1371/journal.pone.0074746

How to cite this article: Cohen Priva, U., & Sanker, C. 2019 Limitations of difference-in-difference for measuring convergence. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1): 15, pp. 1–29. DOI: https://doi.org/10.5334/labphon.200

Submitted: 25 March 2019 **Accepted:** 22 July 2019 **Published:** 11 September 2019

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS