Progress in Regression: Why Natural Language Data Calls For Mixed-Effects Models

Daniel Ezra Johnson, Lancaster University
March 2010 – December 2014

Natural language data – sociolinguistic, historical, and other types of corpora – should not be analyzed with fixed-effects regression models, such as VARBRUL and GoldVarb use. This is because tokens of linguistic variables are rarely independent; they are usually grouped and correlated according to factors like speaker (or text) and word. Fixed-effects models can estimate the effects of higher-level "nesting" predictors (like speaker gender or word frequency), but they cannot be accurate if there exist any individual effects of lower-level "nested" predictors (like speaker or word). Mixed-effects models are designed to take these multiple levels of variation into account at the same time. Because many predictors of interest are in a nesting relationship with speaker or word, mixed models give more accurate quantitative estimates of their effect sizes, and especially of their statistical significance. The problems with fixed-effects models are only exacerbated by the token imbalances that exist across speakers and words in naturalistic speech, while mixed-effects models handle these imbalances well. This article demonstrates these and other advantages of mixed models, using data on /t, d/-deletion taken from the Buckeye Corpus as well as other real and simulated data sets.

Introduction

In recent years, it has become popular to analyze sociolinguistic data with mixed-effects regression

models (Jaeger and Staum 2005, Johnson 2009, Gorman 2010, Drager and Hay 2012, Tagliamonte and Baayen 2012). However, the traditional fixed-effects variable rule (VARBRUL) model – implemented using the software called GoldVarb (Sankoff et al. 2012) – is still being used (Bowie 2012, McQuaid 2012, Pereira Scherre 2012, and). Though the tide may have turned against fixed-effects models, the fundamental reasons to prefer mixed models are not all widely understood. This is because the literature on mixed-effects modeling has tended to discuss data from quite different disciplines, and it has often relied on very technical statistical arguments (Tagliamonte 2011 is an exception). In addition, it has occasionally been suggested (Paolillo 2013) that fixed-effects models fit in GoldVarb can achieve similar results to mixed-effects models.

This article provides a clear, data-based explanation of why mixed-effects models give better results than fixed-effects models. After a theoretical overview, the article goes on to analyze portions of Becker's (2009) data from the Lower East Side of Manhattan, and data from the Buckeye Corpus of Columbus, Ohio, speech (Pitt et al. 2007). These real examples are then crucially supplemented by simulated data sets, where the underlying population parameters are known. Through these examples, the article shows how common configurations of data lead to divergent analyses, and why the mixed-effects approach is almost always more

accurate. The final section considers some other benefits of the approach.

Whether it is a set of sociolinguistic interviews, a collection of historical texts, or a corpus of newspaper articles or Twitter posts, natural language data has a structure that calls for mixed-effects modeling, because of three related issues: grouping, nesting, and imbalance. The several thousand tokens (also known as observations or data points) in a typical naturalistic sociolinguistic data set do not all come from the same speaker, nor does each token come from a separate speaker. Instead, there might be a hundred or so tokens drawn from ten or twenty speakers. A similar grouping structure exists at the level of the word, where there might be several hundred different word types. An infrequent word might be represented by only one or two tokens, but the most common words will occur hundreds or thousands of times. This by-word imbalance will always occur in natural language; imbalance by speaker is also very likely, unless data were thrown away to ensure an equal number of tokens per person.

By itself, such data imbalance would not pose a problem for a fixed-effects regression analysis, if a researcher were asking questions about individual speakers or words. However, linguists are usually more interested in groups of speakers or words: they might want to compare men's and women's pronunciation of /s/, or see how stressed /æ/ is realized differently depending on the following consonant. These group-

level variables have a nesting relationship with the individual-level variables: speaker nests in gender (each speaker is either male or female, at least for the purposes of exposition here), and word nests in following consonant (each word has one particular consonant following /æ/).

The issue of nesting – the fact that the categories of interest (e.g. female speakers) are represented in the data by several individuals, each of whom contributes several tokens – would be benign if there were no individual-level variation. If this were true, there would be no more of a correlation among tokens from different speakers than among tokens from the same speaker. If this were true, the results of fixed-effects regression models would be statistically valid. But the assumption of no individual differences is almost surely false.

It has long been known (Gauchat 1905, Guy 1980) that individual speakers can vary in their overall rate or level of use of linguistic variables, over and above any differences attributable to the social groups to which they belong. On the other hand, it has been argued – or simply assumed – that individuals in the same speech community share linguistic constraints on variable processes (i.e. the following consonant affects /æ/ for all speakers in the same way).

As far as word-level grouping is concerned, variationist sociolinguistics in the Labovian tradition has been more open to the idea of the occasional lexical exception than to the notion of wholesale by-word variation, although the latter is a key prediction of usage-based theories (Pierrehumbert 2001). And some recent work (Baayen and Milin 2010) has shown that social factors can affect individual words differently (e.g. the difference in /s/-pronunciation between men and women might be different, depending on the word).

Fixed-effects models like VARBRUL/GoldVarb, in use since the early 1970s, assume the non-existence – and furthermore impede the discovery – of all four of these types of variation: rate/level by speaker, linguistic constraints by speaker, rate/level by word, and social constraints by word. It now seems likely that all four types of variability do exist. At the very least, the first type was known to exist, but recall that when fixed-effects regression models were first introduced, they were the best statistical tools available. The results they have delivered, while perhaps suboptimal from a modern perspective, are hardly invalidated by the "illegal" pooling together of tokens from disparate speakers.

Today, however, mixed-effects regression models provide a much better alternative for analyzing natural language data. Using random intercepts – the focus of this article – mixed models can accommodate potential rate/level variation, and using random slopes, they can also accommodate constraint variation. But they do not assume that such variation exists. Including random effects makes for far more accurate estimates of the significance and size of the nesting effects, but they also allow the linguist to measure the variation – or lack

thereof – among the nested units (speakers, words, texts, sentences, etc). Thus the methodology of mixed-effects models not only improves the performance of established research designs, but also opens up new research areas of considerable theoretical interest.

Mixed Models for Language Data

Linguists studying natural language typically make many observations of any given linguistic variable, whether phonological, syntactic, or another type.1 They also observe elements of the context in which the variable occurs – not only the linguistic context, but the entire speech setting, including attributes of the speaker. It is then possible to estimate the size and significance of the effects of these contextual elements, known as predictors. For example, one could explore how post-vocalic /r/ is realized differently by men and women (a so-called social or external factor, of which gender, age, socioeconomic class, and ethnicity are typical examples) or how a word-final /t, d/ is affected differently depending on the preceding context (a so-called linguistic or internal factor, which can be a phonological property of the environment, a grammatical property like the morphological/syntactic status of the word, or a lexical property like word frequency).

The "principle of multiple causes" (Bayley 2002) means that the variation observed for any linguistic variable has many sources. Some variation arises because

speakers behave differently from one another. This could be idiosyncratic in origin (where "idiosyncratic" may include effects of unknown causes), or relate to demographic categories, a subset of social/external factors that will be called between-speaker predictors here. Between-speaker variation applies to most, if not all, linguistic variables. Variation can also arise because individual words behave differently. Again, this could be attributed to lexical idiosyncrasy, or else to between-word predictors: phonological and other properties of the words in which the variable occurs.

There are other predictors which are neither between-speaker nor between-word; that is, they can vary even within a given speaker-word combination. These include the effects of adjacent words (usually considered linguistic/internal), and the effects of speech style and speech rate (both usually considered social/external). And even taking these into account, in exactly the same environment, a speaker does not pronounce a word the same way every time: some variation always remains at the level of the token.

Multiple regression is a statistical method that quantifies the simultaneous effects of several contextual predictors on a response. When the response is a measurement on a continuous scale (e.g. of vowel formant frequencies), this is called linear regression.

Linear regression performs perfectly only when several assumptions are met; these include linearity (a given change in a predictor should affect the response a

given amount), independence (the model's errors – the deviations of the observations from the model's predictions – should be independent), homoskedasticity (the variance of the errors should not depend on the values of the predictors), and normality (the errors should be normally distributed). We also assume that no important predictors are omitted (nor any unimportant ones included), that the predictors are measured accurately, and that none of the predictors are collinear (perfectly correlated) with each other.

In practice, regression can never manage to include every relevant predictor, nor will the predictors it does include ever be perfectly uncorrelated, but the results of the technique will be more or less valid if these and the other assumptions are not grossly violated.

With a binary response – the result of a choice (if not always a conscious one) between two alternatives – we use logistic regression. This models the natural logarithm of the odds of the response – the log-odds $\ln(p/(1-p))$, where p is the probability of the response – as a linear function of the predictors. Logistic regression models do not have errors of the same type as linear regression models, so most of the above assumptions do not apply. Instead of a direct linearity assumption, we assume that the log-odds of the response is affected linearly by the predictors. We still assume that the observations, conditional on the predictors, are independent, that there are no important omitted predictors, and that none of the predictors are perfectly collinear.

(A log-odds of zero corresponds to an odds of 1, or 1:1, meaning a 50% chance of an outcome occurring. An increase of 1 in the log-odds brings the probability to 73.1%, and another increase of 1 brings it to 88.1%. Note that equal changes in log-odds do not always correspond to equal changes on the probability scale.)

Though logistic regression has its roots in the 19th century (Cramer 2002), it was developed further in the mid 20th century (Cox 1958) and came to be widely used in the 1970s; VARBRUL 2, the second major version of the variable rule program for sociolinguists, was written in 1975 (Rousseau and Sankoff 1978).

Thirty-five years later, many sociolinguists still use a version of VARBRUL, called GoldVarb. It is limited to logistic regression with categorical predictors, not allowing for continuous dependent or independent variables. Nor does it easily allow for interactions among predictors, among other disadvantages (Johnson 2009). However, GoldVarb does have a flexible method of recoding predictors – and the ability to "slash" or omit some tokens in the estimation of some coefficients.

A flaw in the usual method of analysis using VARBRUL/GoldVarb is that correlations among tokens can lead to a violation of the independence assumption. This assumption says that in a regression, each observation should deviate from the model's prediction randomly and independently. But if tokens are correlated according to individual speaker and/or word, then this assumption cannot be met, unless speaker- and word-level variation

are modeled explicitly, something that users of VARBRUL can do only with difficulty and in a limited range of circumstances.

Of course, from the early years of variationist sociolinguistics, data from multiple speakers and words has been pooled together for analysis. While obscuring individual differences, this method revealed intricate patterns according to higher-level predictors, such as social class (the working class uses less post-vocalic /r/ than the middle class; Labov 1966) and word stress (stressed syllables retain more post-vocalic /r/ than unstressed syllables; Wolfram 1969). But perhaps appreciating the statistical issues involved, these studies did not try to assess the statistical significance of the differences.

To illustrate the subtle but substantial problems arising from pooling, we will examine a corpus of /t, d/-deletion that shows substantial grouping by speaker and word. The corpus was extracted by Josef Fruehwald from the Buckeye Corpus (Pitt et al. 2007), which consists of phonetically transcribed recordings of casual speech from 40 white speakers from the Columbus, Ohio area: 20 older (10 male and 10 female), 20 younger (10 male and 10 female).

In this corpus, the 13,664 tokens of word-final /t/ and /d/ are moderately unbalanced across speaker, ranging from 135 to 519 tokens per person. If we accounted for all the relevant between-speaker predictors – gender, age, social class, etc. – we might

find that speakers did not individually favor or disfavor deletion, and furthermore, that all speakers had the same constraints (predictor effects) on deletion. If not, though, then there would be correlation among each speaker's tokens, violating the independence assumption of a regression model – unless a predictor for individual speaker were included.

There are 905 distinct words in the corpus. As this is naturalistic speech, the data is highly unbalanced across word, with almost half the word types occurring only once while several word types occur more than 1000 times. After taking into account all the between-word predictors we could think of – including lexical frequency, as recommended since Hooper (1976) and in exemplar-theoretic work like Pierrehumbert (2001) – would all word types then behave alike? Perhaps they would, but it seems rash to assume they do without even checking.

The predictors in ordinary regression are called fixed effects, and fixed effects for nested predictors cannot be properly estimated at the same time. Predictors are nested when the value of one is completely predictable from the value of the other. For example, the predictor of speaker is nested in the between-speaker predictor of gender, because any token from "Mary Jones" also comes from the larger "female" grouping.

Regardless of the real magnitude of the gender effect, an ordinary regression model – a fixed-effects model – could fit the data equally well using a gender parameter of any size: the individual-speaker coefficients would

simply shift up and down to compensate for any change in the gender coefficient. Even if speaker identity and a nesting "social factor" like gender actually both influenced the response, a fixed-effects regression's results would be misleadingly arbitrary – even "meaningless" (Guy 1988:128) – because of the two predictors' maximal non-orthogonality.

For example, imagine we measured the voice pitch of three men and three women and obtained mean values of 100 Hz for man A, 120 Hz for man B, 140 Hz for man C, 180 Hz for woman D, 200 Hz for woman E, and 220 Hz for woman F. We might reasonably say that estimated male pitch is 120 Hz, estimated female pitch is 200 Hz, and each gender's speakers diverge from the norm by -20, 0, and +20 Hz. This intuitive solution, like the mixed-effects models below, minimizes the size of the speaker effects.

But a fixed-effects model has no way of privileging this solution above one where, for example, the estimated pitch for both genders is the same, 160 Hz, and the speakers deviate from the norm by -60, -40, -20, +20, +40, and +60 Hz. In fixed-effects regression, nested predictors and nesting predictors compete on an equal footing to account for the same variation. The relative contributions of individual speaker and a between-speaker variable like gender cannot be accurately determined.

Fixed-effects regression encounters the same problem if the nested predictor is the word, and the nesting predictor is a between-word variable. Another common case of nesting involves experimental stimuli or items, which can be nested in between-item predictors, depending on the design of the experiment.

The specific configuration of predictor variables, as derived from a given research design, determines whether nesting relationships exist. If there are no between-speaker predictors, then the speaker variable is not nested and may be modeled as a fixed effect. But if there are between-speaker predictors, speaker nests within them, and fixed-effects regression will be unable to model both levels simultaneously and accurately.

While the nesting problem may have been recognized early on (Rousseau and Sankoff 1978), along with the related issue of temporal correlation among tokens (Sankoff and Laberge 1978), it has received relatively little attention since (but see Van de Velde and van Hout 1998; Sigley 1997, 2003). It was recognized that in using VARBRUL/GoldVarb, one had to choose between including a speaker factor group and one or more between-speaker factor groups. Most researchers would select the latter option, but without recognizing the statistical ramifications. Similarly, researchers included between-word factor groups, necessarily forgoing any factor group for word itself.

For Guy (1980), addressing the relationship between the individual and the group, the only individual differences that matter are differences in constraint estimates and constraint orderings, which are seen to

stem either from insufficient data or dialect differences. Within a given dialect, constraints are thought to be quite uniform across individuals, assuming enough data has been collected to estimate them accurately. Whether to "lump together the data for several people" (20) is decided on the basis of whether they share constraints. Lumping together the data for individuals who differ only in their overall level or rate of a variable is implied to be benign.

As will be shown, there are actually several negative consequences to such lumping or pooling, a practice that may relate to the Labovian emphasis on the primacy of the speech community, with statements like "the community is conceptually and analytically prior to the individual" (Labov 2012: 266) or even "there are no individuals from the linguistic point of view" (Gordon 2006: 341).

Another reason that individual differences have largely been ignored is that they are thought to mainly concern the level, or rate, of variation: a topic often held to be less important than constraints on variation (Erker and Guy 2012: 546). But in order to properly study groups and constraints, we must attend to individual variation in rates. Not doing so can impair our calculation of the significance of group differences as well as our estimation of the magnitude of the constraints themselves.

In another early study of /t, d/-deletion, Neu (1980) analyzes the word and separately, stating that high-

frequency lexical items are more prone to deletion and noting that "[i]f these items are not considered separately, one is likely to conclude that the rule [of deletion] applies with a much higher frequency…or else, for example, that 'preceding /n/' has a much greater effect on deletion than it does" (53). In part, Neu is calling for a word frequency effect in the model, but the point about the interaction between a word effect (and) and a between-word effect (preceding /n/) foreshadows a point made below.

Guy (1991) includes the individual speaker in modeling /t, d/-deletion, but since no between-speaker predictors are considered at the same time, there is no nesting problem. Again, the alternative of lumping together data from multiple speakers who differ in their rates of deletion is viewed as unproblematic.

The statistical theory, and especially the computational means, to better address nesting have existed only recently. In the past, efforts have been made to limit data imbalance across words by discarding tokens from frequent lexical types (Wolfram 1993: 213-214), but this only addresses one of the problems posed by nesting. Some have directly recommended omitting the nested predictor of speaker – and implicitly that of word – from the final models (Guy 1988: 128, Tagliamonte 2006: 182), but, as noted above, this assumes that individual-speaker and individual-word variation do not exist.

VARBRUL practitioners have acknowledged that at least speaker variation does exist, even at times fitting separate models to individual speakers' data (e.g. Guy 1980, 1991), but they have not tended to recognize that by pooling their data, they make a "dangerous aggregation" (Van de Velde and van Hout 1998; see also Gorman 2009). But by including predictors for speaker and word, a properly-specified mixed-effects model – or mixed model for short – is valid whether by-speaker and by-word variation exist or not.

This is possible because while an ordinary regression model has only fixed effects, mixed models have random effects as well. There are several differences between the two types of effect. One distinction is that the fixed effect levels (e.g. male, female) are inherently limited and would likely recur in any extension or replication of a study, while the random effect levels (e.g. Stacy, Rick) might well not. In theory, the random effect levels have been sampled randomly from a larger population, but any units chosen to represent a larger set can work as random effects – especially when we are more interested in accounting for the units' variability than in the units themselves.

It is not always obvious whether to treat some predictors as fixed or random, nor does it always matter much to the results. However, when there is nesting, the nested predictor (e.g. speaker) must be random, while the nesting predictor (e.g. gender) should be fixed, unless it is nested in another predictor. The model-fitting software penalizes the size of the random effects, allowing a principled partition of variance between the levels (see Pinheiro & Bates 2000 for more details).

Although the discussion here often simplifies matters by discussing one fixed effect at a time, real mixed-model analyses will contain several fixed effects (and often their interactions). As in any regression, all relevant predictors must be included. Note that several fixed effects (e.g. gender, class, age) can share one random effect (e.g. speaker).

The statistical theory behind mixed models is not particularly new, but the computational techniques for fitting such models developed rapidly in the 1980's and 1990's. Pinheiro and Bates (2000) achieved a comprehensive implementation of mixed models in the R statistical software environment (R Core Team 2012). A further advance occurred with the 2003 introduction of the package lme4 (Bates et al. 2012). Its modeling function glmer() can handle large data sets, and fit models with crossed random effects, enabling the linguist to consider both speaker and word variation at the same time. This is the function "under the hood" of Rbrul (Johnson 2009), a menu-based front end interface that facilitates mixed-effects modeling (as well as fixed-effects modeling) in R.

The simplest type of random effect is a random intercept. For example, if we have a continuous response, the intercept for each speaker would be an estimate of their deviation from the prediction made for their group

(e.g. old working-class males). If the response is binary, the intercept (measured in log-odds) represents how much an individual favors one or the other outcome, again compared to the group prediction.

Taken together, the speaker intercepts are assumed to follow a normal distribution. The standard deviation or spread of this distribution is the main random effect parameter. The estimated variance of a speaker random intercept can be large or small, or even zero, meaning the speakers in the sample diverge no more than would be expected by chance.

A more complex type of random effect is the random slope, which allows speakers (or words) to differ with respect to their fixed effect constraints. For example, speakers might not only vary in favoring or disfavoring post-vocalic /r/ overall, but also vary in the way they shift their use of /r/ across styles: casual speech, careful speech, reading passage and word list. The first type of variation would be captured with a random intercept, the second type with a random slope. And if the data reflects that all speakers do in fact style-shift in a similar way, then the random slope term would be small, even zero.

When we want to know if any term in a model is significantly different from zero, we can perform hypothesis testing, where we compare two nested models. These models are identical except one includes a predictor that the other does not. This is the predictor whose effect we are testing.

We can compare models that differ in their fixed or random effects, usually to test whether more complex models are justified, and thus whether predictors are significant. In such hypothesis testing, different statistical issues arise depending on whether the model is linear or logistic, and whether we are testing the significance of 1) a fixed effect in an ordinary fixed-effects model, 2) a fixed effect in a mixed model, or 3) a random effect. The following recommendations summarize the usage currently accepted by the R-sig-ME mailing list (FAQ at http://glmm.wikidot.com/faq), although statistical recommendations and software implementations are always evolving.

1) Performing fixed-effects linear regression in R, we would fit the two models with lm() and compare them with an F-test using the (confusingly named) anova() function. For fixed-effects logistic regression, we fit the models with glm() and a perform a likelihood-ratio test with anova(), which is effectively the same thing VARBRUL does.

2) To test a fixed-effect term in a linear mixed model, the Markov chain Monte Carlo (MCMC) method, often implemented by mcmcsamp() or pvals.fnc(), may be preferred over the likelihood-ratio chi-squared test (Baayen et al. 2008; but see Barr et al. 2013). For fixed-effect terms in logistic mixed models, likelihood-ratio tests are considered more acceptable, though they may still be anti-conservative (p-values too high) unless the number of observations (tokens) and the number of

random effect levels (speakers/words) are both large (Bolker et al. 2009). Bootstrapping (Efron 1979) and simulation methods (Jones et al. 2009) are another way to obtain significance estimates, in all cases.

3) When we test a random-effect term, we are testing whether a variance parameter (e.g. the amount that speakers vary) is significantly different from zero. Since the variance cannot be negative, we have to make an adjustment to the likelihood-ratio test, which in the simplest case – testing a random intercept – means dividing the p-value in half (Stram and Lee 1994). The RLRsim package (Scheipl et al. 2008) provides a more general way of testing random effects. Some (e.g. Barr et al. 2013) argue against testing (let alone removing) random effects that reflect a study's design.

If we hold the random effects constant, adding significant fixed effects will generally cause the estimates of individual-speaker and individual-word variation to decrease. Decreasing this variation toward zero may be an attractive goal, but assuming it to be zero from the start – as fixed-effects analyses have unwittingly done – is not logical.

Speakers and words are the most obvious grouping factors in naturalistic linguistic data, and crossed random intercepts for these two factors are generally appropriate, even though fitting such models may require a larger amount of data to be collected.

Whether to use random slopes depends on the fixed-effect predictors involved. For instance, speech style might well have a different effect for different speakers, and plausibly for different words too, while a (phonetically-grounded) following-context effect would seem less likely to affect individual speakers or words differently.

If there is any reason to suspect that individual words or speakers might vary in their average realization of a continuous response variable – or in their rate of use of a binary response – then a random intercept capturing that variation should be included in the model. And if we suspect that speakers or words might vary in their response to a predictor, a corresponding random slope (or slopes) should be included as well (Schielzeth & Forstmeier 2009; Barr et al. 2013), although in practice such "maximal" models can be difficult and/or slow to fit.

The tradition of modeling variation in sociolinguistics has usually proceeded quite differently. While the literature has acknowledged that individual speakers from the same speech community (and demographic group) can vary in terms of rates or input probabilities (intercepts), it has often been claimed that speakers in a community do not vary in their constraints (slopes) (Guy 1991: 5). Both types of variation have been omitted from fixed-effects VARBRUL models. As for by-word variation, it has rarely been considered for rates or constraints. In all these cases, the omissions have substantial consequences.

For the sake of simplicity in exposition, the next sections largely set aside the potential benefits of

random slopes, concentrating on the clear benefits of random intercepts. This is not to be interpreted as saying random slopes are never needed. Indeed, the section dealing with the Buckeye Corpus does include a brief assessment of the use of random slopes.

Fixed-Effects Models Give Worse Results Than Mixed-Effects Models

This section will illustrate four ways in which applying ordinary fixed-effects models to grouped data can cause error. Only individual-speaker grouping will be considered; however, similar pitfalls would apply if we ignore individual-word variation, or any other correlation among observations in a data set. So when the term "speaker" is used from now on, the reader may also wish to imagine "word", "item", or some other repeated unit.

Fixed-effects models overestimate the significance of between-speaker predictors

Perhaps the most important danger of not using mixed models involves the significance of between-speaker predictors. If individual speakers differ greatly, then even randomly-chosen sub-groups can differ substantially, just by chance. So can men and women, old and young speakers, or any other division – again, just by chance.

Ignoring individual-speaker variation "may inflate the significance of statistical tests" (Sigley 2003: 228), leading to a high rate of Type I error, meaning that a

chance effect in the sample is mistaken for a real difference in the population. Mixed models keep the Type I error rate near where it should be (.05 is the usual alpha, or proportion tolerated).

At the same time, there is an unavoidable tradeoff, in that mixed models are more prone to Type II error, where a real population difference does not show up clearly or consistently enough in the sample to be recognized as statistically significant. If individual-speaker variation is at a high level, we cannot hope to discern small population differences without observing a large number of speakers; the smaller the group difference, the more individuals are needed (Johnson 2009).

We start by observing a single predictor, gender, in the Buckeye /t, d/-deletion corpus, where there are 20 male and 20 female speakers. Of course, the results of such a simple analysis will not be as accurate as if we had included other relevant predictors, such as age. But given the various problems that arise with even the simplest fixed-effects models, we can imagine that the problems would be compounded in a more complex analysis, and be more difficult to understand. Working with a single predictor, at the cost of some realism, we can more easily see the improvements offered by mixed models.

The response variable is binary, reflecting tokens of final /t, d/ – preceded by other consonants – that are either deleted, or retained as plain or glottalized stops.

The male speakers deleted the /t, d/ in 3805 of 6962 tokens (54.7 percent), while the female speakers deleted

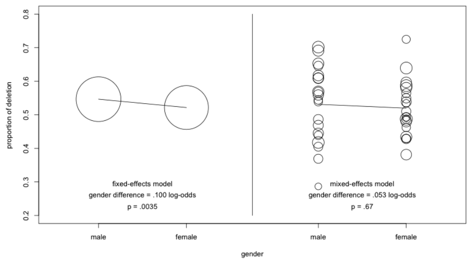
it in 3496 of 6702 tokens (52.2 percent). Ordinary logistic regression returns a coefficient telling us that the male speakers favor deletion by 0.100 log-odds (this follows directly from the raw percentages, although the same difference in percentages does not always correspond to 0.100 log-odds). This quantity is the unstandardized effect size of gender.

(Note: there are also several standardized measures of effect size that make it easier to compare between predictors and between studies: for example, Cohen's d, Hedges' g, and Glass's delta. However, in this article the term effect size is used to simply mean a regression coefficient or the difference between coefficients, that is, the magnitude of an predictor's effect.)

If we perform a likelihood-ratio test, comparing the model with gender to a null model with no predictors, we get a p-value of 0.0035. This implies that it is very unlikely that the observed gender difference is due to chance. That is, according to a fixed-effects model like VARBRUL, gender is a significant predictor of deletion.

The left panel of Figure 1 reinforces this impression. It shows one circle for the male speakers and another, noticeably lower down, for the female speakers. (The area of each circle is proportional to the number of tokens it represents.)

Figure 1: Deletion By Gender In The Buckeye Corpus. Left: Fixed-Effects Model (Pooled Data). Right: Mixed-Effects Model (Data Separated By Speaker).



In the right panel, however, we see the same data broken down by individual. This reveals that both male and female speakers have a wide range of deletion rates, and that the two ranges almost completely overlap. Any gender difference now appears to be quite contingent on the particular speakers in the sample. If a few speakers had been missing, for example, we might not have seen any effect.

We can formalize this by assessing the significance of gender with a mixed-effects model. When we use a random intercept for speaker, the likelihood-ratio test returns a p-value of 0.67, nowhere near the usual 0.05 threshold for statistical significance. The mixed model says that while speakers vary, there is little evidence for a gender difference. While /t, d/-deletion is a stable non-

standard feature, thus one we might expect to be used more by men (Labov 2001: 266), our revised conclusion of no gender difference accords better with the actual patterning of the speakers on Figure 1.

Fixed-effects models inaccurately estimate the effect sizes of between-speaker predictors, when some speakers contribute more data than others

In estimating a difference between two groups of speakers, we should ideally treat each individual about equally (“averaging by speaker”), assuming we have enough data to accurately evaluate the response for each speaker.

Fixed-effects regression distorts group differences by ignoring data imbalance and treating each token equally (“averaging by token”), thereby potentially counting some speakers much more than others. We return to Figure 1 to illustrate this distortion.

The left panel of Figure 1 ignores the fact that different speakers contributed different numbers of tokens. We have an average deletion rate of 54.7 percent (3805/6962) for the data from male speakers, compared with 52.2 percent (3496/6702) for the data from female speakers. The gender effect size is 0.100 log-odds, as noted above.

But if we count speakers equally and simply average their deletion percentages, the gender difference comes

out less than half as large: 53.1 percent for the males vs. 52.0 percent for the females, an effect size of 0.040 log-odds. This happens because the males with higher deletion rates contributed more tokens (a mean of 393 tokens each for the 10 highest-deleting males), and the males with lower deletion rates had fewer tokens (a mean of 303 tokens each for the 10 lowest-deleting males). Whether these differences are due to chance or some relationship between volubility and style, they have the effect of skewing the males’ estimate higher in the fixed-effects model.

A mixed model with a random speaker intercept treats speakers mostly equally; therefore it also returns a much smaller gender difference than the fixed-effects model. The mixed model effect size is 0.053 log-odds.

The inaccuracy of fixed-effects models, faced with token imbalance, is a general problem, but its direction can vary; here, the effect size of gender was overestimated, but with other data, the size of a between-speaker effect could be underestimated.

Another example of effect size misestimation can be seen in the data on which Becker (2009) was based. This comprises 3000 tokens of postvocalic /r/ from seven New York City speakers. The data from the five females has 654 /r/’s out of 1842, or 35.5 percent. The data from the two males has 476/1158 /r/, or 41.1 percent. Working with the pooled data, a fixed effects model estimates the gender effect at 0.24 log-odds.

But this does not take into account that the woman with the lowest rate of postvocalic /r/ (19.9 percent) provided the most data (492 tokens), while one of the women with the highest rates of /r/ (51.6 percent) produced the least amount of data (248 tokens). When the data is pooled, these two women both cause the /r/ rate for females to be underestimated, in turn exaggerating the difference between women and men. By contrast, a mixed model with speaker as a random effect treats speakers more equally, yielding a smaller gender effect of 0.20 log-odds.

Balanced data, with equal numbers of tokens per group, may arise in certain experimental contexts, but sociolinguists’ use of natural speech virtually ensures that balance will be rare in our data sets. We can limit imbalance artificially, by placing a ceiling on the tokens from a given speaker or of a given word, but this approach throws away valuable data arbitrarily, introducing its own problems. One reason mixed models are preferable is because they handle groups in a balanced way, whether or not there is balance at the level of the token.

Fixed-effects models inaccurately estimate the effect sizes of within-speaker predictors, when speakers do not share the same balance of data

The discussion so far has revolved around the consequences of ignoring individual-speaker variation as it relates to between-speaker predictors. But within-speaker predictors – those that are not constant in a given speaker's data – can also be misestimated by failing to take speaker variation into account. This is clearly true if the predictors' effects vary from speaker to speaker – a situation that calls for random slopes – but it can also happen when the variability applies only to speakers' intercepts.

The issue involves another type of data imbalance. Looking at speech style, for example, we might have cause for concern if different speakers were represented by different amounts of data in different styles. For example, suppose we were interested in the pronunciation of a vowel across three speech styles, and the number of tokens in the reading passage and word list were constant across speakers (by design), but the amount of spontaneous speech elicited from each person was (naturally) somewhat different. Such a data set for a speaker is the typical result of a Labovian sociolinguistic interview.

In this example, we are measuring the height of the vowel /ae/ by means of the first formant. Formants are

acoustic resonances in the vocal tract that are characteristic of vowel quality. The first formant, or F1, corresponds inversely to a vowel's height, so high vowels like [i] have lower F1 values than low vowels like [a]. We might measure F1 for the /ae/ vowel in the Northern (U.S.) Cities – e.g. words like trap and bath in Chicago or Detroit, where raising of the /ae/ vowel is a change in progress. Lower F1 values for /ae/ represent more advanced participation in the Northern Cities Shift.

Imagine that some speakers, who happen to have a low F1 (in all styles), also happen to produce more spontaneous speech. If we pool the data, the group estimate for F1 in spontaneous speech will be biased downward. The combination of speaker variability and token imbalance will end up being mistaken for an effect of style.

Using a simulation, we can illustrate this point while ensuring that speakers have the same constraints: speech style affects each speaker in the same way. Unlike real data, the population parameters of simulated data are known, so when we fit both fixed-effects and mixed-effects models to the same data, we can directly observe which estimate is more accurate. Using the R software and the parameters described below, we will run the simulation 1000 times (1000 runs). Each time, we randomly generate the data sets, fit a fixed-effects and a mixed-effects model to the same data, and compare the results. (Note that the parameters of the simulation are

for the purposes of illustration, rather than trying to represent a plausible style effect on the F1 of /ae/.)

In each data set, there are 10 speakers, whose intercepts differ: their average F1 values are normally distributed with a mean of 500 Hz and a standard deviation of 100 Hz. All speakers produce a balanced 50 tokens in word list style and 50 tokens in reading passage style. But for spontaneous speech, there is an imbalance: two speakers produce 25 tokens, six produce 50 tokens, and two produce 75 tokens.

Between styles, all speakers differ in the same way: compared to their reading passage tokens, every speaker's word list tokens average 50 Hz higher in F1, and their spontaneous speech tokens average 50 Hz lower. Within each style, each speaker's productions vary randomly with a standard deviation of 50 Hz.

In the two styles where the data is balanced across speakers, the fixed-effects and mixed-effects coefficients are unbiased and always nearly identical: close to 0 Hz for reading passage, and +50 Hz for word list. For the imbalanced, spontaneous speech style, both models are unbiased, with a mean effect near -50 Hz, but while the mixed model estimate is usually quite close to that figure, the fixed-effects estimate varies widely.

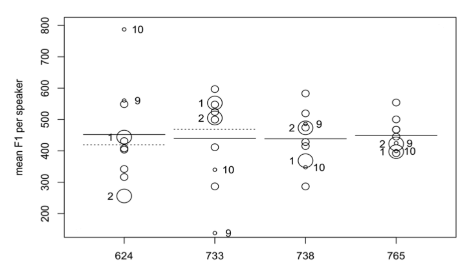
In 821 of the 1000 runs (that is, a large majority), the mixed-effects estimate of the effect of spontaneous speech was closer than the fixed-effects estimate to the underlying parameter of -50 Hz. The median difference between the models was 5.8 Hz. In the other 179 runs,

where the fixed-effect estimate was closer to -50 Hz, the median difference between models was only 1.7 Hz.

The fixed-effects estimate is least accurate when the speakers with more tokens of spontaneous speech have much higher or lower F1 means than those with fewer tokens of spontaneous speech. If the two groups have similar means, there is little difference between models.

Figure 2 shows how token imbalances affect four selected runs. In run 624, the low position of one large circle (speaker with 75 spontaneous tokens) and high position of both small circles (speakers with 25 spontaneous tokens) make the fixed-effects estimate for spontaneous speech too low: -83 Hz. In run 733, the opposite configuration – large circles high, small circles low – makes the fixed-effects estimate too high: -17 Hz.

Figure 2: Four Example Runs From The Simulation (In Each Run, Speakers 1-2 Produce 75 tokens, Speakers 3-8 Produce 50 Tokens, And Speakers 9-10 Produce 25 Tokens)



In run 738, none of the large or small circles have extreme means, so the fixed-effects estimate comes out exactly at -50 Hz. And in run 765, the large and small circles are all on the low side, cancelling each other out; the estimate is -49 Hz. But no matter the pattern of data imbalance, the mixed model adjusts to it, giving a coefficient near -50 Hz.

Whenever we are interested in a within-speaker variable, and the distribution of that variable is different for different speakers, then unless individual-speaker variation is modeled explicitly (using a mixed model), we are at risk of an estimation error.

This problem is most serious when there is a true correlation, not merely a chance association, between speaker intercepts and the distribution of a predictor. This seems likely to occur with stylistic predictors. Speakers who produce more standard variants overall might well produce less spontaneous speech in an interview. A fixed-effects model will then overestimate the style effect. Due to the "missing data" from the more standard speakers, spontaneous speech will appear to be less standard than it really is.

We can illustrate this with the four older female speakers in Becker (2009). They each produced similar numbers of tokens in word list and reading styles (about 15 and 80, respectively) but varied in their production of spontaneous speech. Maggie produced 143 tokens, Ann 228, Lucille 298, and Mae 394. And the more spontaneous speech the women produced, the less they

used post-vocalic /r/ in all styles. Maggie had 52 percent /r/ overall, Ann had 24 percent, Lucille had 31 percent, and Mae had 20 percent.

The data imbalance, where Mae is overrepresented and Maggie is underrepresented in spontaneous speech, causes a fixed-effects model without a speaker term to estimate a lower rate of /r/, and a more negative estimate, for that style. The fixed-effects estimate is -0.32 log-odds for spontaneous speech, whereas a mixed model returns -0.25 log-odds.

This section has shown that if there is data imbalance across a within-speaker variable, as well as overall variation by speaker, the interaction of the two can lead a fixed-effects model (lacking a speaker effect) to misestimate the within-speaker effect. This is much less likely to happen with a mixed-effects model containing a speaker random effect.

Fixed-effects models underestimate the effect sizes of within-speaker predictors in logistic regression

With a binary linguistic variable, we cannot model the response probability p as a linear function of the predictors, at the risk of predicting probabilities outside the legitimate range of 0 to 1. Instead, we typically use logistic regression, which models the log-odds of the response probability – $\ln(p/(1-p))$ – as a linear function of the predictors. The log-odds ranges from $-\infty$ if the probability is 0, to $+\infty$ if the probability is 1.

If we graph the probability as a function of the log-odds, for example $x = \ln(y/(1-y))$, we get a characteristic S-shaped curve: the logistic function. Curves of this shape – representing processes that start slowly, speed up in the middle, then slow down as they approach completion – have been observed for changes in progress, especially in the field of historical syntax (Kroch 1989). In the study of diachronic change, then, logistic regression is fairly well motivated. Indeed, some simple mechanisms of competition between variants (or grammars) predict that rates of change should be proportional to $p(1-p)$, which ensures that a plot of p against time is a logistic curve (Denison 2003).

Logistic regression may not be as well motivated for modeling the synchronic constraints on binary variables. However, its use is all but universal. The following section illustrates a pitfall in applying fixed-effects logistic regression to grouped data.

Imagine that speaker A uses a linguistic variant 50% of the time in a "disfavoring", and 60% in a "favoring" context. This difference works out to 0.41 log-odds. Speaker B uses the variant 79% of the time in the "disfavoring" context, and 85% in the "favoring" context. Speaker B uses the variant more often overall, but the contextual difference is still 0.41 log-odds. Logistic regression will estimate the same effect for both speakers (the same slope, in regression terms), but their rates (or intercepts) will differ.

However, if we pool data from speakers A and B, we observe a contextual effect that is smaller than 0.41 log-odds. For example, if A and B contribute equal amounts of data, their combined "disfavoring" context will show an overall rate of 64.5 percent (the average of 50% and 79%), and their combined "favoring" context will show a rate of 72.5 percent (the average of 60% and 85%). And the difference between 64.5% and 72.5% is only 0.37 log-odds, 9 percent smaller than 0.41.

The greater the individual-speaker variation, the worse a mistake it is to pool the data before estimating a within-speaker logistic effect. Doing so averages speakers' individual rates of variation on the probability scale instead of the log-odds scale (Mood 2010).

Table 1 shows the average effect size from a repeated simulation of 50 speakers. Each speaker's data consists of 100 "disfavoring" and 100 "favoring" tokens, and the difference between them (the underlying effect size) is now 1 log-odds unit. Speakers' intercepts are normally distributed with a standard deviation of 0 (no speaker variation), 0.5, 1.0, 1.5, or 2.0 log-odds.

The table shows the average effect size, over 100 repetitions of the simulation, from a fixed-effects model with context as the only predictor, and from a mixed model that supplements the contextual fixed effect with a random intercept for speaker.

TABLE 1

The Effect of Pooling Binary Data Across Speakers With Different Intercepts

speaker intercept variation standard deviation (log-odds)	fixed-effects model mean effect size (log-odds) response ~ high.low	mixed-effects model mean effect size (log-odds) response ~ high.low + (1|speaker)
0	1.000	1.000
0.5	0.950	1.006
1	0.828	0.998
1.5	0.714	1.004
2	0.604	0.996

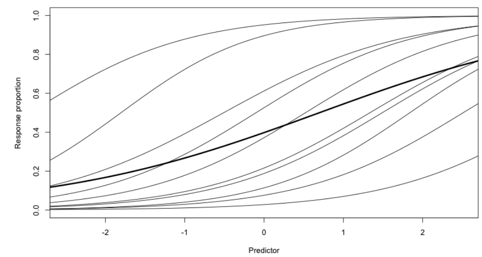
Each simulation has 50 speakers, each with 100 "low" tokens and 100 "high" tokens. Each speaker has a 1.0 log-odds difference between "low" and "high", but speakers vary in their intercept as shown in the left column. Results are the mean of 100 simulations.

As usual, the fixed-effects model is accurate only when speaker intercepts do not vary. As speaker variance increases, its accuracy declines, slowly at first: a speaker standard deviation of 0.5 gives an estimate that is only 5 percent too low. But a speaker standard deviation of 1 gives a result that is 17 percent too low, and a speaker standard deviation of 2 gives a result that is 40 percent too low. By contrast, the mixed model always estimates an effect size that is very close to the ideal value.

Figure 3 is a graphical representation of this same effect. We see ten logistic curves (light-colored lines); each represents one speaker. The curves have very different intercepts (they range from -3.18 to +3.30, with a standard deviation of 2), but they all have quite similar slopes (ranging from 0.89 to 1.19). If we fit a mixed-effects logistic model with a random effect for speaker, the model returns an overall slope of 1.00. This

is very close to the average of the ten individual slopes, which is 1.01. On the other hand, if we pool the data and fit a fixed-effects logistic model, one which ignores the fact that the speakers have different intercepts, then the slope comes out much lower, at 0.59 (dark line). Again we see that in logistic regression, pooling data and ignoring between-speaker intercept variation (or omitting any other relevant between-speaker predictor) will always lead to the underestimation of within-speaker effects.

Figure 3: The Effect Of Pooling Grouped Data In Logistic Regression.



Average slope of curves: 1.01. Mixed-effect estimate: 1.00. Fixed-effect estimate: 0.59.

The same point can be seen in Becker's (2009) data, where speakers used less post-vocalic /r/ when talking about the Lower East Side neighborhood than they did talking about other topics. If we look at each individual's data separately, we can isolate two speakers who show the same size topic effect: Mae and Lindsey each have a

topic effect of 0.41 log-odds. However, Mae produces only 20 percent post-vocalic /r/ overall (16 percent while talking about the LES, 23 percent otherwise), while Lindsey produces 60 percent /r/ overall (55 percent about the LES, 65 percent otherwise).

Part of this difference is due to age – Mae is older – but some of it is likely to be more individual, as both speakers are higher-class, college-educated females. If we model these two speakers with a mixed model, the combined topic effect remains 0.41 log-odds. But if we run a fixed-effects model without speaker, the topic effect falls to 0.21 log-odds.

A less dramatic illustration of the point comes from the seven-speaker /t, d/-deletion study of Guy (1991). With speaker taken into account in a mixed model, the morphological factor weights for deletion are .64 for monomorphemic words (lift), .55 for semi-weak past tense forms (left), and .32 for regular past forms (laughed). But when the speakers are pooled, the factor weights come out as .61, .57, and .33. This compressed span of weights is a consequence of ignoring individual differences in intercepts.

The above discussions are concerned with both statistical significance and effect size. In all cases, we see that failing to model individual speaker variation, when it exists, leads to quantitative error. Analogously, leaving the individual word unmodeled leads to error in the face of individual-word variation.

Many VARBRUL practitioners have indeed omitted these grouping factors from their models, but individual-speaker variation has not been totally ignored. Guy (1980) models each of his speakers independently at first – always a valid if potentially underpowered approach – but although he goes on to pool their data, his intent is not to examine between-speaker predictors, so at least one of the problems of speaker nesting is avoided. Guy (1991) also presents analyses for individuals as well as pooled data; the inherent problem with performing logistic regression on pooled data is revealed, as noted.

More recently, Paolillo (2002, 2013) and Sigley (2003, 2010) have developed elaborate ways to model predictor interactions within the framework of the GoldVarb software, creating what they claim to be hierarchical models. Aside from being extremely complicated to implement in Goldvarb, this elaboration of the fixed-effects approach does not appear to overcome the key problem of how to partition variation between nesting and nested predictors. Preliminary experiments suggest that GoldVarb, even following the method of Paolillo (2013), does not partition the variation in a consistent manner. The effects of the nesting predictors are consistently underestimated.

Actual mixed-effects models, on the other hand, are being adopted more and more widely in many fields of study. They make it very easy to model nested predictors like speaker and word, and thus they represent a clear advance over older techniques.

Fixed-Effects And Mixed-Effects Models Applied To /t, d/-deletion In The Buckeye Corpus

The parameters of simulations have to be manipulated to make desired points clearly. When we use real data sets to compare methodologies, the differences are not always as remarkable, and any given difference may have complex and multiple causes.

Returning to the /t, d/-deletion data from the Buckeye Corpus, this section compares the results of a VARBRUL-style analysis to one employing mixed models. The resulting differences in predictor significances are striking, while those regarding effect sizes are more subtle. Taken together, they recommend the mixed-model approach.

Six predictors will be examined: segment identity, preceding context, following context, morphological category, word frequency, and (as an example of a between-speaker predictor), gender. The coding and ordering of phonological factors is based on Smith et al. (2009). The six predictors are modeled as independent, non-interacting variables. (Erker and Guy's (2012: 545) suggestion that frequency "amplifies" other effects was tested, but not borne out at all in this data.)

Segment identity means whether the /t, d/ would be pronounced /t/ or /d/, if it were not deleted. Preceding context is divided into five categories: sibilant, stop, nasal, non-sibilant fricative, liquid (in decreasing order of

their usual deletion-favoring effect). Following context forms four groups: consonant, glide, vowel (in decreasing order), and pause (the position of pause is dialect-specific; Guy 1980).

Morphological category separates the regular past tense (e.g. laughed) from the irregular past tenses, a miscellaneous group (e.g. left, burnt, cost, held, sent, went). The other two morphological categories are monomorphemes (e.g. lift), and the suffix -n't.

Word frequency was derived from a separate corpus used only for that purpose: 22.8 million words of telephone speech, derived by Kyle Gorman from the Fisher (Godfrey et al. 1992) and Switchboard (Cieri et al. 2004) corpora. The metric used is the base-10 logarithm of the ratio of the frequency of each wordform to that of the median frequency word. Any word with the median frequency of 104 occurrences (like canned) thus receives a frequency score of 0. A word one-tenth as frequent (like institutionalized) receives a score of -1, a word 100 times as frequent (like friend) receives a score of +2, and so forth. The most frequent words are don't at +3.23 and just at +3.22; these two words alone make up 29 percent of the /t, d/-deletion corpus. Words with the minimum frequency score of -2.02 (like annexed, nudist, or whupped) occurred once in the telephone corpus.

Excluding 46 tokens of words missing from the telephone corpus entirely, and 17 tokens without a clear following segment, leaves us with 13,601 tokens of 881 word types.

Our mixed models employ random intercepts for word and speaker, because we have a between-speaker predictor (gender), and several between-word predictors (segment identity, preceding context, morphological category, frequency). Note that following context does not have a nesting relationship with either speaker or word, because each speaker's data contains examples of many different following contexts, and more importantly, each word appears with many different following contexts.

In the mixed model, the 40 estimated speaker intercepts are approximately normally distributed, meeting the assumption of the model. The 881 word intercepts form a leptokurtic distribution, with a pointier peak (and thicker tail) than normal, but this is due to the many words with only a few tokens, which, in logistic regression, are rarely assigned large random effects.

Whenever we do not include by-speaker random slopes (as is the case for most of the discussion here), we are assuming that while speakers may vary in their overall rates of deletion, each speaker is subject to the same constraints (regarding predictors that apply within each speaker's data, like following context).

Similarly, individual words may favor or disfavor deletion, but without by-word random slopes in the model, each word type is assumed to respond in the same way to predictors like gender.

At the end of each of the next two sections, we will briefly assess the effects of relaxing these assumptions and seeing the effect of introducing random slopes.

Differences in significance

Table 2 is a comparison of the significance estimates – p-values from likelihood-ratio tests – returned by fixed-effects and mixed-effects models, regarding the six predictors described above.

Some of the p-values are very small, and so they are all given in scientific notation: for example, 2.06×10^{-17} means .0000000000000000206. The exact size of p-values is meaningful, especially in a methodological comparison like this. Indeed, the idea that p-values must be reported and interpreted as either "significant or not" has been challenged, even by Fisher, the inventor of the p-value, himself (Gigerenzer et al. 2004).

TABLE 2

Significance Of Predictors In Fixed-Effects And Mixed-Effects Models

Predictor	significance (p-value) in fixed-effects model	significance (p-value) in mixed-effects model*
segment identity	2.06×10^{-17}	7.03×10^{-6}
preceding context	1.63×10^{-104}	1.70×10^{-29}
following context	3.70×10^{-107}	1.87×10^{-112}
morphological category	8.54×10^{-27}	7.25×10^{-11}
word frequency	1.50×10^{-70}	2.16×10^{-4}
speaker gender	3.71×10^{-7}	0.258

Both models fit to 13,601 tokens of /t, d/-deletion. *has random intercepts for speaker, word type

The fixed-effect p-values (left column) are all extremely low. Relying on these numbers, we would conclude that the three phonological predictors, as well as morphological category, word frequency, and gender, all influence the probability of /t, d/-deletion.

The p-values from a mixed model (right column) are higher in all cases but one, and usually vastly higher; the exception is following context. Without a nesting relationship with speaker or word, following context did not gain any spurious significance in the fixed-effects model. By contrast, the fixed-effects model overestimated the significance (underestimated the p-value) of the between-word predictors, like preceding context and word frequency, due to unmodeled word variability, while unmodeled speaker variability led to a similarly overstated significance level for the between-speaker predictor, gender.

The fixed-effects model estimated the p-value for gender as 3.71 x 10-7, but the addition of random effects – primarily the speaker random intercept – brought that figure up to 0.258. In other words, gender no longer appeared to be a highly significant of /t, d/-deletion, but rather one whose effect could easily have occurred by chance. If we also add a random by-word slope to consider the possibility that a gender effect could apply differently to different words, the p-value is similar: 0.184.

The idea here is to account for the consistency as well as the size of effects. Above, we saw how a random

speaker intercept helped us see the difference between a hypothetical group of men who all deleted slightly more than a group of women, and the actual situation in the Buckeye corpus: a large range of deletion for both genders with substantial overlap between them. Similarly, a random by-word slope could distinguish a conventionally significant gender effect, for example where men delete slightly more than women regardless of the word being spoken, from the actual situation, where the overall gender difference is not significant, but men do tend to delete more in 491 word types while women favor deletion in 390 others.

We can go on to test if this random slope is itself significant – it is – and identify examples of this lexical interaction. So the common word don't shows more than twice the usual gender effect: 81% deletion for men, 70% for women. Meanwhile, can't shows the reverse effect: 33% deletion for men vs. 40% for women. Mixed models cannot explain a surprising (and statistically significant) pattern like this, but they are indispensable for identifying them. We might have to return to the transcripts or audio to look for other predictors correlated with gender, in order to understand these differences.

On a more basic level, the mixed model reports that speakers vary with a standard deviation of 0.48 log-odds, while words have a standard deviation of 0.59. The model can also tell us which speakers (#19, #11, #13, #37) and words (kind, amount, front) most favor deletion,

and which speakers (#6, #25) and words (can't, saint) most disfavor it.

The infinitesimal fixed-effects p-value for word frequency implies that its relationship to deletion is unquestionable. However, the data does not support such a strong relationship. For example, if we consider old and told, where the preceding context is almost identical, and further constrain the following context to tokens before consonants, we find 61 percent deletion in told (44/72), but only 30 percent in old (20/66), even though old is three times as frequent as told in the telephone corpus.

Such word-level reversals by no means discredit the frequency effect, but taking them into account does lead to a more reasonable significance estimate. Unlike the fixed-effects p-value near 10-70, the mixed-effects p-value near .0002 says that there is a very small, but non-negligible chance that this sample could have come from a population having no real underlying frequency effect on /t, d/-deletion.

With a large data set such as this one, predictor effects that are real – and most of those found here have been detected in previous studies – will remain significant using a mixed-effects model. With a fixed-effects model the significance of many predictors will be exaggerated. This may not matter if we are considering a predictor that actually has a real effect. But fixed-effects regression may also claim "significance" for predictors that have no relationship to the response other than that due to random chance (Type I error).

Differences in effect size

Moving beyond significance levels – which are highly dependent on the size of a data set, as well as on the strength of the effects – this section will compare the estimated effect sizes between a fixed-effects and a mixed-effects model, each of which contain the five predictors that were confirmed as significant by the mixed-effects model above (that is, removing gender, notwithstanding the potential interaction with word type).

TABLE 3

Coefficients Of Predictors In Fixed-Effects And Mixed-Effects Models

predictor (factor group)	level (factor)	coefficient (factor weight) in fixed-effects model	coefficient (factor weight) in mixed-effects model*
segment identity			
	/d/	0.279 (.569)	0.274 (.568)
	/t/	-0.279 (.431)	-0.274 (.432)
preceding context			
	sibilant	0.754 (.680)	0.756 (.680)
	nasal	0.736 (.676)	0.725 (.674)
	stop	0.238 (.559)	0.164 (.541)
	fricative	-0.605 (.353)	-0.336 (.417)
	liquid	-1.123 (.245)	-1.309 (.213)
following context			
	consonant	0.515 (.626)	0.570 (.639)
	glide	0.188 (.547)	0.196 (.549)
	vowel	0.005 (.501)	-0.000 (.500)
	pause	-0.708 (.330)	-0.766 (.317)
morphological category			
	n't	0.272 (.568)	0.548 (.634)
	irregular	0.483 (.618)	0.325 (.581)
	monomorph.	0.007 (.502)	-0.044 (.489)
	regular	-0.762 (.318)	-0.829 (.304)
word frequency			
	+1 log-unit	0.383 (N/A)	0.187 (N/A)
intercept (input prob.)			
	@ median freq.	-1.213 (.229)	-1.074 (.255)

Both models fit to 13,601 tokens of /t, d/-deletion. *has random intercepts for speaker, word type

Table 3 presents these coefficients both in log-odds and as factor weights, except for the continuous predictor of word frequency. The coefficient for frequency represents the estimated change in the log-odds of deletion for a one-unit increase in the frequency score (that is, for a tenfold increase in word frequency).

Each predictor is affected differently by the change from a fixed-effects model to a mixed model with speaker and word intercepts. We will list the similarities and differences, and try to understand why the most important differences come about.

Among the between-word predictors, the models agree on the effect of segment identity: /d/ is slightly more likely to delete than /t/. For the effect of preceding context, the ordering of levels is close to Smith et al. (2009) – except nasals favor deletion here more than stops – but the estimates do change somewhat between the two models. The coefficients for a preceding stop (positive) or fricative (negative) move towards zero in the mixed model, while that for a liquid becomes more negative, disfavoring deletion.

For following context, the mixed model effects are all about 10 percent larger. This is likely caused by the phenomenon discussed above, where pooling data across grouping factors leads to underestimation of effect sizes in logistic regression. (The introduction of random slopes makes the average following-context effects larger still; the individual effects vary somewhat by speaker, and even more according to word.)

This is somewhat surprising in light of Guy et al. (2008), which observed, in a historical corpus from New Zealand, that words which occur more frequently in deletion-favoring environments (e.g. before consonants) show more deletion overall, even in disfavoring environments (e.g. before vowels). The same correlation appears, albeit weakly, in the Buckeye Corpus. The theory is that a word which occurs more often in the deleted form – at first simply due to the balance of environments it occurs in – will acquire a tendency of its own towards deletion. Conversely, if a word tends to occur in contexts that disfavor deletion, it will come to disfavor deletion itself, even in favoring contexts.

A random intercept for word could model this behavior, if it were really this simple. In fact, the correlation is noticeably weaker in the following-vowel environment than the following-consonant environment. Table 4 shows this with the 39 words that have at least 5 tokens before both consonants and vowels (excluding post-nasal tokens, which show an unusually high rate of deletion before vowels), correlating the proportion of following consonants in the context with the deletion rates before consonants and vowels.

TABLE 4

How Overall Following Context Affects Deletion Rates before Consonants and Vowels

percentage of tokens followed by consonants	mean word deletion rate before consonants	mean word deletion rate before vowels
72-90% (top 10 words)	67.4%	21.3%
40-71% (middle 19 words)	46.8%	20.6%
16-38% (bottom 10 words)	39.4%	16.8%
r (Pearson correlation)	.446	.156

39 words, each with at least 5 tokens before consonants and 5 tokens before vowels

Complicating the story even further, it seems that individual words can diverge greatly from the typical following-context effects. If words are individually sensitive to their own contexts, it is hard to see how an overall favoring or disfavoring tendency could develop just from the context (although other such tendencies do develop somehow).

In terms of the deletion-favoring effect of a following consonant compared to a following vowel, the word old is about average: 20/66 = 30% deletion before a consonant (as noted above), 6/44 = 14% deletion before a vowel. The word moved shows an increased sensitivity to the following context: 10/21 = 48% deletion before a consonant, 0/35 = 0% deletion before a vowel. And the word child diverges in the opposite direction: 4/32 = 12.5% deletion before a consonant, 10/30 = 33% deletion before a vowel.

Such findings raise questions about the causes and extent of lexical idiosyncrasy that would take further work to resolve. And we note that while the overall

random slope for following context is significant, the particular differences among moved, old, and child are based on fairly few tokens.

Morphological category is the only predictor where the order of the levels changes between the models. In the fixed-effects model, the irregular past tense category favors deletion most, while in the mixed model, n't favors deletion the most. The reason for the reversal is not entirely clear, but probably reflects the fact that a larger overall n't effect allows the mixed model to have smaller individual-word effects for the few common words in this category.

Both models agree that irregular pasts undergo deletion more than monomorphemes, an unexpected result that deserves further investigation. Regular past forms show the least tendency to delete, a typical finding which may support a functionalist "tendency for semantically relevant information to be retained in surface structure" (Kiparsky 1982:87) or a cycle-based lexical phonology account (Guy 1991) where monomorphemes are exposed to a deletion rule more than rule-generated regular past tense forms.

The largest difference between the two models concerns word frequency, where the mixed model estimate of +0.187 log-odds per tenfold increase in frequency is less than half the size of the fixed-effects estimate of +0.383. That is, more frequent words exhibit more deletion in both models, but in the mixed model this effect is less than half as large.

This change is brought about by the random intercepts assigned to each word, which allow the model to fit the data more closely, along with a weaker overall frequency effect. Words with very high or very low deletion rates can be treated as exceptional, without their behavior necessarily being linked to between-word predictors like frequency.

Mixed models offer a way to handle "outlier" words without throwing away their data. The three highest-frequency words – don't, just and kind – all show more deletion than is predicted from their frequencies and the other factors in the model. If we discarded these three words – one-third of the data! – the fixed-effect frequency slope would drop from 0.383 all the way to 0.100. The mixed model's estimate of 0.187 falls in between; it does not ignore exceptional words, nor does it ignore that their behavior is exceptional.

Also, recall that words with an unusually high or low number of tokens are treated on a fairly equal footing by the mixed model, so the idiosyncratic properties of the most frequent words do not bias our estimates – even our estimates of a frequency effect.

As with any continuous predictor, a careful treatment of word frequency would go on to explore whether some other relationship besides a straight line might fit the data better. But even on an initial pass, we can see that to understand the intricacies of this data set – e.g. that word frequency does favor deletion, but not as much as the

few most frequent words might suggest – the mixed-effects model is a useful, if not essential, tool.

We should also note that by using random effects, mixed models attempt to eliminate idiosyncrasies in a data set that might not apply to another set of data on the same variable, one drawn from different speakers and largely comprised of different words. Fixed-effects models incorporate these idiosyncrasies, making models less comparable.

The Importance Of Mixed Models

The long history of variable rule analysis, including the substantial bibliography on /t, d/-deletion, consists of researchers comparing and contrasting their results in a productive manner. So we know that fixed-effects models' effect sizes are not massively unreliable, nor have shrunken p-values consistently led to a fatal level of Type I error.

Nevertheless, having described several clear advantages of applying mixed-effects models to natural language data, this article recommends that we capture any effects of the individual speaker and/or individual word using crossed random intercepts, at the very minimum, and to consider using random slopes as well,

Our simulations and other analyses have shown how inaccurate our regression estimates can be if we ignore the real structure of our data and act as if each token

were independent and of equal value in determining the effects of the predictors.

The fairly large Buckeye Corpus of /t, d/-deletion showed that substantial differences in effect size, and very large differences in significance, can exist between fixed-effects and mixed models applied to the same data. Of course, the true parameters underlying the Buckeye data, like any real data set, are unknown, but insights taken from the simulations and the investigation of outlier words support the mixed model approach.

Given enough data to fit it, switching to a mixed-effects regression model will cut down on spurious effects, while real effects will usually remain significant. Mixed models also estimate effect sizes more accurately, in a way that abstracts from the idiosyncrasies of the sample at hand. Thus, they offer hope for superior quantitative analysis and are a better tool for comparison with – or replication of – other research.

In the terminology of statistics, mixed model results are generally more conservative. As one linguist puts it, "Using mixed models and adding individual speaker as a random effect results in interesting, logical results for my data. The results are conservative, but I like that. If I don't use speaker as random, I get loads of extra factors as significant, but lots of these make no sense and simply can't be explained. This again gives me confidence in my conservative approach" (Rob Drummond, p.c.).

The other side of this methodological coin is that using mixed models, analysts may need to examine

larger data sets, generally involving more speakers and/or more lexical types, in order for the effects of some predictors to be properly recognized as significant. But preferring Type II error over Type I error, like this, is standard scientific procedure.

Regardless of the specific purpose of our regression analysis, we do not want our models to tell us that irrelevant predictors are significant (which fixed-effects models often do). Our discussions and conclusions are also likely to be improved if we are able to work with the most accurate coefficient estimates possible (which mixed models can provide). In particular, research comparing linguistic varieties – where similar models are fit to different data sets – will benefit from the use of speaker random effects, which help distinguish community differences from purely individual ones. Indeed, revealing the extent of individual-speaker variation – and measuring and comparing it between communities – is itself a valuable insight to be gained from mixed modeling, especially as such variation has been largely overlooked in much VARBRUL practice.

If individual speakers' behavior, or its relationship to group norms, is the focus of investigation (e.g. Drager and Hay 2012), then mixed models are especially valuable. In this case, between-speaker factors (age, gender, etc.) serve as the variables to be controlled, in order to better reveal individual patterns and idiosyncrasies. This is the reverse of the approach employed above, where an improved description of the size and significance of social factors was a goal that was better reached by keeping individual differences under control. Whichever focus a researcher has, mixed models improve their vision.

There are other ways in which linguistic insights can be gained from the use of mixed models, beyond the statistical advantages that have been the focus of this article. As noted, we have fit random intercepts not so much for their own sake, but to obtain more accurate significances and effect sizes for the fixed effects of interest. However, Drager and Hay (2012) show how the random intercepts calculated in one model can be used as predictors in subsequent models, a procedure they call cascading models.

Fruehwald and MacKenzie (2011) propose that if community members show markedly different levels of inter-speaker variability ("cohesion", to use their term) with respect to two phenomena, then the phenomena should be considered grammatically distinct. On the other hand, if a community displays a similar degree of cohesion regarding two processes, the processes might be considered unitary in the grammar. Fruehwald and Mackenzie use this logic to argue that the additional /t, d/-deletion found in English semiweak past tense forms is more variable between speakers – and hence grammatically distinct from – the deletion that affects regular past tenses, even though they occur with a similar average probability. Conversely, the rare contraction of had (e.g. in they'd gone) and the common

contraction of has/have (e.g. in they've gone) may be governed by the same underlying process, because the community has similar cohesion factors (equivalent to speaker intercept standard deviations) with respect to both. While it is far from being proved, such a linguistic hypothesis could hardly have been formulated and tested without mixed models, which are ideally suited for evaluating and comparing inter-speaker variation.

While there exist many other valuable modern statistical methods for the analysis of linguistic data (see Tagliamonte and Baayen 2012), mixed-effects regression models are becoming an essential tool. As long as our data consists of repeated observations from more than one speaker, and of more than one word, the greater accuracy of mixed models with respect to both significance and effect size, as demonstrated in this article, should lead analysts to avoid fixed-effects modeling techniques such as VARBRUL/GoldVarb.

At the same time that they focus a sharper quantitative lens on familiar higher-order social and linguistic predictors, mixed models provide a new type of information about a lower level of variation: they show how speakers and words vary, both as a population and as individuals. The first advantage strengthens the study of variation as we have known it for decades; the second opens new doors for linguistic investigation and insight.

Thanks

The author thanks Ben Bolker, Katie Drager, Josef Fruehwald, Kyle Gorman, Florian Jaeger, John Paolillo, Sali Tagliamonte, and many anonymous reviewers.

References

Baayen, R. Harald, Douglas J. Davidson and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59: 390–412.

Baayen, R. Harald and Petar Milin. 2010. Analyzing reaction times. International Journal of Psychological Research 3(2): 12–18.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language 68(3): 255–278.

Bates, Douglas, Martin Maechler and Ben Bolker. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://cran.r-project.org/package=lme4.

Bates, Douglas M. To appear. lme4: Mixed-effects modeling with R. New York: Springer.

Bayley, Robert. 2002. The quantitative paradigm. In Chambers, J.K, Peter Trudgill and Natalie Schilling-Estes

(eds.), The handbook of language variation and change. Oxford: Blackwell. 117–41.

Becker, Kara. 2009. /r/ and the construction of place identity on New York City's Lower East Side. Journal of Sociolinguistics 13(5): 634–658.

Bolker, Benjamin M., Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Trends in Ecology and Evolution 24(3): 127–135.

Bowie, David. 2012. Early trends in a newly developing variety of English. Dialectologia 8: 27–47.

Cedergren, Henrietta J. and David Sankoff. 1974. Variable rules: performance as a statistical reflection of competence. Language 50(2): 333–355.

Cieri, Christopher, David Miller and Kevin Walker. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In Lino, M. T., Xavier, M. F., Ferreira, F., and Silva, R., editors, Proceedings of the Fourth International Conference on Language Resources and Evaluation.

Cox, David R. 1958. The regression analysis of binary sequences. Journal of the Royal Statistical Society, Series B (Methodological) 20(2): 215–242.

Cramer, J. S. 2002. The origins of logistic regression. Tinbergen Institute Working Paper No. 2002-119/4.

Denison, David. 2003. Log(ist)ic and simplistic S-curves. In Raymond Hickey (ed.), Motives for language change. Cambridge: Cambridge University Press.

Drager, Katie and Jennifer Hay. 2012. Exploiting random intercepts: Two case studies in sociophonetics . Language Variation and Change 24(1): 59–78.

Efron, Bradley. 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7(1): 1–26.

Erker, Daniel and Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: variable subject personal pronoun expression in Spanish. Language 88(3): 526–557.

Fruehwald, Josef and Laurel MacKenzie. 2011. New results from hierarchical models of the community grammar. Paper presented at NWAV 40, Georgetown University, Washington, D.C.

Gauchat, Louis. 1905. L'unité phonétique dans le patois d'une commune. Halle: Niemeyer.

Gigerenzer, Gerd, Stefan Krauss, and Oliver Vitouch. 2004. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (ed.), The Sage Handbook of Quantitative Methodology for the Social Sciences. Thousand Oaks: Sage. 391–408.

Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for

research and development. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, 517–520.

Gordon, Matthew J. 2006. Interview with William Labov. Journal of English Linguistics 34(4): 332–351.

Gorman, Kyle. 2009. On VARBRUL – or, The Spirit of '74. Unpublished manuscript. http://ling.auf.net/lingBuzz/001080.

Gorman, Kyle. 2010. The consequences of multicollinearity among socioeconomic predictors of negative concord in Philadelphia. In Marielle Lerner (ed.), U. Penn Working Papers in Linguistics 16.2: Selected papers from NWAV 38, 66–75.

Guy, Gregory R. 1980. Variation in the group and the individual: the case of final stop deletion. In William Labov (ed.), Locating language in time and space. New York: Academic Press. 1–36.

Guy, Gregory R. 1988. Advanced Varbrul analysis. In Ferrara et al. (eds.), Proceedings from the 16th Annual Conference on New Ways of Analyzing Variation. 124–136.

Guy, Gregory R. 1991. Explanation in variable phonology: an exponential model of morphological constraints. Language variation and change 3(1): 1–22.

Guy, Gregory, Jennifer Hay, and Abby Walker. 2008. Phonological, lexical, and frequency factors in coronal

stop deletion in early New Zealand English. Poster presented at the 11th Conference on Laboratory Phonology, Wellington, New Zealand.

Hooper, Joan Bybee. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In William Christie (ed.), Current progress in historical linguistics. Amsterdam: North Holland. 95–105.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and toward logit mixed models. Journal of Memory and Language 59: 434–446.

Jaeger, T. Florian and Laura Staum. 2005. That-omission beyond processing: Stylistic and social effects. Paper presented at NWAV 34, New York University.

Johnson, Daniel E. 2009. Getting off the GoldVarb standard: introducing Rbrul for mixed-effects variable rule analysis. Language and Linguistics Compass 3(1): 359–383.

Jones, Owen, Robert Maillardet and Andrew Robinson. 2009. Introduction to Scientific Programming and Simulation Using R. Boca Raton FL: CRC Press.

Kiparsky, Paul. 1982. Explanation in phonology. Dordrecht: Foris.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. Language variation and change 1(3), 199–244.

Labov, William. 1966. The social stratification of English in New York City. Washington DC: Center for Applied Linguistics.

Labov, William. 2001. Principles of Linguistic Change, Volume 2: Social Factors. Malden, MA: Blackwell.

McQuaid, Goldie Ann. 2012. Variation at the morphology-phonology interface in Appalachian English. Unpublished Ph.D. dissertation, Georgetown University.

Mood, Carina. 2010. Logistic regression: why we cannot do what we think we can do, and what we can do about it. European Sociological Review 26(1); 67–82.

Paolillo, John C. 2002. Analyzing Linguistic Variation: Statistical Models and Methods. Stanford: CSLI Publications.

Paolillo, John C. 2013. Individual effects in variation analysis: Model, software and research design. Language variation and change 25(1): 89–118.

Pereira Scherre, Maria Marta, Carolina Queiroz Andrade, Edilene Patrícia Dias, Nívia Naves Garcia Lucca, and Adriana Lilia Vidigal Soares de Andrade. 2012. Pronominal syncretism in the urban area of Brasília – Brazil's capital. Paper presented at Sociolinguistics Symposium 19, Freie Universität Berlin.

Pierrehumbert, Janet. 2001. Exemplar dynamics: word frequency, lenition and contrast. In Joan Bybee and Paul

Hopper (eds.), Frequency and the emergence of linguistic structure. Philadelphia: John Benjamins.

Pinheiro, José C. and Douglas M. Bates. 2000. Mixed-effects models in S and S-PLUS. New York: Springer.

Pitt, M. A. et al. 2007. Buckeye Corpus of Conversational Speech (2nd release). Columbus OH: Department of Psychology, Ohio State University.

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/.

Rousseau, Pascale and David Sankoff. 1978. Advances in variable rule methodology. In David Sankoff (ed.), Linguistic variation: models and methods. New York: Academic Press. 57-69.

Sankoff, David and Suzanne Laberge. 1978. Statistical dependencies among successive occurrences of a variable in discourse. In David Sankoff (ed.), Linguistic variation: models and methods. New York: Academic Press. 119-126.

Sankoff, David, Sali Tagliamonte, and Eric Smith. 2012. Goldvarb LION: A variable rule application for Macintosh. Department of Linguistics, University of Toronto.

Scheipl, F., S. Greven and H. Kuechenhoff. 2008. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Computational statistics and data analysis 52(7): 3283-3299.

Schielzeth, Holger and Wolfgang Forstmeier. 2009. Conclusions beyond support: overconfident estimates in mixed models. Behavioral Ecology 20(2): 416-420.

Sigley, Robert. 1997. Choosing your relatives: relative clauses in New Zealand English. Ph.D. thesis, Victoria University of Wellington.

Sigley, Robert. 2003. The importance of interaction effects. Language variation and change 15(2): 227-253.

Sigley, Robert. 2010. Interaction effects using GoldVarb. Workshop paper presented at NWAV 39, University of Texas at San Antonio.

Smith, Jennifer, Mercedes Durham and Liane Fortune. 2009. Universal and dialect-specific pathways of acquisition: caregivers, children, and t/d deletion. Language variation and change 21(1): 69-95.

Stram, Daniel O. and Jae Won Lee. 1994. Variance components testing in the longitudinal mixed effects model. Biometrics 50: 1171-1177.

Tagliamonte, Sali A. 2006. Analysing sociolinguistic variation. Cambridge: Cambridge University Press.

Tagliamonte, Sali A. 2011. Variationist sociolinguistics: Change, observation, interpretation. Malden, Mass.: Wiley-Blackwell.

Tagliamonte, Sali A. and R. Harald Baayen. 2012. Models, forests and trees of York English: was/were variation as a case study for statistical practice. Language variation and change 24(2): 135-178.

Van de Velde, Hans and Roeland van Hout. 1998. Dangerous aggregations: a case study of Dutch (n) deletion. In C. Paradis et al. (eds.), Papers in Sociolinguistics. Québec: Éditions Nota bene. 137-147.

Wolfram, Walter A. 1969. A sociolinguistic description of Detroit Negro speech. Washington DC: Center for Applied Linguistics.