Language Variation and Change, **24** (2012), 59–78.
© Cambridge University Press, 2012 0954-3945/12 $16.00
doi:10.1017/S0954394512000014

Exploiting random intercepts: Two case studies in sociophonetics

KATIE DRAGER
University of Hawai'i at Mānoa

JENNIFER HAY
New Zealand Institute of Language Brain and Behaviour
University of Canterbury

ABSTRACT

An increasing number of sociolinguists are using mixed effects models, models which allow for the inclusion of both fixed and random predicting variables. In most analyses, random effect intercepts are treated as a by-product of the model; they are viewed simply as a way to fit a more accurate model. This paper presents additional uses for random effect intercepts within the context of two case studies. Specifically, this paper demonstrates how random intercepts can be exploited to assist studies of speaker style and identity and to normalize for vocal tract size within certain linguistic environments. We argue that, in addition to adopting mixed effect modeling more generally, sociolinguists should view random intercepts as a potential tool during analysis.

In the study of language variation and change, there is a long tradition of clustering individuals into structured groups, based on social factors such as age, gender, and social class. Across these groups, we observe the productions of (often binary) variables in order to make inferences about the underlying social patterns. Logistic regression models, such as those implemented in Varbrul, have been the method of choice for binary data, and for continuous data, simple linear regression models have been used.

One statistical critique of regression models without random effects is that outliers can affect reported trends. In contrast to simple regression models, mixed effects modeling allows individual speakers to vary in the model as "random effects." As such, we can test whether there are differences among groups that are

This paper has benefited from helpful comments from Daniel Ezra Johnson, Rena Torres Cacoullos, and three anonymous reviewers. We would also like to thank Sali Tagliamonte and the audience of the workshop on using statistical tools to explain linguistic variation at New Ways of Analyzing Variation (NWAV) 38.

robustly present across the dataset, and we can be more confident that the trends are not carried by one or two individuals. This increase in statistical robustness is the primary reason why the field should move beyond simple regression modeling (Baayen, Davidson, & Bates, 2008; Johnson, 2009; Quené & van den Bergh, 2008). But there is an additional reason why the mixed effects model is a useful tool for the sociolinguist, and it is this second benefit that we focus on in this paper.[1]

Simple regression models group individuals together into stratified groups; the models – by their very design – provide no information about individual variation. Yet studies are increasingly focused on the speech of a single individual; speaker style is emerging at the core of the sociolinguistic enterprise (Eckert, 2008; Podesva, 2007; Zhang, 2005). Mixed effects models provide a way of studying group patterns, while also investigating variation at the individual level. That is, we do not have to choose between regression modeling on the one hand (dispensing with the study of individuals) and qualitative analysis on the other (dispensing with statistical rigor). A mixed effects model is a tool that enables us to investigate the individual and the group together.

To demonstrate some possibilities of how researchers could use random effect intercepts in their work, this paper presents two case studies in which we have used the random intercepts from mixed effects modeling to learn about the behavior of individuals. One case study uses random effect intercepts to examine which individuals in an ethnographic study are varying most from observed norms. The second case study uses random intercepts as a method of vocal tract normalization by using the random effects as predictors in subsequent models. That is, once we had an assessment of individual variation along some dimension, that variation could be investigated as a predictor of individual behavior in another dimension. The specific research question that was investigated this way is: Is the acoustic realization of an individual's phoneme in one context predictive of that individual's production of it in another context?

It is important to be clear from the outset that we are not claiming that examining the by-speaker random effects from mixed effects models is the only way to answer these types of questions. However, because there are good independent reasons for the community to be using mixed effects models, it seems worth exploring the benefits of random intercepts. Since our analyses already generate random intercepts, we should exploit them.

MIXED EFFECTS MODELS

The term "mixed effects" model is sometimes also used to refer particularly to hierarchical or multilevel models, which have nested random effects. We are focusing in this paper on relatively simple mixed effects models where random effects are crossed and independent (Baayen, Davidson, & Bates, 2008).

The type of mixed effects model that we are focusing on here is very similar to simple linear or logistic regression models, such as that implemented in Varbrul. All of these regression models allow for multiple predicting factors (called "fixed effects" in a mixed effects model). For each possible predictor, both the magnitude of the effect and the probability of it being due to chance are calculated within the context of all other factors included in the model. Regression models are used to generalize beyond the data to which they are fit, using each of the independent factors to predict what trends there would be if new data were collected. In linear regression, the model bases these predictions on continuous data (e.g., formant values or reaction times), whereas in logistic regression, the log odds of a binary factor (e.g., /t/-/d/ deletion) are calculated. In sociolinguistic work, the predicting factors (or fixed effects) include potential influences on the linguistic variable, such as the following phonological environment or the broad social categories assigned to the speaker.

In addition to the benefits provided by a simple regression model, mixed effects models have the advantage of including "random effects." Random effects can be used in both linear and logistic regression models. Random effects are factors that are sampled from some population; factors should be treated as random effects if they are part of a larger population that has not been sampled exhaustively. In linguistic analyses, random effects are most likely to be individual speakers or stimulus items. Fixed effects, on the other hand, are factors that the researcher expects would be observed when sampling new individuals or items. In mixed effects models, each factor (e.g., an individual speaker) of a random effect (e.g., all participants) is assigned a coefficient, thereby providing a way for the factors to vary from one another. This is desirable because it can reduce the risk of a false positive among the fixed effects. For example, a fixed effect such as GENDER might reach significance as a predictor in a simple linear or logistic regression model, even when only one of the speakers analyzed is behaving differently from all of the others. In a mixed effects model, this speaker is assigned his or her own random intercept, which accounts for the extreme divergence from the norm that was exhibited by this speaker. The structure of the model is such that the significance of GENDER can be assessed while taking inter-speaker variation into account. As long as each individual is coded with one and only one gender term, both the effect of the individual and the effect of gender can be taken into account simultaneously. This contrasts with a simple regression model where the concurrent inclusion of the individual and a category label associated with that individual cannot be done because these variables are, of course, not independent.

FITTING MIXED EFFECTS MODELS

Mixed effects models can be implemented using one of the many statistical tools available. While both authors of this paper use the statistical package R, the methods described herein are not limited to analysis in R. Which statistical package is preferred will vary across researchers depending on how they prioritize currency, ease of use, flexibility of different types of analyses, and cost (R is free whereas, for example, SPSS is not). Rather than discuss the advantages and disadvantages of available tools, we discuss the concepts underlying random intercepts and present case studies of how we have exploited the intercepts for work in sociolinguistics.

When you fit a mixed effects model, the output includes a set of fixed effect coefficients and random effect intercepts. The intercepts for the fixed effects are to be interpreted just as they would be in a standard regression model (see Hay, 2010, for an introduction). The output provides different intercept values for the random effects in your model. Most sociolinguistic models should probably have two sets of random effects: one for the individual, and one for the word or item. In this paper, we focus on random effects for individuals.

INTERPRETING RANDOM INTERCEPTS FOR INDIVIDUAL SPEAKER-HEARERS

Random intercepts are calculated based on how much statistically unexplained variation there is for each factor in a random factor group. When SPEAKER is included as a random effect in a mixed effects model, the model's intercept is the model's "best guess" at how the population (with new speaker-hearers) would behave, and the random intercepts are the by-subject adjustment to the model's intercept. Thus, each speaker's random intercept provides an indication of how much that individual's trend diverges from the predicted trends set forth in the statistical model. Because random intercepts are calculated within the context of the fixed effects included in the model, any trends across the random intercepts of multiple individuals can be understood as being "above and beyond" the trends that may be due to the fixed effects included. For example, if following phonological environment (FOLLOW) is included as a fixed effect, the model's coefficients for both random and fixed effects are calculated while holding FOLLOW constant. Provided that a token is consistent with the default of any predicting factors in the model, speakers have a roughly equal likelihood of producing either variant of a binary factor if the sum of the model's intercept and the speaker's random intercept are equal to 0[2].

When trying to interpret random intercepts, we should look at both direction and degree of the divergence. The direction is indicated by whether the intercept is positive or negative, and the degree is indicated by the intercept's value (which can then be compared to the other random intercepts). For example, speakers with a positive intercept are more likely to use the variant being modeled, and speakers with a negative intercept are less likely to use it. When modeling the likelihood that a speaker will produce an innovative linguistic variant, the leaders of the change in progress would have positive intercepts and the speakers who produce more conservative variants would have negative intercepts. In terms of the value of the intercept, the closer the value is to 0, the less that speaker diverges from the overall pattern captured by the model. So, for linear regression models, a greater value for the intercept indicates a greater estimated value for the dependent variable. For logistic regression models, a larger intercept value indicates a greater likelihood of the dependent variable being Factor A rather than Factor B. (Which one of the factors is Factor A must be determined prior to fitting the model). For example, if a logistic regression model was fit to data investigating variation of (ING) and we were interested in determining the likelihood that a speaker will produce [ɪn] rather than [iŋ], speakers with a positive random intercept would be more likely to produce [ɪn] relative to speakers with a negative random intercept. Speakers with a random intercept value that is closer to 0 diverge less from the model's intercept than those speakers with a value further from 0. Thus, random intercepts provide a gradient measure of examining the extent to which speaker-specific variation is consistent with the trends observed across all of the speakers sampled.

EXPLOITING RANDOM INTERCEPTS

Here, we provide two case studies of sociolinguistics projects recently conducted in New Zealand, both of which used mixed effects modeling in their analyses. We focus on the aspects of the analyses that involved examination or utilization of random intercepts. The first case study is a sociolinguistic ethnography investigating the speakers' construction of identity, and the second case study is a corpus-based study of /r/-SANDHI. Through the case studies, we attempt to demonstrate how random intercepts can be used to investigate the construction of style and as a technique to normalize for vocal tract length in certain restricted contexts.

CASE STUDY 1: SELWYN GIRLS' HIGH

The first dataset we consider was collected by the first author during a year-long ethnography of an all-girls' high school, referred to here as Selwyn Girls' High. In addition to participant observation, casual conversations with the girls were recorded, and some of the girls took part in a series of speech perception experiments. Because both quantitative and qualitative data were collected, mixed effects models can be fit to the quantitative data, and tendencies of the different speakers can be interpreted with the help of qualitative analysis.

The School and the Students

Selwyn Girls' High is located in the city of Christchurch, New Zealand. The student body is a mix of students from different Christchurch suburbs, and students come from a range of socioeconomic backgrounds. All of the girls who took part in the study were in their 13th and final year. One of the privileges of being in the most senior year was that they were the only girls at Selwyn Girls' High who were not required to wear uniforms. They were also the only students with access to the Common Room (CR): a room with a microwave, multiple beanbag chairs, and a stereo. The room was available for all Year 13 students but only some chose to eat lunch there on rainy days.

Drager (2009) argued that while there were a number of different cliques at the school, there was a fundamental difference between the girls whose group ate lunch in the common room (CR girls) and those whose group did not (NCR girls). Where common room girls described themselves as “normal” and viewed being “normal” as a positive attribute, non–common room girls described themselves as different from the other girls and viewed being “normal” as a negative attribute. While behavior among non–common room girls’ cliques diverged from the common room girls’ behavior in different ways, non–common room girls shared the stance that they were “different”; they were not common room girls.

Of course, not all common room girls ascribed to common room norms to the same degree and not all non–common room girls rejected the common room norms outright. In fact, there were non–common room girls, such as Holly, who wanted to be friends with common room girls, and others, such as Tania, who had previously been in a common room group. One way that the range of stances could be analyzed would be to assign a numerical value to each girl, representing something akin to “the amount that she ascribes to common room values.” Using this technique would permit the researcher to include the scale as a fixed (predicting) factor in a statistical model, removing the possibility of exploiting the random effect intercepts. While scales can certainly be worthwhile, there are also dangers in trying to quantify qualitative data. Quantification loses the detail and descriptive features that make qualitative data so valuable; numbers on a scale can only go so far in explaining who someone is. While biographical descriptions can also only go so far, they can evoke images and emotions usually lacking in numbers: images and emotions that are important in understanding how a person identifies.

A second danger in using a scale is that the researcher may overlook some difference between speakers or words, a difference that may nonetheless pattern with the linguistic variable and, more importantly, be socially or linguistically meaningful. Examining the random effects in the same way as described for this case study allows researchers to explore potentially meaningful differences between their random factors; a researcher may notice a difference in the distribution of some factor (e.g., one syllable words have positive intercepts and multisyllabic words have negative intercepts) and then revise their model to include a fixed effect for syllable number. Thus, the random intercepts are a useful tool in the process of exploratory data analysis and model fitting. This kind of data trimming may not always be desirable for social data; keeping the data encoding individual variability as a description rather than reducing it to a number can potentially provide a richer account of the social information and present the readers with more information so that they may be better able to determine whether or not they agree with the author. Ultimately, what method is appropriate depends on the nature of the data and the author’s goals.

Rather than using a scale, descriptions of speakers alongside quotes from the speakers themselves could be used to give the reader a sense of how the individuals relate to each other and to society more generally, without turning the social factor into a number or a label. When such descriptions are desirable, random intercepts can be used. Using speaker descriptions and quotes to represent social data is certainly nothing new in sociolinguistic work (see, e.g., Eckert, 1996a; Mendoza-Denton, 2008), but combining these descriptions with random intercepts is new, and it is this technique that is presented in the first case study. Two alternatives would be to use percentages (which would mean that other predicting factors would not be controlled statistically) and fitting separate models to data from each speaker. But through exploiting random intercepts, a researcher can combine statistically controlled quantitative analysis of a linguistic variable with qualitative treatment of social factors.

Phonetic Variation at Selwyn Girls' High

To examine whether the socially based distinction between common room and non–common room girls was made evident in the girls' speech, acoustic phonetic analysis was conducted on recordings of casual speech produced by 28 different girls, 14 of whom were non–common room girls. The analysis focused on variation of the word *like*, a word with a number of different grammatical functions including the discourse particle as in (a) and the quotative as in (b).

a. Lily was LIKE checking out my brother. (Kanani, CR girl)
b. And Mum's LIKE "turn that stupid thing off." (Marama, NCR girl)

At Selwyn Girls' High, the use of discursive functions of *like* is highly salient. Words such as quotative *like* that are themselves socially meaningful may serve as foci of socially meaningful phonetic variation (Eckert, 1996b). Upon examining the phonetic realizations of tokens of *like* from Selwyn Girls' High, it became evident that /k/ realization was linked to the girl's identity but that it was linked differently for the different discursive functions. Because /k/ realizations pattern differently across two different discursive functions, we can be confident that the observed pattern is not a result of /k/ realizations more generally across the different groups.

The analysis revealed phonetic variation that depended on the relationship between the grammatical function of the token of *like* (TYPE), whether or not the /k/ was released (K-REL), and whether the speaker was a common room or non–common room girl (GROUP). A detailed description of the analysis and the phonetic variation observed is presented in Drager (2009). Here, we describe only the result that is relevant for the discussion of exploiting random intercepts.

Due to the correlation of different phonetic factors, TYPE was first treated as the dependent variable and K-REL was treated as a predicting factor. In predicting whether the token of *like* was the discourse particle or the quotative (TYPE), a token was more likely to be quotative *like* than discourse particle *like* if it was produced by a common room girl and the /k/ was not released. K-REL appears to be related to a combination of both a token's function and the speaker's group. This provides evidence that socially conditioned phonetic variation can be observed across different lemmas, even when they share a word form. The use of discursive functions of *like* was highly salient at Selwyn Girls' High; the girls were aware that they used the discursive functions and they believed that some girls (all of whom were common room girls) used them more often than everyone else at Selwyn Girls' High. Given the salience of the discursive functions, it is possible that the girls used the realizations of *like* to index their stance as "normal" or "different" from other girls at the school.

But how do trends in the speech of each individual relate to the tendency for common room girls to release the /k/ in the discourse particle and non–common room girls in the quotative? And are the nonlinguistic components of an individual's style consistent with their ranking in terms of /k/ realization? To investigate the ranking of a speaker in terms of their likelihood of releasing the /k/ in the discourse particle as opposed to the quotative, the random intercepts for each speaker were calculated by fitting a mixed effects model to the data for K-REL, with following phonological environment (FOLLOW) included as a predicting factor (Drager, 2009). Included as a random effect in the model was an interaction between TYPE and SPEAKER. Including the interaction as a random effect produced two intercepts per speaker: one indicating their use of /k/ release for the discourse particle and one indicating their use of /k/ release for the quotative.

Although GROUP would have been a significant predictor in the model, it was not included as a fixed effect because it would complicate the current purpose of fitting the model to examine the random intercepts. If GROUP was included in the model, the random intercepts would still provide an indication of divergence. However, for each girl, we would have an estimate of the degree to which she deviated from the trends predicted for her by the fixed effects in the model. Thus, high intercepts would indicate how much more likely a girl is to release /k/ than is her group in general. We conceived of group membership as situated within a greater continuum of stance, where common room girls tended to value being "normal" and non–common room girls tended to value being "different." Thus, with GROUP omitted from the fixed effects, the random effects could potentially represent a continuum of the speakers' stances in regard to Selwyn Girls' High norms.

For each speaker, the model provides a value that reflects the difference between the random intercept for the discourse particle and the quotative (K-REL:TYPE), thus indicating the likelihood that a given speaker will release the /k/ in one function of *like* when compared to the other function. The value does not provide any insight into how frequently that speaker generally releases the /k/ in *like*. The girls' K-REL:TYPE values are shown in Table 1. Keep in mind that these values were calculated without including GROUP in the model; any patterns relating to GROUP arise strictly from the ranking of K-REL:TYPE.

The values in Table 1 are listed in increasing order so that we can examine a speaker's likelihood of releasing /k/ in the discourse particle rather than the quotative, relative to the other speakers. Looking first at the direction of the trends, we compare positive and negative values. Those which are positive indicate that the speaker is more likely to release /k/ in the quotative (i.e., the "non–common room trend"), and those which are negative indicate that the

TABLE 1. *Likelihood of an individual releasing the /k/ in discourse particle* like *compared to quotative* like *(K-REL:TYPE), listed from smallest to largest. K-REL:TYPE is the difference between random intercepts when the token is a discourse particle and when it is a quotative, for each speaker. All of the names provided are pseudonyms*

Speaker	Group	K-REL:TYPE
Barbara	CR	− 1.84791
Clementine	CR	− 1.71651
Rochelle	CR	− 1.47306
Rose	CR	− 1.33595
Holly	NCR	− 1.26401
Betty	CR	− 1.01303
Meredith	NCR	− .85033
Juliet	CR	− .74285
Tracy	CR	− .72341
Bianca	NCR	− .67159
Emma	CR	− .65743
Tania	NCR	− .59702
Katrina	CR	− .40379
Sarah	NCR	− .38485
Justine	CR	− .38075
Mariah	NCR	− .14698
Theresa	NCR	− .09286
Christina	CR	.015748
Jane	CR	.13346
Marissa	NCR	.281689
Kanani	CR	.561684
Marama	NCR	.589017
Patricia	CR	.746599
Isabelle	NCR	.967743
Vanessa	NCR	1.024424
Esther	NCR	1.130588
Joy	NCR	1.789199
Santra	NCR	1.994998

speaker is more likely to release /k/ in the discourse particle (i.e., the “common room trend”). Roughly 64% of the girls with the “non–common room trend” are non-common room girls, and roughly 59% of the girls with the “common room trend” are common room girls. But what of the exceptions, the girls exhibiting patterns not associated with their group? To interpret their “unexpected” behavior, we can use information gleaned from interviews and participant observation. For example, the best friends of Patricia, a common room girl with non–common room K-REL:TYPE trends, were girls who did not attend Selwyn Girls’ High; Patricia expressed feeling as though her common room clique was not really her group. And Kanani, another common room girl with non–common room K-REL:TYPE trends, had switched from a non–common room group at the beginning of the year. In contrast, Holly (a non–common room girl) has a negative K-REL:TYPE value, meaning that in regard to her patterns of /k/ release, she behaved similarly to common room girls; she was more likely to produce the /k/ in discourse particle *like* than in quotative *like*. Although she did not eat lunch in the common room, Holly might have been more accurately categorized as a common room girl because she ascribed to their norms. For example, the clothes she wore were similar to those worn by common room girls, she attended some of the same parties, and Holly talked about some of the common room girls as though they are friends despite not being seen together at school. While the K-REL:TYPE trends of Patricia, Kanani, and Holly may at first glance appear to be exceptions in regard to their group, qualitative data help to interpret why their speech patterns may not be consistent with their classification as common room or non–common room. Examining the random intercepts provided a way to compare the quantitative data, which could then be interpreted with the help of qualitative data.

The discussion thus far has treated the data as binary (i.e., negative values versus positive values). One advantage of using random intercepts rather than another possible technique is the ability to view divergence from the norm as a continuous variable. For example, Santra, Joy, and Esther (all of whom were non–common room girls) have the largest values for K-REL:TYPE; they were more likely to release the /k/ when producing quotative *like* than when producing discourse particle *like*. This trend is consistent with the pattern associated with the non–common room girls. In contrast, Barbara, Clementine, and Rochelle (all of whom were common room girls) have the smallest values; they were more likely to release the /k/ in the discourse particle than in the quotative. This trend is consistent with the pattern associated with the common room girls. Marissa and Theresa (both non–common room girls) and Jane and Christina (both common room girls) have K-REL:TYPE values near 0. This means that, according to the model, there is a roughly equal likelihood that they will release /k/ when producing quotative *like* and discourse particle *like* (though it does not give any indication about whether they are likely to release the /k/ when producing either function).

Some speakers' K-REL:TYPE trends are stronger than those of their friends, despite being in the same direction. Treating the data as continuous allows us to make comparisons between friends' speech patterns and speculate as to why one speaker may demonstrate a stronger trend in some dimension. For example, Santra has the largest K-REL:TYPE value. Her ranking of K-REL:TYPE among the other girls indicates that she is the speaker who was most likely to drop the /k/ in discourse particle *like* and produce the /k/ in quotative *like*. In other words, she produced the strongest trend in the direction that is statistically associated with non–common room girls. Santra was one of the central members of the Goths, a non–common room group. In fact, she was the only member to wear all black; she was the reason that girls in other groups referred to her clique as the Goths. She was also the only openly bisexual girl in her year and she actively challenged any political or social views with which she disagreed. Her extreme K-REL:TYPE value reflects her active rejection of other norms at Selwyn Girls' High, relative even to other girls with positive K-REL:TYPE values. Compare Santra's behavior to that of her friend, Marissa. Like Santra, Marissa was one of the Goths and was a non–common room girl. Both Santra and Marissa have positive K-REL:TYPE values, but Marissa's value is closer to 0; in terms of /k/ release, she diverges less from the common room girls than does Santra. Is this tendency of less divergence consistent with Marissa's non-linguistic behavior? Yes, from the clothes she wore to the stories she told, Santra was more outrageous than Marissa was. Like Santra, Marissa promoted the idea that she was different from the majority of girls at Selwyn Girls' High, but she did so with a more subtle style than Santra did. This method allows for comparison of speech styles across different speakers without dispensing with the analysis of social categories to which the speakers have been assigned.

Calculating the random effect intercepts (or, as done here, the difference between intercepts from interacting factors) provides a way to examine the extent to which different individuals' behaviors converge on and diverge from the behaviors of the other speakers. It may, therefore, prove useful for analyses where the intercepts can be used to predict behavior in some other dimension. For example, Drager (2009) tested the intercept values from the production of *like* as a predicting factor in a model from a speech perception experiment in order to gauge whether speakers' production patterns were related to their behaviors in perception. The greater number of analytical devices that we as sociolinguists have at our disposal, the more ways we can approach the relationship between speaker identity and language use. This case study demonstrates how random intercepts can be one more tool in our sociolinguistic toolbox. To illustrate how a speaker's random coefficients might be used as a vocal tract normalization technique for consonants in certain contexts, we will turn to our second case study.

CASE STUDY 2: /R/-SANDHI

For the second case study, we turn to a corpus-based investigation of /r/-SANDHI in New Zealand English (NZE), discussed in more detail in Hay and Maclagan (forthcoming). In this study, the speakers' random intercepts were used to offset effects of differing vocal tract lengths.

The dependent variable was F3, measured at the lowest value of F3 during the /r/. F3 is a resonance from cavities anterior to the palatal constriction (see: Alwan, Narayanan, & Haker, 1997; Epsy-Wilson, Boyce, Jackson, Narayanan, & Alwan, 1997; Stevens, 1998) and manifests with lower values when there is a greater constriction.

Because acoustic analysis in sociolinguistics has tended to focus on vowels, there is a large literature addressing the problem of vocal tract normalization for vowel formants. The problem is that speakers have different length vocal tracts, leading to different overall values of F1 and F2 across speakers. A value of F2, for example, that counts as a relatively front vowel for one speaker, might actually be a relatively back vowel for another. A range of techniques has been developed to try and "adjust" vowel values, so they can be sensibly compared across speakers (Adank, Smits, & van Hout, 2004; Fabricius, Watt, & Johnson, 2009; Lobanov, 1971).

These kinds of techniques have not been explored for consonants, although for many consonants the same problem applies. F3 of /r/ is prone to variation by vocal tract length. Some studies have compared raw F3 values across speakers (see, e.g., Hirson & Sohail, 2007), but this is nonideal for the reasons outlined above. It is possible that some techniques could be adapted from the vowel normalization literature, but as far as we know this has not yet been attempted for /r/. Here we describe how we used mixed effects modeling to bypass the vocal tract length normalization problem for our analysis of New Zealand English /r/-SANDHI. We did this by feeding the random intercepts of one model as fixed effects in a second model.

Remember, the random intercept for a speaker represents the degree to which that speaker diverges from the overall trends of the model. Thus, it can be used to compare how divergence in one model (such as one from production) might be related to patterns of other variables (such as one from perception). We will refer to this practice of using values from one model as fixed effects in a different one as *cascading models*.

Before looking at data from the case study, there are some cascading model "rules" we should discuss.

1. Do not include the second model's dependent variable (y) as a predictor in the first model. The first model will hold constant any factors included as fixed effects. Therefore, the fixed effect (y) will be factored out of the first model's random intercepts, rendering them unusable as predictors in future models of y.
2. Recognize that any significant relationship between the speaker's random intercept and the second model's dependent variable is reliant on the other fixed effects included in the two models. All modifications to the first model will lead to changes in the random intercept values. All modifications to the second model can lead to changes in both the fixed effect coefficients as well as the reported p values.
3. Only use by-subject random intercepts from one model as predicting factors in another model if both models are fit to data from the same subjects. Likewise, by-item random intercepts can only be used if the data for the two models are based on the same items.

New Zealand English is nonrhotic, and displays /r/-SANDHI across word and morpheme boundaries. /r/-SANDHI is a grouping term used to describe two related phenomena: linking /r/ and intrusive /r/. Linking /r/ is the label used when there is an <r> in the orthography but the [r] is realized only when followed by a vowel, as in *soaring* or *soar again*. Intrusive /r/, on the other hand, occurs between vowels where an <r> is not represented in the orthography. For example, in New Zealand English, a phrase like *ma and pa* is sometimes realized as *ma-r-and pa*. Intrusive /r/, like linking /r/, can be found word-internally across a morpheme boundary, so that the word *pawing* might be pronounced *paw-r-ing*. Intrusive /r/ and linking /r/ are variably present across word boundaries in NZE. Across morpheme boundaries, intrusive /r/ is variable, and linking /r/ is near-categorical. The rate of /r/-SANDHI differs dramatically across different speakers and is influenced by a speaker's socioeconomic status.

Previous work on New Zealand English has demonstrated that intrusive /r/ is phonetically gradient (Hay & Maclagan, 2010). Hay and Maclagan (2010) found that the value of the third formant (F3), which is an acoustic correlate of the degree of constriction of the tongue when producing an /r/, was predicted by the frequency with which a speaker uses intrusive /r/. In other words, speakers who tend to use intrusive /r/ more often, produce more "/r/-ful" [r]s than speakers who tend to use intrusive /r/ less often.

Formant values are strongly influenced by vocal tract length; a lower F3 in /r/ is found in the speech of people who have a longer vocal tract. In an attempt to account statistically for variability in vocal tract length, Hay and Maclagan (2010) measured F3 in tokens of the word *Sarah* that were produced by the same speakers as those who produced the tokens containing /r/-SANDHI. If all of the interspeaker variation of F3 in intrusive /r/ was due to vocal tract differences, then it should be predictable from the measurements of F3 of /r/ in *Sarah*. Indeed, F3 in *Sarah* was a significant predictor of a speaker's F3 in intrusive /r/. It was not, however, the only predictor. A speaker's rate of /r/-production was a significant predictor even when differences in the F3 of /r/ in *Sarah* were taken into account.

Because Hay and Maclagan (2010) analyzed read tokens, recordings of *Sarah* produced in identical contexts were available across all speakers. This level of control, however, is unlikely when dealing with natural speech.

For example, Hay and Maclagan (under review) conducted a second study in which gradience of F3 in /r/-SANDHI was investigated, but this time the variable was extracted from a corpus rather than read speech. Therefore, a similar approach was taken, but this time the study used random intercepts to assess variation across speakers' "real" /r/.

Hay and Maclagan (under review) were interested in investigating whether the frequency with which a word tends to occur before a vowel can predict the value of F3 in word boundary /r/-SANDHI. Because /r/ does not surface before consonants in New Zealand English, words that usually occur before consonants are encountered less often with /r/. For example, the word *further* is likely to occur before a vowel (because it is likely to be followed by words such as *on, away*, or *apart*) and the word *longer* is likely to occur before a consonant (e.g., very often before *than*). As a result, most New Zealanders have encountered the word *further* more often with [r] realized than they have encountered the word *longer* with [r] realized. Because of this difference between the number of times different words have been encountered with linking /r/, Hay and Maclagan hypothesized that the /r/ in words usually followed by a consonant will tend to have a higher F3 than the /r/ in words usually followed by a vowel. In other words, when in an environment where [r] is more likely to be realized, the [r] may be realized with a more "r-like" pronunciation. This would be a kind of word-based analog of the speaker-based effect observed by Hay and Maclagan

(2010). Put simply: speakers and words which are associated with high rates of /r/-SANDHI might also be associated with lower F3 in the /r/ when it is produced.

To investigate the hypothesized relationship between the F3 value of /r/ and the frequency at which the word occurs before a vowel, Hay and Maclagan (under review) conducted a corpus-based study of /r/-SANDHI in New Zealand English. The tokens analyzed came from the Intermediate Archive, a corpus of spontaneous speech produced by New Zealanders and held at the University of Canterbury (Gordon, Campbell, Hay, Maclagan, Sudbury, & Trudgill, 2004). The research was conducted on the speech of 13 males and 14 females, all of whom were born in New Zealand between 1900 and 1935. For the part of the analysis where random intercepts were used, Hay and Maclagan focused on the subset of the data for which they had the most tokens where [r] was realized: linking /r/ in word-final position.

The 27 speakers who produced the analyzed tokens differ in terms of vocal tract size, which means that the tokens differ in regard to their formant values. One option is to fit a mixed effects model to these data, which will assign each speaker a random intercept and enable us to look at effects over and above the speaker-specific variation. This is a significant advance on a simple regression model, where including non-normalized data from different speakers would be problematic. However, there is a sense in which the random intercept here has the danger of being too powerful. What we want to model is not the overall variation in F3 across speakers and words, but how low the F3 is in each token *for that particular speaker*. For example, a value of 1800 might be very low for one speaker (given their vocal tract length), but very high for another speaker. The model, however, does not have information about the different speakers' vocal tracts and treats both of these values as equivalent. What is needed is an estimate of each speaker's F3 in other contexts, akin to the average F3 for /r/ in the word *Sarah* used by Hay and Maclagan (2010).

To address this, Hay and Maclagan fit a mixed effects model to the F3 value of intervocalic /r/ (as in the word *parent*) from the same speakers upon whose speech the first analysis was based. Because we are dealing with a corpus, there are many tokens of the speakers producing /r/ in a variety of phonological and prosodic contexts. This is problematic because the contexts affect F3 values of /r/, and their distribution is different for every speaker. To remove some of this variation statistically, the predicting factors in the model of intervocalic /r/ were stress pattern (STRESS), the preceding phonological environment (PRECED), and the following phonological environment (FOLLOW). Importantly, GENDER was not included as a fixed effect, even though it would have been significant; males tend to, overall, have lower F3 values in /r/ than females do because they tend to have longer vocal tracts. Refraining from including GENDER in the model is done for the same reason that GROUP was not included in the model in Case Study 1; it was not desirable for this particular model to hold GENDER constant.

To see more concretely why this is true, consider the by-speaker random effect intercepts shown in Table 2. Both sets of intercepts are returned by a model of F3 of /r/ (as in *parents*) and are based on the same dataset. Shown on the left are the

TABLE 2. *Random effects from models of F3 in intervocalic /r/ taken from two models. One of these included a fixed effect for GENDER (shown on the left) and one excluded this fixed effect (shown on the right)*

Including Fixed Effect for Gender			Excluding Fixed Effect for Gender		
Speaker	Random intercept	Gender	Speaker	Random intercept	Gender
Elsie Robinson	−210.995951	f	Dennis Kemp	−258.4538096	m
Dennis Kemp	−164.296758	m	Thomas Ryan	−217.9661735	m
Marjory Gillespie	−155.925678	f	Erik Laytham	−178.685933	m
Thomas Ryan	−128.865476	m	Elsie Robinson	−143.4114129	f
Erik Laytham	−90.671392	m	Marjory Gillespie	−91.9558795	f
Pauline Grither	−72.645094	f	Lance Blackman	−78.70626019	m
Jean Atkinson	−59.186297	f	Elliott Atkinson	−76.37226811	m
Violet Eccles	−42.693805	f	Bill Gillespie	−63.38669711	m
Erna Blackman	−12.223297	f	Thomas McConnell	−54.74834359	m
Lance Blackman	−10.170907	m	John Johnson	−44.37897063	m
Millie Harris	−7.66853	f	David Moore	−38.36916776	m
Elliott Atkinson	2.332993	m	Cap Jardine	−31.08596193	m
Bill Gillespie	17.75938	m	Colin Nicholson	−12.68476126	m
Jocelyn McNae	20.831472	f	**Pauline Grither**	−.08282423	f
Thomas McConnell	23.531232	m	Percy Cox	12.05207983	m
John Johnson	32.580546	m	Jean Atkinson	18.02029132	f
David Moore	38.685201	m	Basil Grither	18.63864374	m
Cap Jardine	48.577169	m	Violet Eccles	21.15628497	f
Colin Nicholson	64.00104	m	Millie Harris	35.743453	f
Marie Dunn	76.284533	f	Erna Blackman	61.12350643	f
Vera Hayward	80.411758	f	Jocelyn McNae	99.91003978	f
Percy Cox	82.459641	m	Marie Dunn	153.9741287	f
Basil Grither	84.077332	m	Nan Hay	157.6235464	f
Mary Direen	84.6478	f	Vera Hayward	164.5654661	f
Nan Hay	88.051334	f	Mary Direen	166.5387607	f
Elizabeth Arnott	99.348861	f	Elizabeth Arnott	183.8342036	f
Kathleen Fountain	111.762892	f	Kathleen Fountain	197.1080589	f

random effect intercepts returned by a model that includes GENDER as fixed effect, and on the right are the random effect intercepts from a model that does not. The speakers in each column are sorted from the lowest to highest value of their random effect intercept.

The dependent variable in this model (F3) is a formant value and, as would be expected when comparing non-normalized formant values, GENDER is highly significant when included as a fixed effect. In a model where it is included as a fixed effect (as in the left column), males and females are more or less equally dispersed through the ordered list of random intercepts. Each speaker's gender is "taken care of" by the fixed effect; the random effect intercepts from this model represent how much each individual varies from the expected values, *given their gender*. In the model where gender is not included as a fixed effect (as in the right column), the random effect intercepts encode the cross-speaker variation, including that driven by gender; the random effect intercepts from this model represent how much each individual varies from the expected values, across all speakers and *regardless of their gender*. Thus, most of the low values are associated with men and most of the higher values are associated with women. For example, compare the position of Pauline Grither across the two lists. When GENDER is in the model, she has a fairly low random effect intercept because her formant values are quite low for a woman. When GENDER is not included in the model, her random effect intercept falls in the middle range because her formant values are not low in the context of all the speakers being considered together.

The fixed effects that are included in a model influence the by-speaker random intercepts outputted by the model; whether it is appropriate to include a fixed effect such as GENDER in your model will depend on the ultimate goals of your study. For Hay and Maclagan, including GENDER in the model would have been the right thing to do if the sole agenda was to understand influences on F3 in the words being modeled. This, however, was not the sole agenda. Instead, Hay and Maclagan wanted to obtain a set of random effect intercepts from this model that could then be used to predict the formant values produced in linking /r/. The random effect intercepts taken from the model that included GENDER could be an appropriate predictor, provided that GENDER was also included as a fixed effect in the model of linking /r/. But Hay and Maclagan desired a more pure estimate of individual variation in F3. They took the random effect intercepts from the model that excluded the fixed effect of GENDER, as this set of intercepts more directly encodes the full set of variation observed across individuals.

Hay and Maclagan (under review) found that when put in their linking /r/ mixed effects model as a fixed effect, the random intercepts from the model of F3 in intervocalic /r/ (shown in the right column) were highly predictive of the F3 values in linking /r/. This is expected; speakers who produce more [r] with lower F3 in intervocalic position also produce linking [r] with lower F3. In this model, GENDER was not a significant predictor of F3, providing evidence that there was no socially based gender variation of linking /r/ beyond that which might be found for /r/ in intervocalic position more generally. Thus, using random intercepts appears to have worked as a type of vocal tract normalization device for this variable, effectively avoiding the need for more explicit normalization procedures. In addition to the intervocalic /r/ random intercept, other significant factors included the rate of use of linking /r/ by the speaker (consistent with Hay & Maclagan, 2010), and the proportion of time a word occurs before vowels (consistent with the hypothesis under investigation).

Note that by using this methodology, Hay and Maclagan (under review) were able to establish that there was no gender-based variation that was particular to the production of linking /r/. If they had opted to use the random effects from the original model that included GENDER as a fixed effect (i.e., the values from the left column of Table 2), GENDER would have been significant in the resulting model of linking /r/. Under these circumstances, it would not have been possible to test for any separate effect of GENDER on the linking /r/ data.

Thus, we can conclude that speakers and words which are associated with high rates of /r/-SANDHI are also associated with lower F3 in /r/-SANDHI tokens. This was established by finding an estimate of speaker F3 in other contexts. The estimate which was used was the random intercept from a model of F3 in intervocalic /r/, which included fixed effects for linguistic environment in the model. Thus, the intercepts provide an estimate of F3 that is comparable across speakers because it holds constant, as much as possible, the variable influence of those other contexts.

While crude, this was an effective and efficient way for Hay and Maclagan to include an estimate of vocal tract length in their model. Of course, it relies heavily on the "estimator" being an appropriate one, and in most cases this will be a difficult thing to establish. There may well be social variation in the realization of intervocalic /r/ of which we are unaware, thereby contaminating its viability as a normalizing value. Additionally, there may be linguistic predictors missing from our model of intervocalic /r/, meaning that they are unaccounted for in the by-speakers random effects and may still influence the model's estimates.

Normalization

Before concluding, it is worthwhile to say more about vocal tract normalization and the degree to which mixed effects models provide some relief in this area. In mixed effects models, each speaker gets his or her own intercept. As such, it becomes statistically legitimate to include, within the same model, data from speakers who have values that span quite different ranges. Thus, one could theoretically take raw (non-normalized) values for F2 of /e/, and include data from many different speakers to examine what predicts this F2 value. Whether this can provide interpretable results, however, depends on your research question. If your question relates to variation across individuals – does social class, for example, affect frontness of /e/ – then this method of normalization should not be used. This is because what counts as a "front" /e/ differs across different speakers depending on vocal tract length. Without the inclusion in the model of some index of vocal tract length, the distinction between whether the variation is socially meaningful or not cannot be made.

However, your research question may be speaker-internal; you may have data from multiple speakers and want to compare two types of tokens from a single individual. For example, in a recent study conducted in New Zealand, Hay and Nilson (2009) were interested in whether New Zealanders use different vowels when talking about Australia than when talking about non-Australian topics. To investigate this, they examined a corpus to find speakers who discussed Australia and then compared variation in two clips of speech (one where Australia was discussed and one where it was not) of equal duration and from the same speaker. Therefore, although there were many speakers, there were observations of both "Australian" and "non-Australian" speech from each speaker. Because the research question related to variation within individuals rather than across individuals, it was valid to use the raw formant values in a single model. Speaker was included as a random effect – which has the approximate effect of centering all of speakers' formant values around a comparable mean – and the topic of speech was included as a fixed effect in the model (where it was significant). Here, the random effect for speaker serves as a normalization tool in the sense that it enables the inclusion of data from many speakers within a single statistical model. It is effective in this way because the question of interest does not involve comparisons across speakers. However, any time that interspeaker comparisons are desired, there needs to be some way of understanding what counts as an extreme value *given those speaker's vocal tract length/vowel space*. This requires either normalization in the classical sense or the inclusion of an appropriate predictor in the model, such as that described above and used by Hay and Maclagan (under review).

CONCLUSION

The goal of this paper is to demonstrate some of the different ways in which random intercepts might be used by sociolinguists. To do this, we present case studies from sociophonetic work conducted in New Zealand. The first study shows how random intercepts can be used to rank speakers according to their degree of divergence from a trend, relative to the other speakers analyzed. The second case study demonstrates cascading models, using random intercepts as an alternative technique for vocal tract normalization in the context of a study of linking /r/.

The first case study provides an example of how random intercepts can be used to shed light on how patterns in an individual's speech relate to patterns produced by others in the community; individuals with the smallest values tended to be girls who actively rejected norms, and individuals with the largest values tended to be girls who contributed to the construction of what it meant to be normal at Selwyn Girls' High. Comparing speakers' random intercept values has many possible applications in the sociolinguistic domain. For example, when combined with qualitative data, this technique may assist in research that examines speakers' styles and the construction of personae, or it might be used in studies investigating the effect of dialect contact or a speaker's exposure to certain variables.

Additionally, this paper discusses how random intercepts can be used as predicting factors in cascading models. In the second case study, it is used as an alternative tool for vocal tract normalization, using production patterns in one context as a predictor for patterns in a second context. By cascading models, one could also potentially examine the relationship between the production and perception of a variant or it could be used to test a link between two linguistically unrelated variables, such as trends in a speaker's vowel realizations and the grammatical variants that speaker tends to use.

While the examples provided in this paper come from sociophonetic work, random intercepts can be used for all levels of the grammar as well as for social variables. They may prove helpful when investigating speakers' attitudes or an individual's integration in a social group or network.

As discussed earlier, the types of questions discussed here could be approached using a range of different methods. The investigation of random intercepts is but another tool in our toolbox. The overall case for using mixed effects models in sociolinguistic research is compelling, and mixed effects models have the added benefit that they yield random effects. Random intercepts are not just a superfluous by-product of the model-fitting process; they are interpretable, useful, and could help to shed light on our analyses.

NOTES

[1] This paper focuses on methodology for analyses rather than on the findings. The results discussed in this paper are presented in more detail elsewhere.

[2] Here we are assuming "treatment contrasts" – the default mode in R. When results are reported with treatment contrasts, one factor is chosen as the default for each categorical independent variable. The default factor is then assigned a coefficient of 0. It is also possible to report coefficients using "sum contrasts," which is closer to Varbrul. When using sum contrasts, no factor is chosen as a default, but the values are instead anchored around a mean.

REFERENCES

Adank, Patti, Smits, Roel, & van Hout, Roeland. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America* 116 (5):3099–3107.

Alwan, Abeer, Narayanan Shrikanth, & Haker, Katherine. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *Journal of the Acoustical Society of America* 101:1078–1089.

Baayen, R. H., Davidson D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59:390–412.

Drager, Katie. (2009). A sociophonetic ethnography of Selwyn Girls' High. Unpublished PhD dissertation, University of Canterbury.

Eckert, Penelope. (1996a). (ay) goes to the city: Exploring the expressive use of variation. In G. R. Guy, C. Feagin, D. Schiffrin, & J. Baugh (eds.), *Towards a social science of language: Papers in honor of William Labov*. Vol. 1. Philadelphia: John Benjamins, pp. 47–68.

______. (1996b). Vowels and nail polish: The emergence of linguistic style in the preadolescent heterosexual marketplace. *Proceedings of the 1996 Berkeley Women and Language Conference*. Berkeley: Berkeley Women and Language Group.

______. (2008). Variation and the indexical field. *Journal of Sociolinguistics* 12(4):453–476.

Epsy-Wilson, Carol Y., Suzanne E. Boyce, Michel Jackson, Shrikanth Narayanan, & Abeer Alwan. (1997). Acoustic modeling of American English /r/. *Proceedings of Eurospeech* 1:393–396.

Fabricius, Anne H., Dominic Watt, & Daniel Ezra Johnson. (2009). A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change* 21:413–435.

Gordon, Elizabeth, Lyle Campbell, Jennifer Hay, Margaret Maclagan, Andrea Sudbury, & Peter Trudgill. (2004). *New Zealand English: Its origins and evolution*. Studies in English Language, Cambridge: Cambridge University Press.

Hay, Jennifer. (2010). Statistical analysis. In M. Di Paolo & M. Yaeger-Dror (eds.), *Sociophonetics: A student's guide*. New York/Abingdon: Routledge.

Hay, Jennifer, & Maclagan, Margaret. (2010). Social and phonetic conditioners on the frequency and degree of "intrusive /r/" in New Zealand English. In D. Preston & N. Niedzielski (eds.), *A reader in Sociophonetics*, pp. 41–70. Berlin/New York: De Gruyter Mouton.

______. (under review). /r/-sandhi in early 20th Century New Zealand English.

Hay, Jennifer, & Nilson, Elissa. (2009). Self-priming in New Zealanders' speech about Australia. Paper presented at: Linguistic Society of New Zealand 18th Biennial Conference. November 30–December 1, 2009. Palmerston North, New Zealand.

Hirson, Allen, & Sohail, Nabiah. (2007). Variability of rhotics in Punjabi-English bilinguals. *Proceedings of ICPhS* 16:1501–1504.

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

Lobanov, Boris M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America* 49(2):606–607.

Mendoza-Denton, Norma. (2008). *Homegirls: Language and Cultural practice among latina youth gang girls*. Malden/Oxford: Blackwell.

Podesva, Robert J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics* 11(4):478–504.

Quené, Hugo, & van den Bergh, Huub. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59:413–425.

Stevens, Kenneth N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Zhang, Qing. (2005). A Chinese yuppie in Beijing: Phonological variation and the construction of a new professional identity. *Language in Society* 34(3):431–466.