1pSCa2. Multilevel models, covariates, and controlled factors in experimental speech research: Unified analyses of highly structured data

Noah Silbert*, Jared A. Linck and Mark VanDam

*Corresponding author’s address: Center for Advanced Study of Language, University of Maryland, College Park, MD 20742, nsilbert@umd.edu

Experimental speech research often makes use of complex experimental designs, but even when multiple experimental factors are manipulated, measured outcomes may be influenced by non-controlled and incompletely controlled factors. Multilevel models (of which mixed-effect models are a special case) enable unified analysis of the relationships between, on the one hand, trial-level data and, on the other, experimental factors and potentially important non-controlled variables. Fitted multilevel models allow us to draw inferences simultaneously about group-level experimental effects and covariates (the typical focus of experimental work) as well as individual subject and item properties (both of which can be important in applied research). The utility of multilevel models will be illustrated with analyses of data from a number of studies. We present models of phonological structure, gender differences, and within-gender subject variability in the acoustics of spoken English consonants; simultaneous modeling of experimental factors, subject and item variability, and second language proficiency in bilingual lexical processing; and modeling of the effects of age, hearing loss, phonological/lexical properties, subject and item variability, and multiple vocabulary-related covariates in early language development.
STRUCTURE IN SPEECH, DATA, AND MODELS

Speech is complex, and the experimental study of speech frequently involves the manipulation and measurement of complexly related factors. Yet even the most careful, diligent speech researcher cannot always measure every relevant variable. Because of this, interesting and worthwhile speech research questions are not always amenable to simple statistical procedures. Multilevel statistical models (of which mixed-effects models are a special case) are powerful, flexible tools that enable simultaneous analysis of trial-level data, subject- and item-level factors, as well as controlled and non-controlled variables at various levels. Multilevel models allow researchers to account for the structure of complex experimental designs and draw inferences about experimental effects while simultaneously taking into account variability between subjects and among items.

There are compelling mathematical reasons to use highly structured multilevel models (e.g., to minimize Type I errors, to prevent violations of various statistical assumptions in data; Barr, Levy, Scheepers, & Tily, 2013). The focus of this work is to illustrate the utility of multilevel models in extracting interesting information from highly structured data sets. More specifically, we present applications of multilevel models to three data sets from speech research projects. First, we present an analysis of the (negative) effect of second-language (L2) immersion on first-language (L1) lexical processing while taking overall L2 proficiency into account (Linck, Kroll, & Sunderman, 2009). Second, we use a simple mixed-effects linear regression model to analyze the acoustics of fricatives and how spectral properties vary between and within sex categories (de Jong, Silbert, Regier, & Albin, 2011; Silbert, de Jong, Regier, & Albin, 2012). Finally, we use a use a more complex mixed-effects model to analyze the effects of age, hearing loss, and lexical status in children’s accuracy in a word-repetition task. In addition to providing measures of the robustness of these experimental factors, this model provides additional information about item difficulty, which can have important implications for clinical research (VanDam, Silbert, & Moeller, 2012).

L2 IMMERSION AND L1 LEXICAL PROCESSING

Linck et al. (2009) report a study of the effect of learning context on interference between first and second language lexical items. We consider a subset of their data and analyses here, focusing on the results of the translation recognition task and the relative influence of lexical and semantic neighbor distractors on L1 English students studying Spanish in an immersion program compared with L1 English students studying Spanish in a typical language classroom. On each trial of the translation recognition task, a participant was presented with two words in sequence, the first in Spanish and the second in English. The participant was instructed to indicate, as quickly as possible, whether the words were translation equivalents of one another by pressing one of two buttons.

In order to test the effect of immersion on L1 access, two crucial (English) distractor types were employed in a subset of trials. First, some distractors were orthographic lexical neighbors of the Spanish word. For example, the Spanish word ‘cara’ (‘face’ in English) would be paired with the word ‘card’ in the lexical neighbor distractor condition. Second, some distractors were semantic neighbors. For example, the Spanish ‘cara’ would be followed by ‘head’ in the semantic distractor condition.

Of course, individual subjects can vary substantially in their lexical processing speed and their second language (L2) proficiency, and particular word pairs may be more or less difficult to recognize as correct translation equivalents. Analyzing the translation recognition response times with a multilevel linear regression model enables analysis of the interaction between distractor type and learning context (i.e., the experimentally manipulated factors) while simultaneously modeling variation between individual subjects and between specific word pairs (i.e., random intercepts for subjects and items, respectively) and taking into account subjects’ L2 proficiency.

The three way interaction between distractor type, learning context, and distractor relatedness (i.e., whether the distractor was a control or a lexical/semantic neighbor) accounted for a statistically significant proportion of the variance in the translation recognition reaction time (RT) data. Model comparisons indicated that including random intercepts for subjects and items also improved the model fit significantly. Figure 1 shows the mean interference magnitudes (i.e., the mean RT to related distractors minus the mean RT to unrelated control words, in ms) produced by lexical and semantic neighbors for the immersed (blue) and classroom (red) students. Critically, these effects remained, even after controlling for individual differences in L2 proficiency, thereby indicating a fundamental shift in L1 lexical processing in the immersion context that did not simply reflect an increase in L2 proficiency. This theoretically informative inference is based on a complicated set of relationships in the data that was readily detected with the multilevel linear regression model.
Silbert et al. (2012) reported analyses of a speech production corpus involving a subset of English obstruent consonants produced 10 times by each of 20 native speakers (10 male, 10 female; see also de Jong et al., 2011). The focus in the original work was on the mapping between multiple phonological dimensions (e.g., voicing, place, manner) and multiple acoustic dimensions (e.g., consonant duration, spectral mean, noise amplitude). Here we re-analyze a subset of this acoustic data, focusing on the effects of voicing, place of articulation, and speaker sex on the spectral mean of alveolar and alveo-palatal fricatives.

The nesting of production repetitions within speakers and of speakers within sex categories highlights the utility of multilevel modeling. A multilevel model allows sources of variability to be partitioned in accordance with the structure of this nesting. First, for any given speaker, there is variability between repetitions. The repeated measurement of spectral mean values for each speaker allows the effects of voicing and place to vary across subjects (along with baseline [intercept] values). Second, within each sex there is variability between individual speakers. At the group level, the parameter estimates for speaker sex, consonant voicing, and consonant place (as well as an interaction between sex and place) allow us to draw inferences about how spectral mean varies as a function of these experimental variables.

Figure 2 illustrates the multilevel structure of the data and the group-level fitted-model estimates of spectral means. Each small symbol indicates the observed spectral mean of a single token produced by a single speaker. The x-axis indicates individual speakers, with red symbols indicating male speakers and blue symbols indicating female speakers. The filled markers correspond to voiced fricatives, the open markers to voiceless fricatives. The triangles

**FIGURE 1.** Translation recognition interference (in ms) as a function of distractor type (lexical vs. semantic) and learning context (immersed vs. classroom).
correspond to alveolar fricatives and the circles to alveo-palatal fricatives. The large symbols indicate the model-estimated group-level spectral means.

![Spectral Mean of Fricatives](image)

**FIGURE 2.** Spectral mean of fricatives as a function of place, voicing, and sex. Triangles indicate alveolar [s] and [z]; circles alveopalatal [ʃ] (‘sh’) and [ʒ] (‘zh’). Filled symbols indicate voiced fricatives, unfilled voiceless. Red symbols indicate male speakers, blue symbols female. Large symbols indicate fitted model group-level spectral mean estimates.

There is substantial variability both within and between individual speakers, though the general pattern is that the spectral means for alveo-palatal fricatives are (approximately) equivalent across speakers and speaker sex, while spectral mean is significantly higher for alveolars produced by females than for those produced by males. A multilevel model provides a rigorous statistical treatment of the hierarchical structure of the dataset, allowing the analysis to, on the one hand, take into account the complex variance relationships among subsets of the data (e.g., trial-level observations within subjects, between-subject differences within sex) and, on the other, draw inferences about differences in spectral mean values induced by differences in voicing, place of articulation, and speaker sex.
AGE, HEARING, AND LEXICAL PROPERTIES IN LEXICAL DEVELOPMENT

VanDam et al. (2012) presented a study of the effects of age, hearing loss, lexical status, phonotactic probability, vocabulary size, and articulatory ability on children’s accuracy in a word repetition task. Each of 61 children listened to and repeated 68 words and 32 non-words that varied with respect to their phonotactic probabilities. Thirty of the participants were normal hearing, and 31 had mild to severe hearing loss without secondary disabilities. All children with hearing loss were hearing aid users and used their hearing aids during testing. Half of the children with hearing loss were 4 years of age, and half were 7 years of age; 17 of the normal hearing children were 4 years of age, and 14 were 7 years of age. Repetition accuracy (as judged by a trained phonetician) was tallied for each phonological form for each child. Vocabulary size and articulatory abilities were also measured for each child.

A multilevel logistic regression model (fit using the lme4 package in R; Bates, Maechler, & Bolker, 2012; R Development Core Team, 2012) examined the relationships between trial-level accuracy, the experimental variables age, hearing loss, lexical status, and phonotactic probability, and the measured, non-controlled variables vocabulary size and articulation ability. More specifically, the multilevel structure enable simultaneous analysis of (a) group-level effects of age, hearing loss, vocabulary, and articulation, (b) subject-specific and group-level effects of lexical status and phonotactic probability, and (c) variability among stimuli (items) and with respect to child-specific effects (subjects). Figure 3 illustrates the effects of age, hearing loss, and lexical status (word vs. non-word).

**FIGURE 3.** Effects of age, hearing loss, and lexical status. Gray symbols indicate trial-level fitted model accuracy estimates. Large circles and large triangles indicate group-level fitted model accuracy estimates. Upward pointing triangles indicate word accuracies, downward pointing triangles non-word accuracies. Red symbols indicate accuracies for normal hearing children, green symbols accuracies for children with hearing loss. Small (red/green) triangles indicate younger children’s accuracies, larger (red/green) triangles older children’s accuracies. The in-figure text provides group-level parameter estimates (S.E.s).
The main effects of age and hearing loss and the interaction between age and hearing loss were statistically significant. Older children were more accurate than younger children, and children with normal hearing were more accurate than children with hearing loss. The interaction between age and hearing indicates that the older children with hearing loss are less accurate than would be expected based on the main effects of age and hearing loss. The effect of lexical status varies across subjects, though it is consistently positive, with repetition of real words more accurate than repetition of non-words. As in the previous two examples, a multilevel model allows us to address variability between subjects and among items, take non-controlled covariates into account, and simultaneously test for effects of experimentally controlled variables.

The multilevel model also allows us to directly inspect the relative difficulty of different stimulus words. For basic scientific research, it is often enough to allow a model to account for such differences 'under the hood,' in a fashion that does not examine item-level effects. However, direct consideration of item properties can be important for stimulus selection for clinical tests, since extremely difficult and extremely easy items may not provide as much useful diagnostic information as less extreme items.

Figure 4 shows a histogram and estimated kernel density of the item-specific intercept offsets (i.e., deviations from the overall intercept). Most of the items deviate very little from the overall level of accuracy (i.e., most are near 0 in the figure), and a few deviate substantially. As discussed by Gelman and Hill (2006), the estimated maximum change in accuracy probability associated with a unit change in a predictor in a logistic regression model is obtained by dividing the predictor’s parameter by four. Using the same logic to compare maximum differences in expected accuracy across items, the items with intercept deviations near +1 should be up to, but no more than, 0.25 more accurate than the average item, and items near −1 should be no more than 0.25 less accurate. Depending on how much consistency is desired in the stimulus set, it may be worthwhile to exclude items that deviate more than this in either direction. Note that one item is extremely difficult – the intercept offset is nearly −2, corresponding to a maximum possible difference of −0.5 in accuracy probability relative to the average item. Further work along these lines in both applied and theoretical contexts may benefit from the exclusion of this item.

![Figure 4](image_url)

**FIGURE 4.** Item-specific intercept offsets. Positive offsets indicate easier items, negative offsets more difficult items.

**CONCLUSION**

The three examples considered here demonstrate a number of the benefits of multilevel statistical models in experimental speech research. In the first example, individual differences in proficiency (whether of theoretical interest or merely acting as a nuisance variable) were explicitly accounted for in the model, thereby directly assessing whether a potential confound impacts the inferences to be drawn from the experimental manipulations and subsequent statistical analysis. In the second and third examples, complex hierarchical structures and associated dependencies among subsets of the full dataset were explicitly modeled, leading to a richer, more accurate understanding of the phenomenon under examination. In the third example, we also saw that hierarchical models allow a direct examination of variability driven by specific stimuli used in an experiment, which can inform the selection (or exclusion) of specific items to enhance measurement in future studies or clinical applications. By explicitly incorporating a dataset’s multilevel structure into statistical analyses, researchers can examine the relationships between variables more fully and more rigorously, leading to better understanding of the myriad phenomena that influence speech.
ACKNOWLEDGMENTS

This work was supported by grants from the National Institute on Deafness and Other Communication Disorders (grants R01DC006681 and T32 DC000013-31). Opinions and statements are the authors’ only.

REFERENCES


