METHODOLOGICAL ISSUES IN THE ANALYSIS OF PHONOTACTIC PROBABILITY EFFECTS IN NONWORDS

Mary E. Beckman[a], *Benjamin Munson*[b] *& Jan Edwards*[c]

[a]Ohio State University, USA; [b]University of Minnesota, USA;
[c]University of Wisconsin–Madison, USA
mbeckman@ling.osu.edu; munso005@umn.edu; jedwards2@wisc.edu

ABSTRACT

When an unfamiliar word consists of phoneme sequences that are attested in many of the words that a listener already knows, it should be easier to incorporate into the mental lexicon. Results of many studies with nonword materials support this idea, showing that a nonce word form with such high *phonotactic probability* sounds more like a possible word, and is repeated and learned more easily. However, the inferential statistical tests that are typically used make it difficult to generalize beyond the particular set of nonword items used in any one study, and there are few replications. This paper addresses these methodological issues. It uses a partial replication of one earlier study in previously unreported data for filler items from a second, to describe how the use of mixed-effects models can resolve problems with prior analyses.

Keywords: phonotactic probability, nonwords, replication, mixed-effects models

1. INTRODUCTION

Nonwords are useful in designing experiments to study how the mental lexicon is organized, because they let us control other properties of words, such as familiarity or age of acquisition, so as to focus on the phonetic form. Experiments using nonwords suggest that having a phonetic form that is more typical of the other words in the language makes a new word easier to repeat [6] and to learn [5]. Such experiments have also suggested that a form can be more or less typical in several ways. For example, in English, more disyllables are stressed on the first syllable than on the second, and more words contain the phoneme sequences [ik] or [ft] than contain [auk] or [fk]. These aspects of typicality combine, so that native speakers of English can be more accurate when repeating nonwords with the more typical stress pattern that also contain more typical phoneme sequences than when repeating nonwords with just one of these two properties [6].

The typicality of a phoneme sequence is often estimated by its log frequency in a sample lexicon, a quantity that is called *phonotactic probability*. For example, using the *Hoosier Mental Lexicon* as a sample English lexicon, we calculate phonotactic probabilities of -9.8 and -11.8 for [ik] and [ft] as opposed to -14.6 and -15.6 for [auk] and [fk]. In one study [2], we made recordings of 22 pairs of nonwords that contained such high- versus low-probability sequences, and played them to 104 English-speaking children aged 3 through 8 years to repeat. The children made more mistakes on the low-probability items. The magnitude of this probability effect differed across the subjects and was correlated with vocabulary size. A later study with an older group of children [3] shows a similar effect, and also that children with specific language impairment (LI) make even more errors on low-probability items than do their age peers.

While the later study found similar results, we cannot call it a replication. The studies differ not just in using different samples of children, but also in selecting different high- versus low-probability sequences embedded in different nonword items. Both sets of items also differed from the items in [5] and [6]. Moreover, the statistical tests in all four studies evaluate the contrast between high- and low-probability items using ANOVA and other related ordinary least squares (OLS) regression models in ways that make the inferences subject to the "language-as-fixed-effect fallacy" [1]. Thus, we cannot generalize confidently from any one study without an exact replication. These facts lead us to ask whether the different studies do in fact show the same effect of phonotactic probability.

2. REPLICATION

To address the question, we reanalyzed data from the studies reported in [2] and [3], including data for 10 filler items in [3] which happened to be 5 of the 22 target nonword pairs in [2] (Table 1). These shared items constitute a kind of replication of the first study, although it is not exact, because the

nonwords were recorded by a male speaker of the Midlands dialect in [2] and by a female speaker of the Northern dialect in [3]. So the audio stimuli as well as the children's responses were obtained in a multi-level sampling scheme, with items selected first from the "population" of possible types and then stimulus tokens sampled from a "population" of possible utterances by a phonetically-trained talker selected from the population of speakers.

Table 1: The 10 nonword types (with target low- vs high-probability VC or CC sequence underlined), the mean wordlikeness of audio stimuli used in [2] (rated on a 1-5 Likert scale in a norming study), and the position-specific log probabilities of target sequences calculated using the *HML*, as described in [2].

(underlined) target sequence, nonword		wordlike-ness (1-5)		log sequence frequency	
low	high	low	high	low	high
dug̲nəted	tʌg̲nədit	2.68	3.03	–14.59	–10.53
au̲ftəgɑ	au̲ntəko	2.43	3.11	–14.59	–8.96
au̲kpəde	i̲kbəni	2.41	2.06	–14.59	–9.77
næfk̲ətu	gʌf̲tədaɪ	2.73	2.44	–15.57	–11.79
dɛg̲dəne	tik̲təpo	2.43	2.54	–15.57	–9.45

Table 2: Number of children in each group, and mean values (and standard deviations) for their ages and scores from tests of receptive (PPVT) and expressive vocabulary (EVT in [2], EOWPVT in [3]).

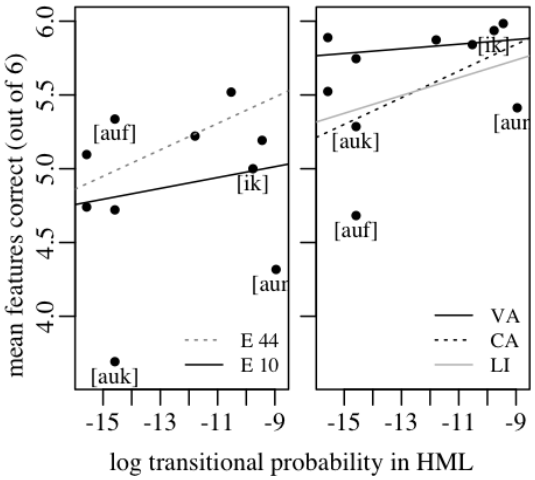
group (N)	age months	PPVT raw	PPVT standard	E(...)VT raw
E (104)	66 (19)	86 (25)	114 (13)	86 (25)
VA (20)	100 (25)	127 (24)	115 (6)	102 (19)
CA (22)	124 (20)	145 (15)	114 (15)	112 (17)
LI (21)	131 (18)	126 (19)	94 (11)	90 (17)

We analyzed productions by four groups of participants (Table 2). From [3], we had data for children with LI (as assessed by the CELF) and for children who were recruited to provide control groups matched for chronological age (group CA) or for raw receptive vocabulary size (group VA) as measured by the PPVT. From [2], we had data for 104 children (group E), of whom 66 were younger than the youngest child in group VA. We scored the target VC or CC of each filler item in [3] as in [2], by counting correct features, evaluating place, manner, and voicing for each C, and high/mid/low, front/central/back, and long/short for each V.

3. ANALYSIS BY ITEMS

Fig. 1 plots mean diphone accuracy as a function of phonotactic probability for the ten shared items. Data for group E are on the left. The solid line is an OLS regression fit to the 10 datapoints. This is the same by-items analysis we used in [2] for the larger set of 44 items, shown here by the dashed line. The correlation was significant for the larger set ($F[1,42]=10.78$, $p<0.01$, $R^2=0.19$), but not for the smaller subset of 10 means shown in the plot. The discrepancy is due to the datapoints for [auk] and [aun]. We identified these as outliers in [2] as well, and their inclusion in these 10 shared items dilutes the relationship between the accuracy score and the diphone's phonotactic probability.

Figure 1: Mean diphone accuracy for each nonword as a function of the sequence frequency, with lines for regression curves fit to means for different groups.



These two items are closer to the main cloud of datapoints in the right panel of Fig. 1, which plots mean accuracy scores for the audio stimuli in the replication study. The relationship again is positive but not significant when evaluated using the same statistical model. Here the failure to reproduce the results of the larger study reported in [2] could be a ceiling effect. The means for these mostly older children are closer to the maximum possible score.

Our failure to reproduce even the primary result of the main by-items analyses in [2] reinforces the importance of replication, as well as the need for cautious inference from the types of statistical tests used in the earlier study. The apparent ceiling effect highlights another problem as well. An OLS test assumes that the response variable is normally distributed. Our six-point accuracy scores violate this assumption. They are count data, and will be approximately normal only if they are not too far from the middle number of 3 features correct. Most of the datapoints are well above this middle region.

4. ANALYSIS BY SUBJECTS

For the earlier report on these data in [2], we used OLS methods in two types of analysis by subjects, for which we treated phonotactic probability as a nominal variable by dividing items into high- and

low-probability groups in order to make it easier to estimate interaction effects. First we applied a repeated-measures ANOVA to the raw accuracy scores. For this analysis, we also treated age as a nominal variable, by dividing the data into three unbalanced groups for 3- and 4-year-olds (N=43), 5- and 6-year-olds (N=38), and 7- and 8-year-olds (N=23). This test showed significant main effects both of probability and of age group, as well as a significant interaction. The low-probability items were less accurate than the high-, the youngest children were less accurate than the older, and they were especially so on the low-probability targets.

For the second type of by-subjects analysis, we took better advantage of the careful design of our nonwords. We had created the 44 items to pit high- versus low-probability in pairs of sequences that differed minimally (e.g., [ft]:[fk]) and embedded them in paired nonword frames that we chose to minimize other potentially confounding differences in properties such as the prosodic position of the matched sequences within their frames and the paired nonwords' mean wordlikeness ratings. We calculated a new response variable by subtracting the accuracy of the low-frequency item from that of the paired high-frequency item. This difference score has the advantage that the values cluster just above 0, the middle of the possible range from -6 to +6, making OLS analyses less inappropriate.

We evaluated the differences scores by a t-test and by several OLS regressions. The t-test showed that the mean difference is positive, confirming that the 104 children's productions tend to be more accurate for the high-probability items than for the low. The first regression analysis confirmed that the magnitude of this effect correlates with age, which we could treat as a continuous variable, overcoming the problem of unbalanced age group sizes in the RM-ANOVA. A second regression showed further that the size of the probability effect is related to the size of the child's expressive vocabulary, as estimated by the log of the EVT raw score. Because the number of words that a child knows how to say grows tremendously in early childhood, these two relationships could be the same result. We therefore did a third regression that included both age and the log EVT raw score. The EVT explained a significant proportion of the variation in children's difference scores even after age was partialled out, but age was not significant after EVT was partialled out.

When we used these two types of by-subjects analysis to probe the weaker effect of phonotactic probability in the subset of 5 item pairs in the left-hand panel of Fig. 1, we reproduced some, but not all of our results reported in [2] for the full set of 22 item pairs. Specifically, the ANOVA showed significant main effects of phonotactic probability (F[1,101]=35.6, p<0.001) and age (F[2,101]=16.7, p<0.001), but no significant interaction. The two-tailed t-test confirmed that the difference between paired high- and low-probability items is generally positive (mean=0.33, t[103]=5.89, p<0.001), but the tests regressing this child-by-child estimate of the size of the probability effect against age and EVT raw score found no relationship with either.

Our failure to reproduce that result from [2] is because the outlier pair [aun]:[auf] has a huge effect on each child's mean difference in accuracy between high- and low-probability items. The fact that there are just as many subjects in these tests as in the analogous by-subjects analyses for the larger set of items makes this a very clear example of the "language-as-fixed-effect fallacy" [1].

5. MIXED-EFFECTS MODELS

As noted in [4], we can avoid the fallacy by using mixed-effects models such as (1), which is the base or "empty" model for analyzing difference scores.

(1) $$difference_{ij} = \gamma_{00} + \pi_{0i} + \omega_{0j}$$

Here, the difference score for child i producing the high- versus low-probability nonwords in pair j is modeled in terms of an intercept γ_{00} (which is the estimated grand mean of all the difference scores) plus the deviations from γ_{00} for that child (π_{0i}) and for that word pair (ω_{0j}). Each of the vectors of N deviations modeled by π and 5 deviations modeled by ω is assumed to be a sample from a normally distributed population of deviations for that grouping factor. The fit of the model to the data is evaluated for a succession of choices of values for the γ_{00}, π, and ω model parameters, to converge on the model that is maximally likely to have generated the set of N*5 difference scores.

We can use this approach in analyzing the raw scores, too, by treating them as count data as in (2).

(2) $$\log\left(p/(1-p)\right) = \gamma_{00} + \pi_{0i} + \omega_{0j}$$

This base uses the logit function to link the counts to the estimates for the model parameters via p, the estimated probability that features will be correct in productions of word j by child i. If the variance

estimated for each random effect in (2) is not too close to 0, we can build a second, more complex model that adds another effect. The two models then can be compared by the likelihood ratio test (LRT), which uses the X^2 distribution to see if one model is significantly more likely than the other to have generated the data, taking relative complexity into account. For example, in order to evaluate whether target features are more likely to be correct in high-probability sequences than in low-, we can build model (3), which is just like the base model in (2) except that it includes γ_{prob}, the fixed effect of probability, treated as a nominal variable.

$$\log\left(p/(1-p)\right) = \gamma_0 + \gamma_{prob} + \pi_{0i} + \omega_{0j} \quad (3)$$

When we built model (3) for the original group E data for all 44 nonword items, we found the estimated value of γ_{prob} to be positive (features are more likely to be correct in high-probability items) and the LRT comparison to the base model in (2), showed (3) to have a better fit to the data (X^2=6.4, df=1, p=0.01), confirming the result of our original t-test but here using an analysis that can generalize to other sets of items as well as to other subjects.

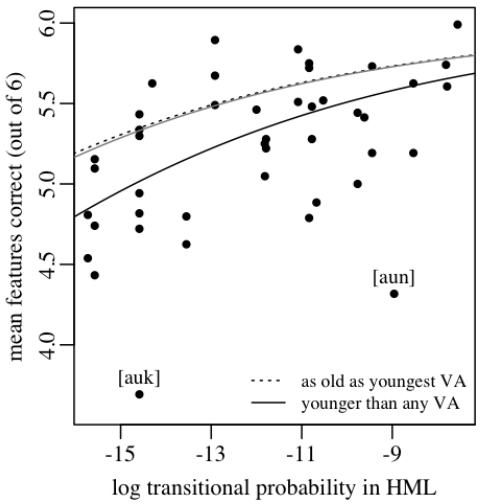
When we built (2) and (3) for the data plotted on the right in Fig. 1, γ_{prob} again was positive and model (3) had a marginally better fit (X^2=3.7, df=1, p=0.05). Thus, this partial replication with a small subset that includes the two items identified as outliers in [2] reproduces what seems to be a "true" probability effect even for the older children in [3], which was obscured by a ceiling effect in Fig. 1.

Another series of models for all 44 items in [2] added fixed-effect terms for probability (treated as a continuous variable), then for age (treated as a three-group nominal variable as in our earlier ANOVA), and finally for their interaction. The last model was not significantly better than the simpler model without the added term. This suggests that the interaction between probability and age that we found in the ANOVA in [2] may be an artifact of a ceiling effect for the older children from that study.

We explored this suggestion by re-building the series with just two age groups, differentiating the 38 children in [2] who were at least as old as the youngest child in [3] from the 66 children who were younger. Fig. 2 plots the estimated effects. Each curve is a straight line when plotted in the logit-transformed space of the models, and bends here only because it approaches the asymptote of 100% correct. The distance between the solid and dashed black curves is the size of the age effect in the simpler of the two models. The distance between the dashed curve and the solid gray curve is the size of the non-significant interaction effect in the most complex model in the sequence.

In future work, we plan to see if these ceiling effects can be circumvented by designing age-graded materials so that older children repeat the target diphones in more complex nonword frames.

Figure 2: Mean diphone accuracy for each of the 44 nonwords in [2] as a function of sequence frequency.



6. ACKNOWLEDGEMENTS

We thank Jennifer Windsor, Beth Kurtz, and Kathryn Kohnert, and reiterate other acknowledgments in [2, 3], including the support of NIDCD grants R01 02932 (to Edwards), R01 004437 (to Windsor), R03 005542 (to Kohnert), and R03 005702 (to Munson).

7. REFERENCES

[1] Clark, H.H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *J. Verbal Learning & Verbal Beh.*12, 335-359.

[2] Edwards, J., Beckman, M.E., Munson, B. 2004. The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *J. Speech Lang. Hear. Res.* 47, 421-436.

[3] Munson, B., Kurtz, B.A., Windsor, J. 2005. The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *J. Speech Lang. Hear. Res.* 48, 1033-1047.

[4] Quené, H., van den Bergh, H. 2008. Examples of mixed effects modeling with crossed random effects and with binomial data. *J. Mem. Lang.* 59, 413-425.

[5] Storkel, H.L. 2001. Learning new words: Phonotactic probability in language development. *J. Speech Lang. Hear. Res.* 44, 1321-1337.

[6] Vitevitch, M.S., Luce, P.A., Charles-Luce, J., Kemmer, D. 1997. Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Lang. Speech* 40, 47-62.