Example 6b: Subjects Crossed with Items as Crossed Random Effects (complete syntax, data, and output available for R electronically)

This example illustrates multilevel models with crossed random effects for an observed continuous outcome in the context of repeated measures data from Locker, Hoffman, and Bovaird (2007). Response time (*RT*, in ms) for a lexical decision task (in which subjects decide as quickly as they can whether each item is a word or a non-word) was collected for 39 items from 38 subjects (total possible observations = 1482; total actual observations = 1392 after removing RTs for inaccurate responses). Items are words that varied systematically in two features: semantic frequency (*freq01*: 0 = low, 1 = high) and neighborhood size (*size01*: 0 = small, 1 = large). The data are in "long" (or "stacked") format, in which each row contains one trial (per subject per item). The example uses several custom functions to process output and add additional computations, as shown below (all in the online files).

R Syntax for Packages; Importing and Preparing Data for Analysis

```
# Install and load required packages
if (!require(pacman)) install.packages("pacman")
pacman::p_load(readxl, TeachingDemos, lme4, lmerTest, emmeans, ggplot2)
# readxl: import .xls or .xlsx files; TeachingDemos: save output to text file
# lme4/lmerTest: MLMs with crossed random effects (Satterthwaite DDF)
# emmeans: model-implied conditional means and simple slopes (emmSat, emtrends)
# ggplot2: figure of predicted values

# Set working directory (to import and export files to)
setwd("C:/Dropbox/26_EDF9850/Unit6/Example6b")

# Load R functions for this class from R file in folder above working directory
source("../../EDF9850_Functions.R")  # Remove ../ for same folder
source("https://www.lesahoffman.com/EDF9850/EDF9850_Functions.R") # From the web

# Import Example 6b excel data file in sheet "Data"
Example6b = read_excel(path="Example6b_Data.xlsx", sheet="Data")
# Convert to data frame to use for analysis
Example6b = as.data.frame(Example6b)

# Filter to only cases complete on all variables to be used below
Example6b = Example6b[complete.cases(Example6b[ , c("RT","freq01","size01")]), ]

print("Descriptive Statistics for RT across all Trials")
univSummary(data=Example6b, vars="RT")
```

```
 Variable    N    Mean      SD  Variance     Min      Max  Median
 RT       1392 632.383 146.083 21340.288 352.000 1806.000 604.000
```

```
print("Number of Items per Item Predictor Level")
# (!duplicated keeps one row per item, so each item is counted only once)
freqSummary(data=Example6b[!duplicated(Example6b$ItemID), ],
            vars=c("freq01","size01"))
```

```
 Variable Level  N Proportion
 freq01   0     20      0.513
 freq01   1     19      0.487
 size01   0     20      0.513
 size01   1     19      0.487
```

```
# Cross-tabulation of the item predictors (also one row per item)
table(Example6b$freq01[!duplicated(Example6b$ItemID)],
      Example6b$size01[!duplicated(Example6b$ItemID)])
```

```
    0  1
  0 10 10
  1 10  9
```

Model 1: Empty Means, Random Intercept for Subjects Only

In the notation below, the subscripts are t = level-1 trial, i = level-2 item, and s = level-2 subject. This first model contains only a fixed intercept γ_{000} and a random intercept U_{00s} for each subject, so it accounts only for the correlation among trials from the same subject (ignoring that trials are also from the same items). A model with no random effects at all (a single level) is not shown separately, because the LRT below already compares Model 1 to that model.

$$\boldsymbol{RT_{tis} = \gamma_{000} + U_{00s} + e_{tis}}$$

```
print("Model 1: Empty Means, Random Intercept for Subjects Only")
Model1 = lmer(data=Example6b, REML=TRUE, formula=RT~1+(1|SubjectID))
lmerSummary(Model1, ICC1=TRUE, LRT=TRUE, randCI=TRUE)
```

```
Model Fit for Predicting RT:
      AIC       BIC    logLik -2*logLik
17546.265 17561.980 -8770.132 17540.265

Model for the Variance:
 Groups     Name        Variance Std.Dev
 SubjectID  Intercept  5167.096  71.883
 Residual             16307.206 127.700
Number of obs: 1392 , groups: SubjectID, 38

Fixed Effects:
 Term           Est     SE     DF t.value p.value
 Intercept  631.417 12.154 37.004  51.951   <.001

ICC1 intraclass correlation:
    Adjusted ICC: 0.241
  Unadjusted ICC: 0.241

LRT for random effect(s):
 Random Effect Dropped  Npar    LL  -2LL   AIC   BIC     LRT DF p.value p.mix
 <none>                    3 -8770 17540 17546 17562
 Intercept | SubjectID     2 -8910 17821 17825 17835 280.439  1   <.001 <.001
```

ICC1 = .241 is the expected correlation of trials from the same subject (i.e., 24.1% of the variance in RT is due to mean differences between subjects). The LRT for the random intercept variance tests whether trials from the same subject are more alike than trials from different subjects, $-2\Delta LL(1) = 280.44$, $p < .001$. The fixed intercept (631.4) is now the weighted mean of the subject means, so it differs a bit from the original RT mean (632.4), because subjects contributed different numbers of correct trials.

Model 2: Empty Means, Crossed Random Intercepts for Subjects and Items

The random intercept for items, U_{0i0}, creates a separate correlation among trials for the same item (beyond that for the same subject). Crossed random effects are specified by simply including both grouping variables in the model, as in (1|SubjectID)+(1|ItemID). Trials are the unique combination of a subject and an item, so they are nested within both level-2 dimensions.

$$\boldsymbol{RT_{tis} = \gamma_{000} + U_{00s} + U_{0i0} + e_{tis}}$$

```
print("Model 2: Empty Means, Crossed Random Intercepts for Subjects and Items")
Model2 = lmer(data=Example6b, REML=TRUE, formula=RT~1+(1|SubjectID)+(1|ItemID))
lmerSummary(Model2, ICC1=TRUE, LRT=TRUE, randCI=TRUE)
```

```
Model Fit for Predicting RT:
      AIC       BIC    logLik -2*logLik
17447.865 17468.819 -8719.933 17439.865

Model for the Variance:
 Groups     Name        Variance Std.Dev
 ItemID     Intercept   2414.113  49.134
 SubjectID  Intercept   5166.722  71.880
 Residual              14343.043 119.762
Number of obs: 1392 , groups: ItemID, 39; SubjectID, 38

Fixed Effects:
 Term           Est     SE     DF t.value p.value
 Intercept  635.328 14.434 59.435  44.016   <.001

ICC1 intraclass correlation:
    Adjusted ICC: 0.346
  Unadjusted ICC: 0.346

Proportion of Variance by Source (adj = adjusted after fixed effects):
 Source         Variance Prop.Adj Prop.Unadj
 ItemID         2414.113    0.110      0.110
 SubjectID      5166.722    0.236      0.236
 Residual      14343.043    0.654      0.654
 Fixed Effects     0.000                0.000

LRT for random effect(s):
 Random Effect Dropped  Npar    LL  -2LL   AIC   BIC     LRT DF p.value p.mix
 <none>                    4 -8720 17440 17448 17469
 Intercept | SubjectID     3 -8879 17758 17764 17780 317.925  1   <.001 <.001
 Intercept | ItemID        3 -8770 17540 17546 17562 100.399  1   <.001 <.001
```

All sources of variance are orthogonal, so we can sum them to compute the proportion of variance due to each sampling dimension, as shown in the table above: ICC1 for the same subject = .236 (23.6% of the variance is between-subject mean differences), ICC1 for the same item = .110 (11.0% is between-item mean differences), and the remaining 65.4% is trial-specific deviation. ICC1 = .346 is the expected correlation of trials with both the same subject and the same item (as given in the ICC1 table). The subject ICC is similar to that of Model 1 (.241), because the item variance came out of the level-1 residual variance. Each random intercept variance is needed.

```
Random Effect 95% CIs (Fixed Effect +/- 1.96*SQRT(random effect variance)):
 Term                 FixedEffect Variance     SD   Lower   Upper
 Intercept | ItemID       635.328 2414.113 49.134 539.026 731.630
 Intercept | SubjectID    635.328 5166.722 71.880 494.443 776.212
```

The fixed intercept (635.3) is now the weighted mean of the subject means AND the item means. The random effect 95% CIs (fixed intercept ± 1.96*SQRT(random intercept variance)) show the expected range of the intercepts across 95% of subjects and items.

Model 3: Add Fixed Effects of Item Frequency, Item Size, and their Interaction

Next, we add fixed slopes for the item predictors *freq01* and *size01* and their interaction to the crossed model to explain between-item differences in mean RT. The fixed intercept γ_{000} is the predicted RT for a low-frequency, small-size item; γ_{010} is the simple effect of frequency for small-size items; γ020 is the simple effect of size for low-frequency items; and γ_{030} is the interaction (the difference in the frequency slope across size, or equivalently, the difference in the size slope across frequency).

$$RT_{tis} = \gamma_{000} + \gamma_{010}(freq01_i) + \gamma_{020}(size01_i) + \gamma_{030}(freq01_i)(size01_i) + U_{00s} + U_{0i0} + e_{tis}$$

```
print("Model 3: Model 2 + Item Frequency, Item Size, and their Interaction")
Model3 = lmer(data=Example6b, REML=TRUE,
              formula=RT~1+freq01+size01+freq01:size01+(1|SubjectID)+(1|ItemID))
lmerSummary(Model3, LRT=TRUE)
# Shorter code to include an interaction and all its lower-order main effects:
# formula=RT~1+freq01*size01+(1|SubjectID)+(1|ItemID)
```

```
Model Fit for Predicting RT:
      AIC       BIC    logLik -2*logLik
17416.447 17453.117 -8701.224 17402.447

Model for the Variance:
 Groups     Name        Variance Std.Dev
 ItemID     Intercept   1692.411  41.139
 SubjectID  Intercept   5168.480  71.892
 Residual              14341.039 119.754

Overall F-test for All Fixed Slopes:
 Model  NumDF  DenDF F.value p.value
 Model3     3 31.782   5.343   0.004

Fixed Effects:
 Term               Est     SE     DF t.value p.value
 Intercept      615.783 18.575 60.661  33.151   <.001
 freq01          70.021 20.597 32.401   3.400   0.002
 size01           4.435 20.422 31.392   0.217   0.829
 freq01:size01  -72.031 29.378 31.781  -2.452   0.020

LRT for random effect(s):
 Random Effect Dropped  Npar    LL  -2LL   AIC   BIC     LRT DF p.value p.mix
 <none>                    7 -8701 17402 17416 17453
 Intercept | SubjectID     6 -8860 17721 17733 17764 318.202  1   <.001 <.001
 Intercept | ItemID        6 -8733 17466 17478 17510  63.991  1   <.001 <.001
```

The *F*-test shows that the two item features and their interaction together significantly predicted RT, $F(3, 31.8) = 5.34$, $p = .004$. The LRT shows that significant item random intercept variance remains after controlling for the item features, $-2\Delta LL(1) = 63.99$, $p < .001$, so items differ in ways beyond their frequency and size.

```
print("Predicted RT for each Frequency by Size Combination")
# at= is needed because freq01 and size01 are numeric 0/1 predictors, not factors
emmSat(Model3, ~freq01*size01, at=list(freq01=c(0,1), size01=c(0,1)))
```

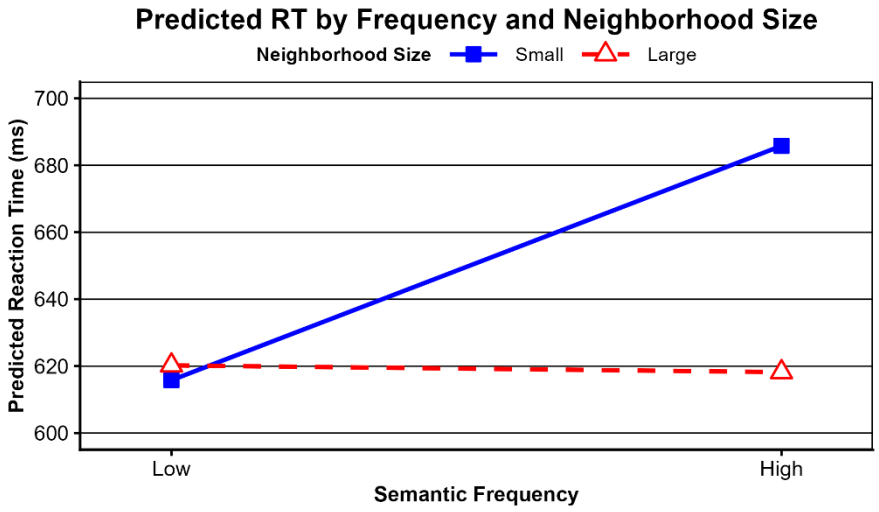
```
 freq01 size01  emmean     SE     df lower.CL upper.CL
 0      0      615.783 18.575 60.661  578.636  652.929
 1      0      685.803 18.742 62.663  648.346  723.261
 0      1      620.217 18.549 60.333  583.118  657.317
 1      1      618.208 19.151 58.826  579.884  656.532
```

```
print("Simple Slopes of Frequency at each Size")
FreqSlopes = emtrends(Model3, ~size01, var="freq01", at=list(size01=c(0,1)),
                      lmer.df="satterthwaite")
summary(FreqSlopes, infer=TRUE) # infer=TRUE adds CIs, t, and p
```

```
 size01 freq01.trend   SE   df lower.CL upper.CL t.ratio p.value
      0        70.02 20.6 32.4     28.1    112.0   3.400  0.0018
      1        -2.01 20.9 31.2    -44.7     40.7  -0.096  0.9242
```

```
print("Simple Slopes of Size at each Frequency")
SizeSlopes = emtrends(Model3, ~freq01, var="size01", at=list(freq01=c(0,1)),
                      lmer.df="satterthwaite"); summary(SizeSlopes, infer=TRUE)
```

```
 freq01 size01.trend   SE   df lower.CL upper.CL t.ratio p.value
      0         4.43 20.4 31.4    -37.2     46.1   0.217  0.8295
      1       -67.60 21.1 32.1   -110.6    -24.6  -3.201  0.0031
```



See R code online for plot for this Figure 1

The frequency effect is positive only for small-size words (blue line) and ~0 for large-size words (the red dashed line).

```
PseudoR2(baseModel=Model2, altModel1=Model3)
```

```
Pseudo-R2
 Term                   Model   Variance Pseudo.R2
 Intercept | ItemID     Model2  2414.113
 Intercept | ItemID     Model3  1692.411     0.299
 Intercept | SubjectID  Model2  5166.722
 Intercept | SubjectID  Model3  5168.480    -0.000
 Residual               Model2 14343.043
 Residual               Model3 14341.039     0.000
```

```
TotalR2(altModel1=Model3)
```

```
Total R2
 Model  Total.R2
 Model3    0.030
```

Given that 11% of the total RT variance was due to item mean differences, explaining 29.9% of it translates to about 3% of all RT variance, which is the total-R^2 given by the fixed effects.

Next we examine whether subjects differ in their effects of the item predictors, which is possible because each subject sees items of every kind. The random slopes below are added one at a time, each with its covariance with the subject random intercept.

Model 4a: Model 3 + Random Item Frequency Slope across Subjects

$$RT_{tis} = \gamma_{000} + \gamma_{010}(freq01_i) + \gamma_{020}(size01_i) + \gamma_{030}(freq01_i)(size01_i) + U_{00s} + U_{01s}(freq01_i) + U_{0i0} + e_{tis}$$

```
print("Model 4a: Model 3 + Random Item Frequency Slope across Subjects")
Model4a = lmer(data=Example6b, REML=TRUE,
               formula=RT~1+freq01+size01+freq01:size01
                          +(1+freq01|SubjectID)+(1|ItemID))
lmerSummary(Model4a, LRT=TRUE, randCI=TRUE)
Model for the Variance:
 Groups     Name        Variance Std.Dev  Corr
 ItemID     Intercept   1700.404  41.236
 SubjectID  Intercept   4266.150  65.316
            freq01       371.707  19.280 0.693
 Residual              14244.399 119.350

LRT for random effect(s):
 Random Effect Dropped  Npar    LL  -2LL   AIC   BIC    LRT DF p.value p.mix
 <none>                    9 -8699 17398 17416 17463
 freq01 | SubjectID        7 -8701 17402 17416 17453  4.874  2   0.087 0.057
 Intercept | ItemID        8 -8731 17462 17478 17520 64.809  1   <.001 <.001

Random Effect 95% CIs (Fixed Effect +/- 1.96*SQRT(random effect variance)):
 Term                  FixedEffect Variance     SD   Lower   Upper
 Intercept | ItemID        615.850 1700.404 41.236 535.027 696.672
 Intercept | SubjectID     615.850 4266.150 65.316 487.831 743.868
 freq01 | SubjectID         69.845  371.707 19.280  32.057 107.633
```

The new random slope U_{01s} allows individual differences in the RT difference between low- and high-frequency words. The fixed freq01 slope is now 69.8 (it was 70.0), and now represents the mean of the subject-specific frequency slopes. The random slope variance is 371.7, correlated r = .69 with the subject random intercepts (subjects slower to low-frequency words had larger frequency effects). However, the model fit did not significantly improve from adding the random slope variance and covariance, $-2\Delta LL(2) = 4.87$, p = .087 (mixture p = .057), so we would (likely) not retain them. The 95% random slope CI (69.8 ± 1.96*SQRT(371.7) = 32.1 to 107.6) conveys the expected variability in the frequency effect across subjects (all positive slopes).

Model 4b: Model 3 + Random Item Size Slope across Subjects

$$RT_{tis} = \gamma_{000} + \gamma_{010}(freq01_i) + \gamma_{020}(size01_i) + \gamma_{030}(freq01_i)(size01_i) + U_{00s} + U_{02s}(size01_i) + U_{0i0} + e_{tis}$$

```
print("Model 4b: Model 3 + Random Item Size Slope across Subjects")
Model4b = lmer(data=Example6b, REML=TRUE,
               formula=RT~1+freq01+size01+freq01:size01
                          +(1+size01|SubjectID)+(1|ItemID))
lmerSummary(Model4b, LRT=TRUE, randCI=TRUE)
```

```
Model for the Variance:
 Groups     Name        Variance Std.Dev  Corr
 ItemID     Intercept  1692.994  41.146
 SubjectID  Intercept  5103.813  71.441
            size01       84.446   9.189 0.067
 Residual             14319.258 119.663

LRT for random effect(s):
 Random Effect Dropped  Npar    LL  -2LL   AIC   BIC    LRT DF p.value p.mix
 <none>                    9 -8701 17402 17420 17468
 size01 | SubjectID        7 -8701 17402 17416 17453  0.082  2   0.960 0.868
 Intercept | ItemID        8 -8733 17466 17482 17524 64.064  1   <.001 <.001

Random Effect 95% CIs (Fixed Effect +/- 1.96*SQRT(random effect variance)):
 Term                  FixedEffect Variance     SD   Lower   Upper
 Intercept | ItemID        615.800 1692.994 41.146 535.153 696.446
 Intercept | SubjectID     615.800 5103.813 71.441 475.775 755.824
 size01 | SubjectID          4.415   84.446  9.189 -13.596  22.427
```

The new random slope U_{02s} allows individual differences in the RT difference between small- and large-size words. The fixed size01 slope is essentially unchanged (4.42), but now represents the mean of the subject-specific size slopes. The model fit decidedly did not improve from adding the random slope variance and covariance, $-2\Delta LL(2) = 0.08$, $p = .960$ (mixture $p = .868$), so we would not retain them. The 95% random slope CI includes both negative and positive size effects.

Sample Results Section starts here

[indicates notes about what to customize or also include; note that SE and p-values are not needed if you provide tables for the model solutions; these are wordy on purpose to help show how to interpret each effect]

The extent to which semantic frequency (coded low = 0, high = 1) and phonological neighborhood size (coded small = 0, large = 1) could predict response time (RT) in milliseconds in a lexical decision task was examined for 39 items administered to 38 subjects. Because RTs for incorrect responses were not included, the data were unbalanced, such that each subject had a different number of trials included for each condition. Accordingly, rather than aggregating the individual trial RTs into potentially biased item condition means (that would assume items are fixed) and conducting an analysis of variance, all possible RTs were examined instead in a multilevel model with crossed random effects for subjects and items (Locker et al., 2007), in which level-1 trials (i.e., the unique combination of each subject with each item) were nested within level-2 subjects and within level-2 items (as crossed random factors). Restricted maximum likelihood within the lmer function in the lme4 R package (v. 2.0.6; Bates et al., 2015) was used to estimate all models; denominator degrees of freedom were estimated with the Satterthwaite method using the lmerTest R package (v. 3.2.1; Kuznetsova et al., 2017), and conditional means and simple slopes were obtained using the emmeans R package (v. 2.0.4; Lenth & Piaskowski, 2026). Likelihood ratio tests (i.e., the difference in model −2LL values) were used to evaluate new random effect variances and covariances, in which a mixture of χ^2 distributions (with the two mixture degrees of freedom given in parentheses below) was used to determine the significance of the new random effect variances bounded at 0.

The extent to which systematic variability in mean RT existed for each dimension of sampling was first examined in a series of empty means models (i.e., only a fixed intercept and no predictors). Adding a random intercept variance for subjects to a model with only a residual variance significantly improved model fit, $-2\Delta LL(0;1) = 280.44$, $p < .001$, indicating significant differences between subjects in mean RT, and that trials from the same subject were positively correlated. The addition of a random intercept for items also significantly improved model fit, $-2\Delta LL(0;1) = 100.40$, $p < .001$, indicating significant differences between items in mean RT as well, and

that trials for the same item were also positively correlated. Of the total estimated RT variance, 24% was due to between-subject differences in mean RT (given by the subject random intercept), 11% was due to between-item differences in mean RT (given by the item random intercept), and the remaining 65% was due to the subject by item interaction (i.e., residual variance). Construction of 95% random intercept confidence intervals revealed that 95% of subject mean RTs are expected to fall between 494 and 776 ms, whereas 95% of the item mean RTs are expected to fall between 539 and 732 ms. Thus, there was relatively more variability across subjects than across items. The extent to which the main and interaction effects of semantic frequency and neighborhood size could account for between-item differences in mean RT was then examined in a conditional model.

ANOVA-like description of the results: There was a significant semantic frequency by neighborhood size interaction, $t(31.8) = -2.45$, $p = .020$; the pattern of the interaction is shown in Figure 1 and was interpreted by examining simple main effects of each predictor. First, with respect to the effect of frequency, for small neighborhood words, responses were significantly faster to words of low than high frequency ($M = 615.78$, $M = 685.80$), $t(32.4) = 3.40$, $p = .002$, whereas for large neighborhood words, there was no significant difference between words of low or high frequency ($M = 620.22$, $M = 618.22$), $t(31.2) = -0.10$, $p = .924$. Second, with respect to the effect of neighborhood size, for low frequency words, there was no significant difference between words with small or large neighborhood size ($M = 615.78$, $M = 620.22$), $t(31.4) = 0.22$, $p = .829$, whereas for high frequency words, responses were significantly slower to words with smaller than larger neighborhoods ($M = 685.80$, $M = 618.21$), $t(32.2) = -3.20$, $p = .003$.

Regression-like description of the same results, which refer to the linear model equation: The fixed intercept for the predicted RT for a word of low frequency and small size was $\gamma_{000} = 615.78$. High frequency words of small size were responded to significantly slower than low frequency words of small size of $\gamma_{010} = 70.02$ ($p = .002$). Small and large size words of low frequency did not differ significantly, $\gamma_{020} = 4.44$ ($p = .829$). However, there was a significant frequency by size interaction of $\gamma_{030} = -72.03$ ($p = .020$), such that relative to the frequency effect for small words of $\gamma_{010} = 70.02$, the frequency effect for large words was significantly less positive by $\gamma_{030} = -72.0$ (yielding a nonsignificant simple effect of frequency for large words of $\gamma_{010} + \gamma_{030} = -2.01$, $p = .924$). Similarly, relative to the size effect for low frequency words of $\gamma_{020} = 4.44$, the size effect for high frequency words was significantly more negative by $\gamma_{030} = -72.0$ (yielding a significant simple effect of size for high frequency words of $\gamma_{020} + \gamma_{030} = -67.60$, $p = .003$). Thus, as shown in Figure 1, a positive frequency effect was found only for words of small size, and a negative size effect was found only for high frequency words.

The effects of frequency and size explained 29.9% of the item intercept variance. Given that 11% of the total RT variance was due to mean differences between items, this translates into a total reduction in all RT variance of about 3%. The extent to which these effects were sufficient to describe all between-item differences in mean RT was then examined by removing the item random intercept variance from the conditional model. The resulting significant decrease in model fit, $-2\Delta LL(0;1) = 63.99$, $p < .001$, suggested that significant differences remain between items after controlling for their primary design features, or that items should not be treated as fixed.

Finally, the potential for individual subject differences in the frequency slope was examined by adding a random subject frequency slope (and its covariance with the subject random intercept). Model fit did not significantly improve, $-2\Delta LL(1,2) = 4.87$, $p = .057$, indicating that each subject does not need their own random deviation from the fixed effect of frequency. Likewise, the potential for individual subject differences in the neighborhood size slope was examined by adding a random subject size slope (and its covariance with the subject random intercept) to the model. Model fit did not significantly improve, $-2\Delta LL(1,2) = 0.08$, $p = .868$, indicating that each subject does not need their own random deviation from the fixed effect of size, either.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26.

Lenth, R. V., & Piaskowski, J. (2026). emmeans: Estimated marginal means, aka least-squares means (R package version 2.0.4) [Computer software]. https://CRAN.R-project.org/package=emmeans

Locker, L., Jr., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods, 39*(4), 723–730.