Linear Mixed Effects Models

Vera Demberg

Saarland University

WS 2021/2022

Today: **What to do if we have repeated measures**

1. Introduction to linear mixed effects modelling

2. Linear mixed-effects models – example

3. Linear mixed-effects models - A final example

(using examples from Chiara Gambi and Bodo Winter)

Table of Contents

Linear mixed effects models

Linear mixed effects models

- are also known as multilevel models in other fields
- have gained in popularity in the last few years, as estimating such models has become possible with increased computational power
- can be thought of as an extension of linear regression
- take into account the way in which the data are structured (repeated measures)

Data strata

Very often, our data (individual observations) are structured in at least two ways:

- we collect multiple observations from each subject
  (because each subject responds to more than one item)
- we collect multiple observations from each item
  (because each item is responded to by more than one subject)
- this means that individual observations are not completely independent from one another, but they are grouped under subjects and under items

The importance of the independence assumption

- Failing to meet conditional independence assumptions (as dictated by the design) is one of the worst statistical errors one can make in confirmatory analysis
- more important than normality, outliers, homogeneity of variance etc.
- ... but is also one of the most poorly understood aspects in the use of LMEMs
- **General rule:** It's the design that determines what test to use, not whether repeated measurements from the same subjects (or items) are actually correlated or not!

The traditional solution

Traditionally, within the ANOVA framework, this problem is handled in the following way:

- we group together all observations for a given subject (corresponding to different items) to compute an average for each subject
- and enter these averages, instead of the raw data, in an ANOVA, which we call by-subject ANOVA (its outcome is usually reported as F_1)
- we group together all observations for a given item (corresponding to different subjects) to compute an average for each item
- and enter these averages, instead of the raw data, in an ANOVA, which we call by-item ANOVA (its outcome is usually reported as F_2)

Averaging by subjects and items

Subject	Item	D.V.	
1	1	345	
1	2	570	
1	3	485	Subj 1
1	4	365	
1	5	444	
2	1	...	
2	2		
2	3		Subj 2
2	4		
2	5		
3	1	...	
3	2		
3	3		Subj 3
3	4		
3	5		

Item 1

Item 2

Fixed and random effects

Instead of averaging and calculating two separate models, linear mixed effects models extend linear regression to include random effect as well as fixed effect parameters (hence, mixed effects)

Fixed effect

Refers to parameters that do not vary by group
(i.e., they are the same for all subjects and all items)

Random effect

Refers to parameters that vary by group
(i.e., they represent adjustments to the fixed effect parameters that are meant to capture differences between subjects and between items).

Understanding random effect structure

- **Toy data set**
- Hypothetical lexical decision experiment examining the effect of "word type" (A vs. B) on reaction times (RT)
- within-subject / between-item manipulation
- 4 subjects $\times$ 4 items $=$ 16 data points

from Barr, D.J., Levy, R., Scheepers., C., & Tily, H.J. (2013). *Random effects structure for confirmatory hypothesis testing: Keep it maximal.* Journal of Memory and Language, 68(3), 255-278.

The data

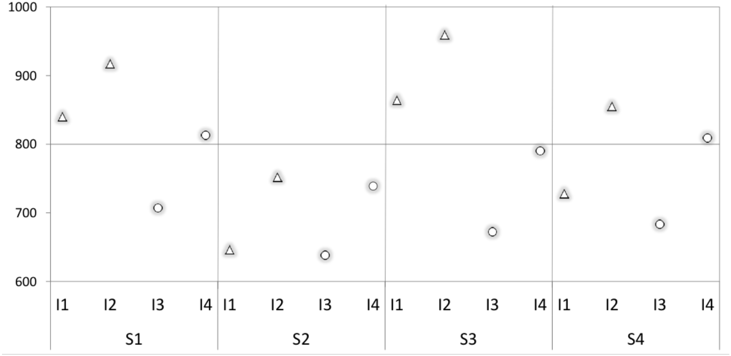


Figure: ◯ word type A; △ word Type B

Simple linear model, no mixed effects

$$Y_{si} = \beta_0 + \beta_1 X_i + e_{si}, \quad e_{si} \sim N(0, \sigma^2)$$

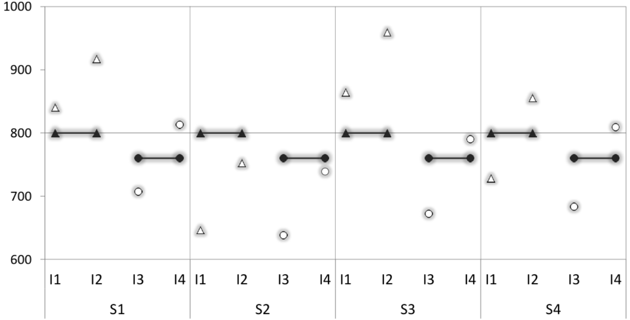


Figure: ◯ word type A; △ word Type B; model predictions in black

In R: `lm(Y ∼ X, data=d)`

Model with by-subject random intercept

$$Y_{si} = \beta_0 + S_{0s} + \beta_1 X_i + e_{si}, \quad S_{0s} \sim N(0, \tau_{00}^2), \quad e_{si} \sim N(0, \sigma^2)$$

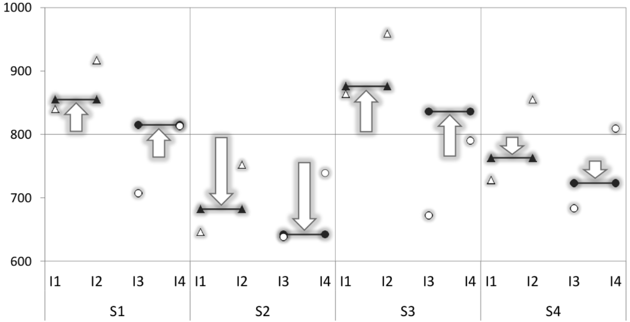


Figure: ◯ word type A; △ word Type B; model predictions in black

In R: `lmer(Y ~ X + (1|Subject), data=d)`

Model with by-subject random intercept and slope

$$Y_{si} = \beta_0 + S_{0s} + (\beta_1 + S_{1s})X_i + e_{si},$$

$$(S_{0s}, S_{1s}) \sim N\left(0, \begin{bmatrix} \tau_{00}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & \tau_{11}^2 \end{bmatrix}\right)$$

$$e_{si} \sim N(0, \sigma^2).$$

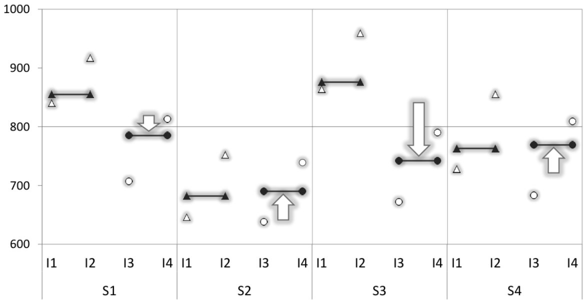


Figure: ◯ word type A; △ word Type B; model predictions in black

In R: `lmer(Y ~ X + (1 + X |Subject), data=d)`

without correlation between random intercept and slope

$Y_{si} = \beta_0 + S_{0s} + (\beta_1 + S_{1s})X_i + e_{si}$,
$S_{0s} \sim N(0, \tau_{00}^2)$,
$S_{1s} \sim N(0, \tau_{11}^2)$,
$e_{si} \sim N(0, \sigma^2)$.

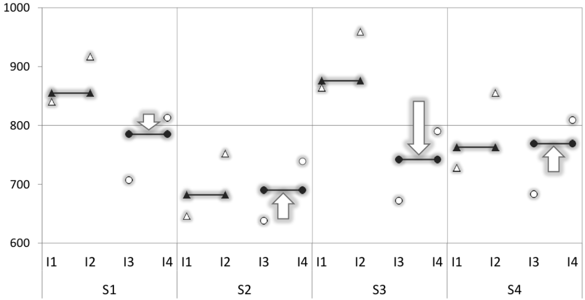


Figure: ○ word type A; △ word Type B; model predictions in black

In R: `lmer(Y ~ X + (1 + X || Subject), data=d)`
or: `lmer(Y ~ X + (1 | Subject) + (0 + X | Subject), data=d)`

Model with by-item random intercept

$$Y_{si} = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s})X_i + e_{si},$$

$$(S_{0s}, S_{1s}) \sim N\left(0, \begin{bmatrix} \tau_{00}^2 & \rho\tau_{00}\tau_{11} \\ \rho\tau_{00}\tau_{11} & \tau_{11}^2 \end{bmatrix}\right),$$

$$I_{0i} \sim N(0, \omega_{00}^2),$$

$$e_{si} \sim N(0, \sigma^2).$$

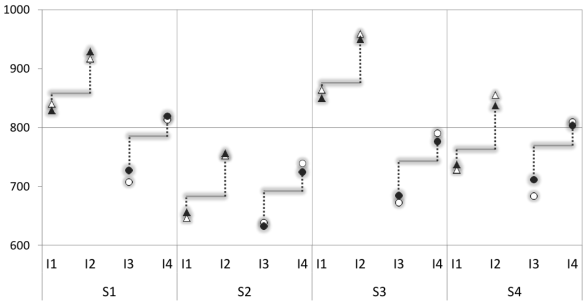


Figure: ◯ word type A; △ word Type B; model predictions in black

In R: `lmer(Y ~ X + (1 + X |Subject) + (1|Item), data=d)`

By subject/item intercepts or also slopes?

- You should not use models with only random intercept but no random slopes.
- This increases the power while inflating Type I error risk!
- Without random slopes, you're not properly taking into account the dependence between data points.

Mixed effects models

- Individual observations are the unit of analysis (each row in the figure)
- Subjects and items are treated as crossed random effects
- In a random intercept model, this means that we take into account intrinsic "baseline" differences between different items and/or subjects

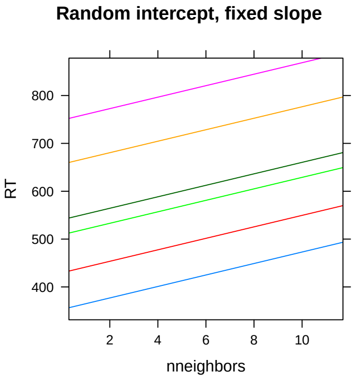


Figure: Visualization from Gelman and Hill

Mixed effects models

- In a random intercept, random slope model: we take into account intrinsic "baseline" differences between different items and/or subjects
- and in addition, we also take into account differences between subjects and/or items in the way they are affected by the D.V. of interest

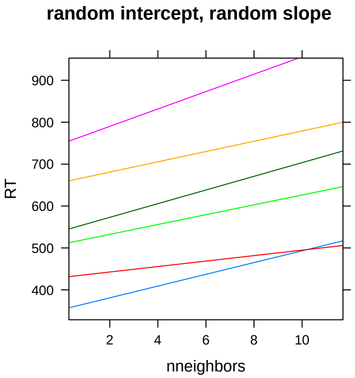


Figure: Visualization from Gelman and Hill

Advantages of the new approach over regression without random effects

- This approach is sometimes called a *partial pooling* approach (Gelman and Hill, 2007)
- it strikes a balance between ignoring the variation between groups (subjects, items) and fitting a separate model for each group
- it allows to define a model at the observation level and at the group level simultaneously
- NOTE: only predictors that vary within subjects and/or items can be entered in the group-level model (i.e., we can only have random slopes for those D.V. that have repeated observations for subjects and/or items)

Advantages of the new approach over ANOVA

- it is possible to generalize over subjects and items simultaneously
- it is possible to include observation level predictors into the model at the same time as group level predictors
- it is possible to analyse imbalanced datasets (that have very different number of observations per subject or item)
- it is possible to analyse continuous and binary outcome variables within the same framework

Keeping it maximal

Let me re-emphasize:

- Choice of random effect structure for confirmatory analysis should be design-driven
- A maximal model encodes the dependencies that are likely to be created by the sampling method itself (e.g., repeated measures over sampling units)
- Assumption: Variability is the norm
  - Different subjects (items) are differentially sensitive to (differentially able to create) the experimental manipulation of interest
  - No effect is constant across subjects or items!
  - Random slopes capture precisely that variability

Table of Contents

A first example from Bodo Winter

We have a dataset of pitch of voices for different people, and we know whether people were male of female. We want to see whether attitude (polite vs. casual) and gender have an effect on pitch.

`pitch` $\sim$ `gender + attitude +` ϵ

A study investigating this question would typically collect data from a number of participants such that each person would have speech recorded in a few polite and a few casual settings. That is, we would have repeated measures by subject and by item.

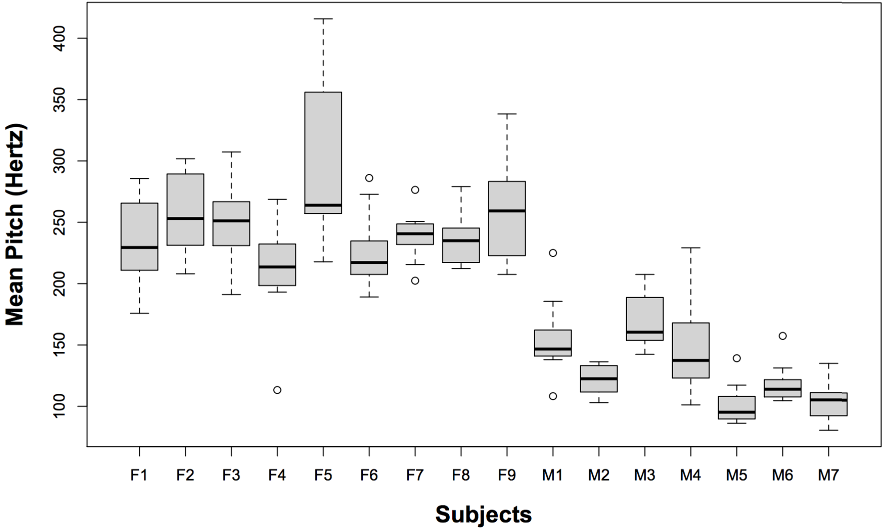
Example

16 participants, who each speak in 7 different scenarios, once in a polite and once in a casual setting (i.e., we have 14 samples for each person)

Problem with repeated measures

With our repeated measures, we **violate the assumptions** of linear regression models where all **observations should be independent**.
Our observations which come from the same subject or are in the same scenario are not independent of one another!
For instance, each person has a different voice and hence also may have different pitch.

Let's take a look at variation between subjects



Random intercept for each subject

Quite obviously, males have lower pitch than female speakers.
We can model further individual differences by assuming different random intercepts for each subject. That is, each subject is assigned a different intercept value, and the mixed model estimates these intercepts for you.

Random intercept for each subject

Quite obviously, males have lower pitch than female speakers.
We can model further individual differences by assuming different random intercepts for each subject. That is, each subject is assigned a different intercept value, and the mixed model estimates these intercepts for you.

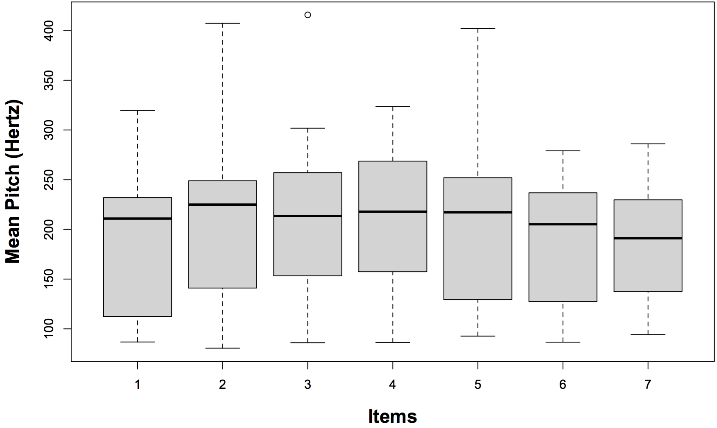
in R, we can specify this as follows:

```
pitch ~ gender + attitude + (1|subject)
```

We can't do this with the standard `lm()` function, but need to use `lmer()` from the `lme4` package instead.

variation by item

Just like people have different voices, the scenarios in which they were asked to speak may lead to different pitch (e.g. because one scenario is more embarrassing than another one).



Model with crossed random effects (on smaller dataset with 6 data points)

```
lmer(pitch ~ attitude + gender + (1|subject) + (1|scenario),
data=politeness)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept) 205.2    14.33
 subject  (Intercept) 417.0    20.42
 Residual             637.4    25.25
```

Model with crossed random effects (on smaller dataset with 6 data points)

```
lmer(pitch ~ attitude + gender + (1|subject) + (1|scenario),
data=politeness)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept) 205.2    14.33
 subject  (Intercept) 417.0    20.42
 Residual             637.4    25.25

Fixed effects:
            Estimate Std. Error t value
(Intercept)  256.847     13.827  18.576
attitudepol  -19.722      5.547  -3.555
genderM     -108.517     17.572  -6.176
```

Comparing alternative models

Question: Does a predictor (here: attitude) improve model fit?

Comparing alternative models

Question: Does a predictor (here: attitude) improve model fit?

```
anova(politeness.null,politeness.model)
```

This is the resulting output:

```
Data: politeness
Models:
politeness.null: frequency ~ gender + (1 | subject) + (1 | scenario)
politeness.model: frequency ~ attitude + gender + (1 | subject) + (1 | scenario)
                 Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
politeness.null   5 816.72 828.81 -403.36   806.72
politeness.model  6 807.10 821.61 -397.55   795.10 11.618      1  0.0006532 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Let's have a look at the coefficients of the model by subject and by item:

```
coef(politeness.model)
```

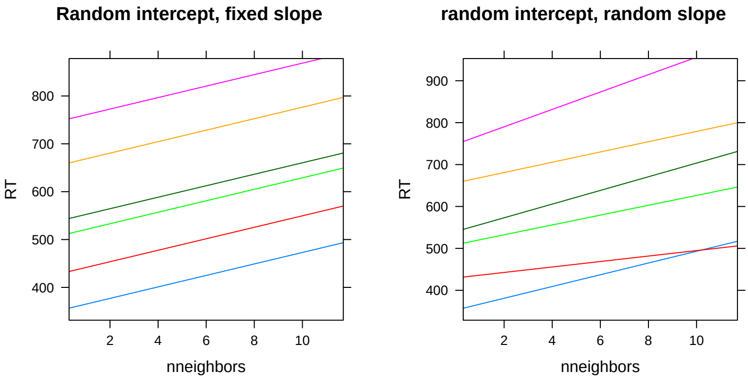
Here is the output:

```
$scenario
  (Intercept) attitudepol    genderM
1    243.4859   -19.72207 -108.5173
2    263.3592   -19.72207 -108.5173
3    268.1322   -19.72207 -108.5173
4    277.2546   -19.72207 -108.5173
5    254.9319   -19.72207 -108.5173
6    244.8015   -19.72207 -108.5173
7    245.9618   -19.72207 -108.5173

$subject
   (Intercept) attitudepol    genderM
F1    243.3684   -19.72207 -108.5173
F2    266.9443   -19.72207 -108.5173
F3    260.2276   -19.72207 -108.5173
M3    284.3536   -19.72207 -108.5173
M4    262.0575   -19.72207 -108.5173
M7    224.1292   -19.72207 -108.5173

attr(,"class")
[1] "coef.mer"
```

So far, we only have random intercepts, but not yet random slopes.



Random slopes for attitude under subject allow each subject to change their pitch to different degrees in polite vs. casual scenarios.

```
politeness.model = lmer(frequency ∼ attitude + gender +
(1+attitude|subject) + (1+attitude|scenario),
data=politeness, REML=FALSE)
```

```
coef(politeness.model)
```

Here's a reprint of the output that I got:

```
$scenario
  (Intercept) attitudepol   genderM
1    245.2603   -20.43832 -110.8021
2    263.3012   -15.94386 -110.8021
3    269.1432   -20.63361 -110.8021
4    276.8309   -16.30132 -110.8021
5    256.0579   -19.40575 -110.8021
6    246.8605   -21.94816 -110.8021
7    248.4702   -23.55752 -110.8021

$subject
   (Intercept) attitudepol   genderM
F1    243.8053   -20.68245 -110.8021
F2    266.7321   -19.17028 -110.8021
F3    260.1484   -19.60452 -110.8021
M3    285.6958   -17.91950 -110.8021
M4    264.1982   -19.33741 -110.8021
M7    227.3551   -21.76744 -110.8021

attr(,"class")
```

Writing up the results (1)

- You need to describe the model to such an extent that people can reproduce the analysis.
- Another important thing is to give enough credit to the people who put so much of their free time into making lme4 and R work so efficiently.
- Write down which version of R you used: `citation()` to get the R package version
  `citation("lme4")` to get the lme4 version

Writing up results (2)

"We used R (R Core Team, 2012) and *lme4* (Bates, Maechler & Bolker, 2012) to perform a linear mixed effects analysis of the relationship between pitch and politeness. As fixed effects, we entered politeness and gender (without interaction term) into the model. As random effects, we had intercepts for subjects and items, as well as by-subject and by-item random slopes for the effect of politeness. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question."

Writing up results (3)

- Many (published!!) articles just state
  - *...we used a linear mixed effects model with random effects for subjects and items...*
  - Not sufficient!
- Report the final random effect structure and how you got there
  - If a random slope required by the design had been dropped for whatever reason, explain to non-experts what that means, e.g. *...by dropping the subject random slope for effect X, the model assumes that there is absolutely no subject-specific variation in the size and direction of effect X in the population...*

Table of Contents

The data

- for this example, we will use the dataset `lexdec`, which is part of the package `languageR`
- it contains lexical decision times of 21 subjects for a set of 79 words (items); this will be our continuous outcome variable

The research question

We want to know whether **lexical decision times** are different for low vs. high frequency words and for native vs. non-native speakers of English

`RT ~ Frequency * NativeLanguage`

Within and between

- first of all, notice that Frequency and Native Language are both categorical predictor variables, and they both have two levels
- however, they differ in that Frequency varies *within* subjects but *between* items,
- whereas Native Language varies *within* items but *between* subjects
- this will be important when we define the random effect structure!

Data cleaning and transformations

- The lexical decision times have been transformed into the logarithmic scale, so we have log(decision times)
- The reason for this is that they were not normally distributed! Remember that this is a fundamental assumption of linear regression models (with gaussian link function)
- In addition, all incorrect responses have been removed, as well as all decision times that are very long (over 7 on the logarithmic scale) and therefore likely to be outliers (due to fatigue, distraction, etc.)

Means

Frequency	English	Other	Tot
High	6.28	6.38	6.32
Low	6.35	6.50	6.41
Tot	6.32	6.44	**6.37**

- Lexical decision times seem to be higher for low frequency words
- Also decision times are higher for non-native speakers.
- But are these differences statistically reliable?

How to fit linear mixed effects models in R

Again,

- we use the function `lmer()`, which is part of the package `lme4`
- the "1" represents the intercept
- then fixed effects are listed
- and then random intercepts by subjects and words

Now let's get to our question ...

Qs: are lexical decision times affected by word frequency and by the language status (native vs. non-native) of the responder?

- to answer this question, we fit a model with predictors Frequency, Native Language and their interaction as fixed effects
- we also include random intercepts for both words and subjects
- and we include the following random slopes:
  - by-subject random slopes for Frequency
  - by-word random slopes for Native Language
- this way we have a "maximal random structure" (Barr et al., 2013)

```
model = lmer(RT ~ 1 +Freq.cat*Native.Language+
(1+Freq.cat|Subject) + (1+Native.Language|Word),
data=lexdec)
```

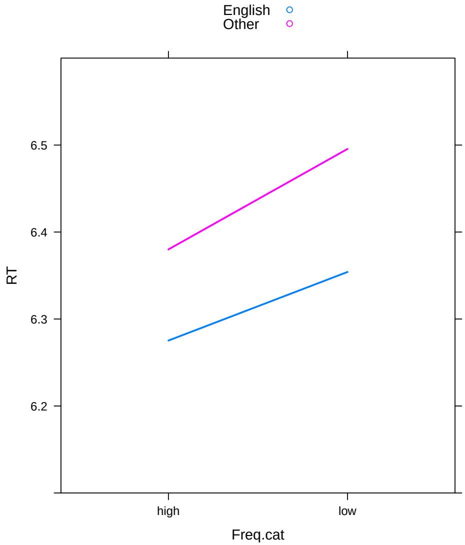
Summary of fixed effects

```
model = lmer(RT ~ 1 +Freq.cat*Native.Language+
(1+Freq.cat|Subject) + (1+Native.Language|Word),
data=lexdec)
```

Fixed effects:

Coef.	Estimate	se	t
Intercept (m_{TOT})	6.37	0.03	232.00
Freq.catLow	0.10	0.01	7.36
NativeLanguageOther	0.14	0.05	2.51
Freq.cat*NativeLanguage	0.05	0.02	2.68

Visualizing the interaction



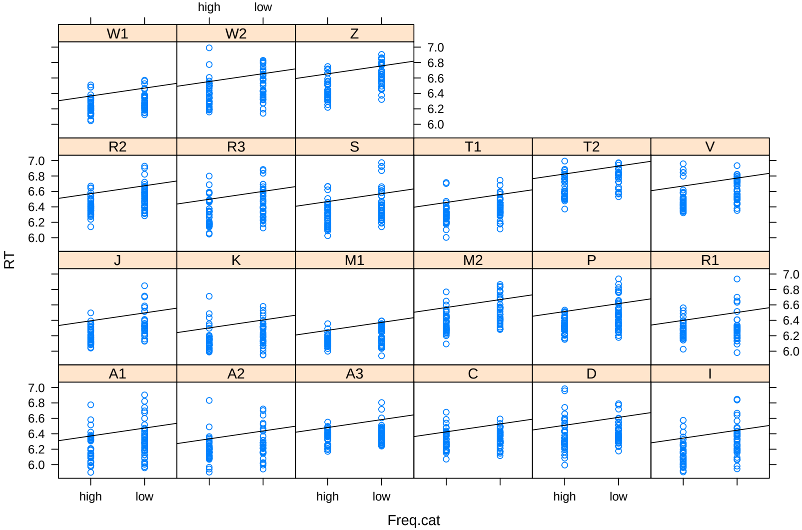
Summary of fixed effects

```
model = lmer(RT ~ 1 +Freq.cat*Native.Language+
(1+Freq.cat|Subject) + (1+Native.Language|Word),
data=lexdec, family="gaussian")
```

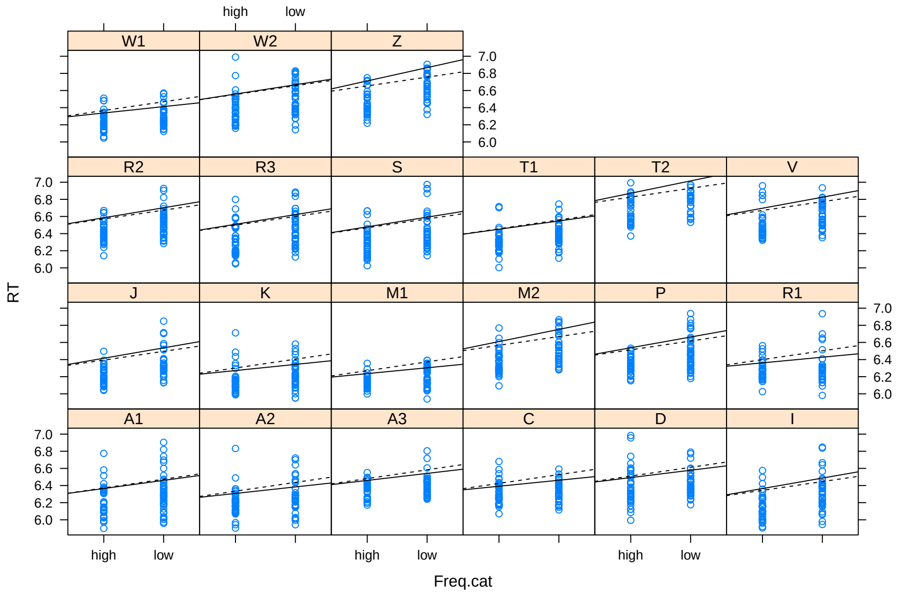
Fixed effects:

Coef.	Estimate	se	t
Intercept (m_{TOT})	6.37	0.03	232.00
Freq.catLow	0.10	0.01	7.36
NativeLanguageOther	0.14	0.05	2.51
Freq.cat*NativeLanguage	0.05	0.02	2.68

Let's take a look at the random intercepts and slopes for subjects.



By-subject intercepts. You could create a similar graph for words, but there's too many of them to display on a single slide


W1
W2
Z
R2
R3
S
T1
T2
V
J
K
M1
M2
P
R1
A1
A2
A3
C
D
I
high
low
RT
Freq.cat
7.0
6.8
6.6
6.4
6.2
6.0

Further discussion

- how to decide between what is a fixed vs. a random factor in the model
- nested random effects

Summary

- Understand the observational dependencies created by your design
- Effects (if they exist) inevitably vary across sampling units
- Random slopes account for that variability in repeated-measures designs with more than one observation per condition
- Be suspicious of any method (parametric or non-parametric) that uses the same model for different sampling designs
- For confirmatory analysis, it's better to "keep it maximal" (Barr et al., 2011)