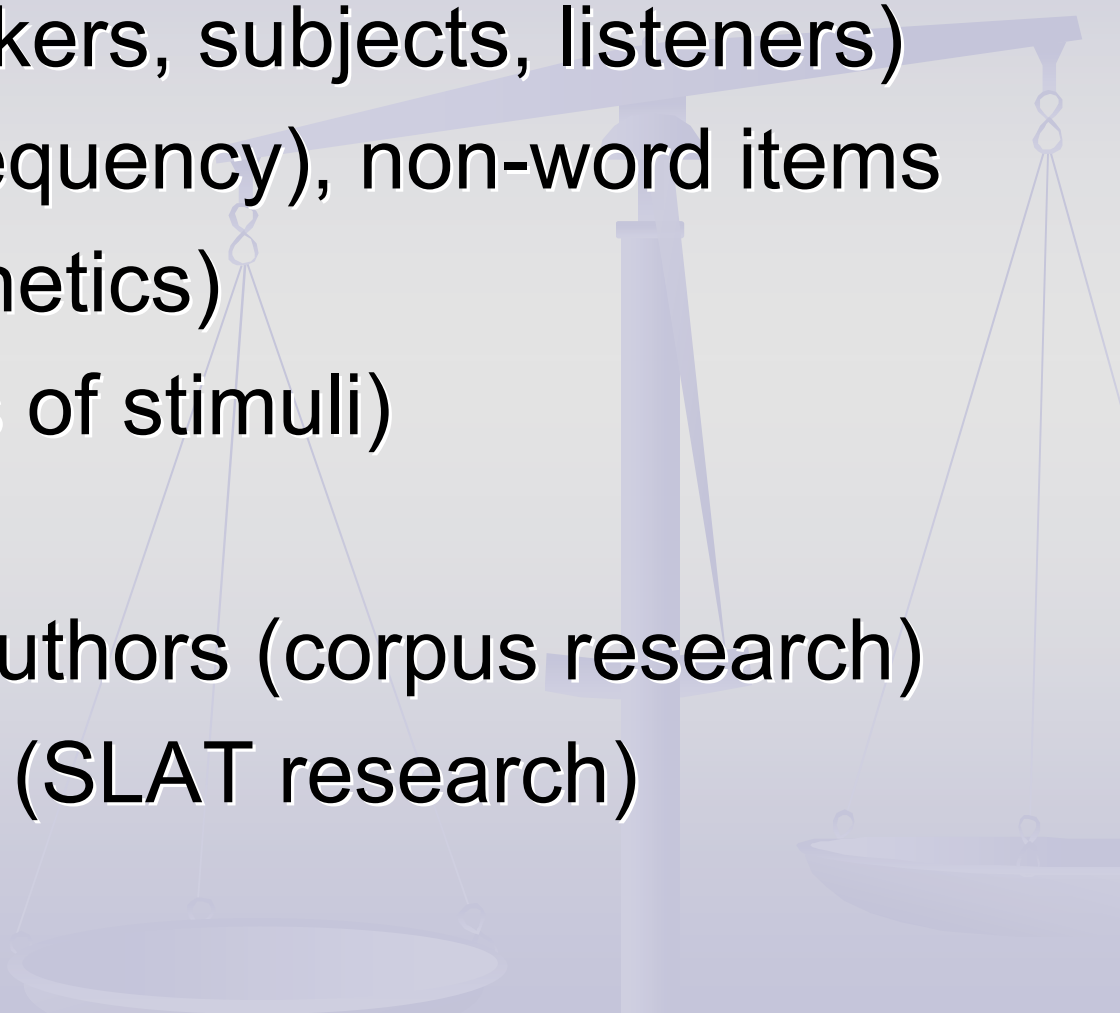


Sources of variability in linguistic data:



Methods for analyzing
random factors

What causes variability in linguistic data?

- Individuals (speakers, subjects, listeners)
 - Words (lexical frequency), non-word items
 - Repetitions (phonetics)
 - Voices (speakers of stimuli)
 - Languages
 - Newspapers or authors (corpus research)
 - Classes, schools (SLAT research)
- 

Random vs. fixed factors

- Random factors: some things selected randomly from a larger population
- Different from fixed factors (e.g. gender, place of articulation, scope type, ...)
- ANOVA and multiple regression are not made to handle random factors, at least not more than one of them
- Why not? Autocorrelation. Clusters of related variability.

Problem: how many do we have?

- Phonetics, psycholinguistics: usually at least 2 random factors (subjects, items)
- May have more: voices producing stimuli, repetitions, counterbalanced group/order
- Various random factors occur in other kinds of ling. research (e.g. document or newspaper in corpus research, conversation in discourse analysis)

What's being done about this? 1

- One possibility: **ignore the problem!**
- Treat each data point (each repetition of each item by each speaker) as a separate, independent data point
- Run ANOVA as if those were separate subjects
- Not good: artificially inflates likelihood of getting a significant result, by a **LOT**.
- Somewhat common in phonetics, at least in talks.

Toy example: Quené

- Quené & van den Bergh (2004) give a hypothetical dataset with 12 speakers and 3 repetitions in each condition.
- Analyzed as if each repetition were a subject (disaggregated):

$$F(2, 105) = 5.15, p = 0.007$$

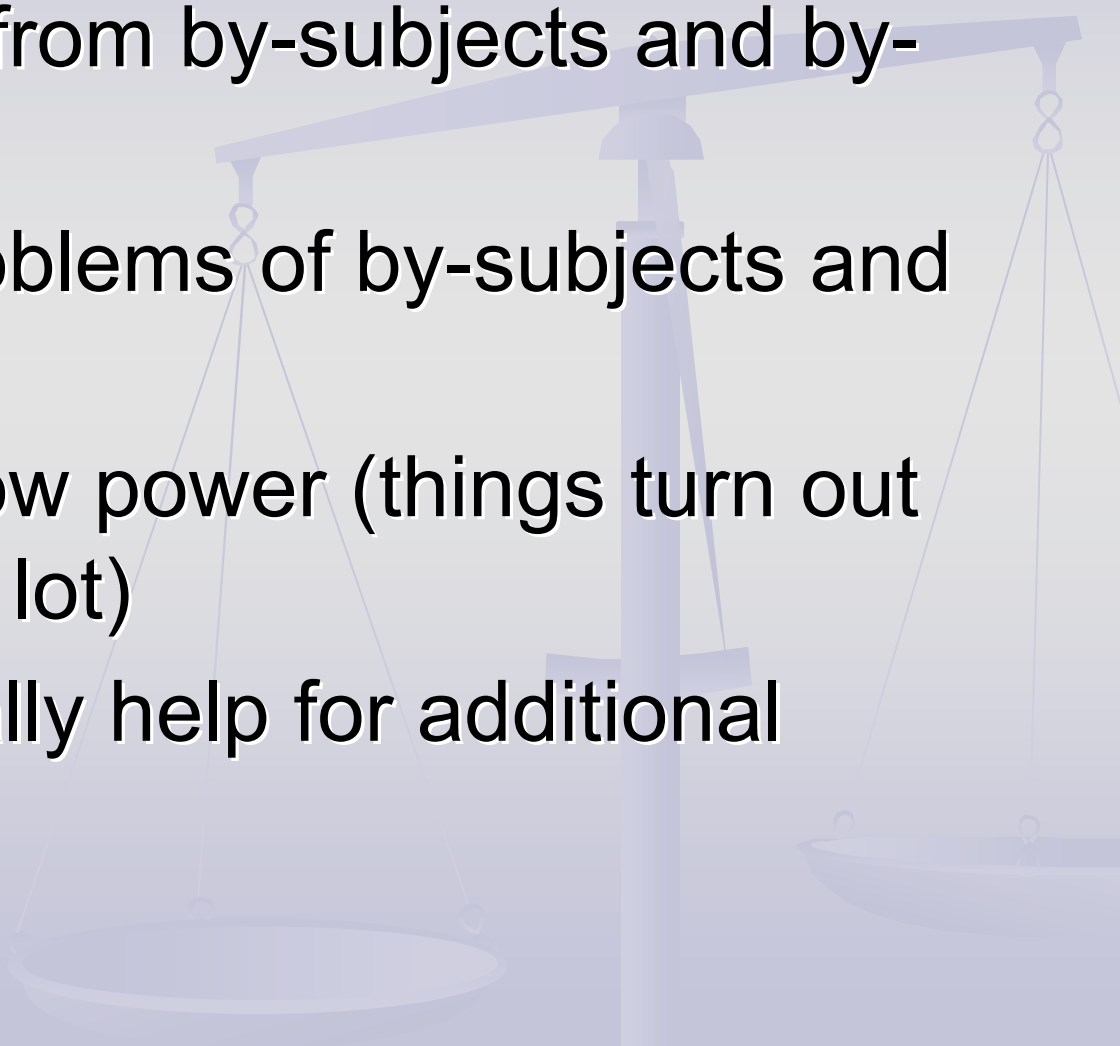
2: By-subjects and by-items

- Average over items and do ANOVA with subjects as the unit of measurement. Then average over subjects and do the same with items. Must be significant on both.
- Standard in psycholinguistics.
- There are problems with this method (see Forster's work), but it's standard.
- Phonetics often does just by-subjects.
- What about other random factors????

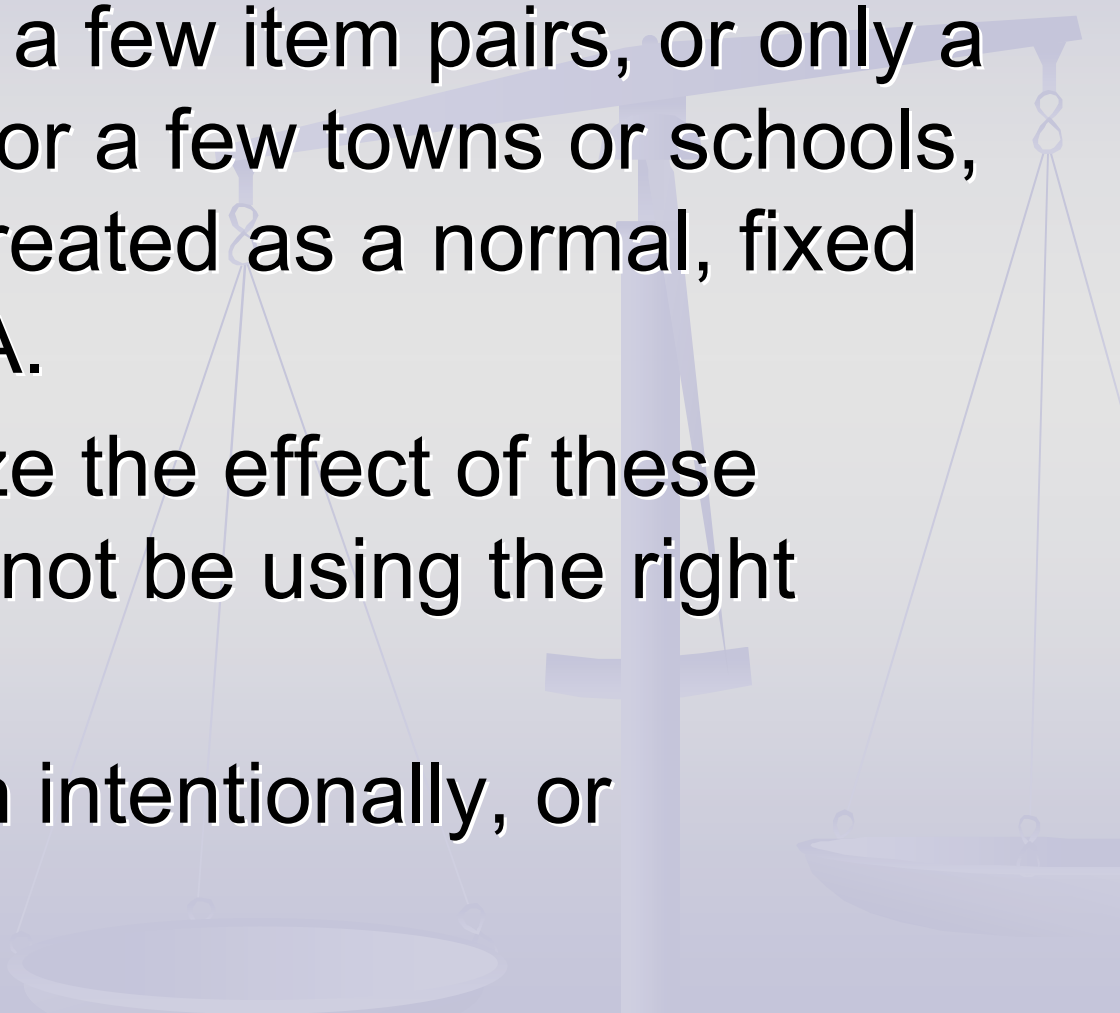
Same example

- Same data analyzed as by-subjects (averaged over repetitions).
- $F(2,22)=3.58, p=0.045$
- But there's a sphericity violation, when accounted for, $F(1.25,13.75)=3.58, p=.073$

3: F'

- F' is calculated from by-subjects and by-items ANOVA
 - Corrects the problems of by-subjects and by-items
 - But may have low power (things turn out not significant a lot)
 - This doesn't really help for additional random factors
- 

4: Treat random factors as fixed

- If there are only a few item pairs, or only a few languages, or a few towns or schools, these may get treated as a normal, fixed factor in ANOVA.
 - Let's you analyze the effect of these things, but may not be using the right math.
 - Are they chosen intentionally, or randomly?
- 

Same example

- Treatment and repetition number both as fixed factors

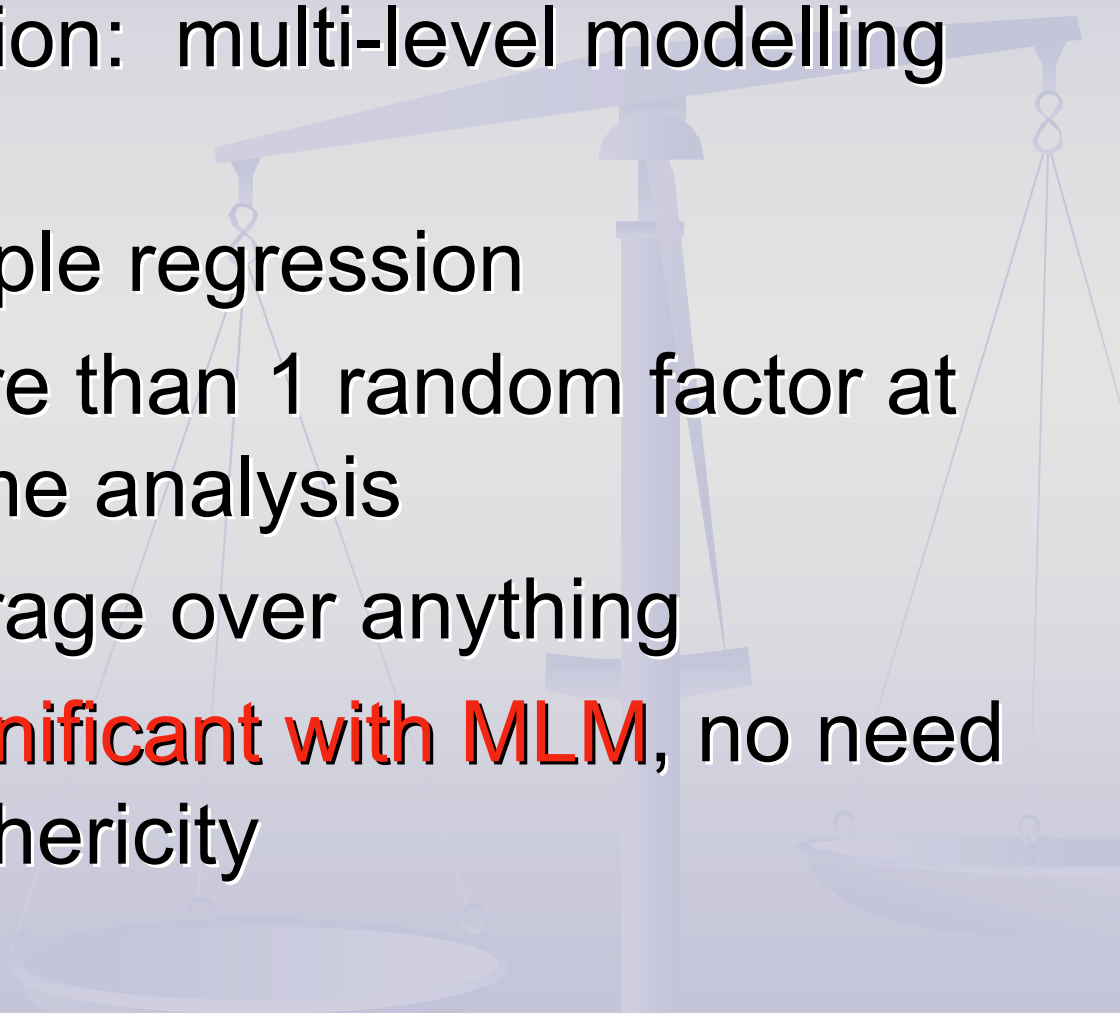
Treatment: $F(2,22)=6.67, p=0.005$

Repetition: $F<1$

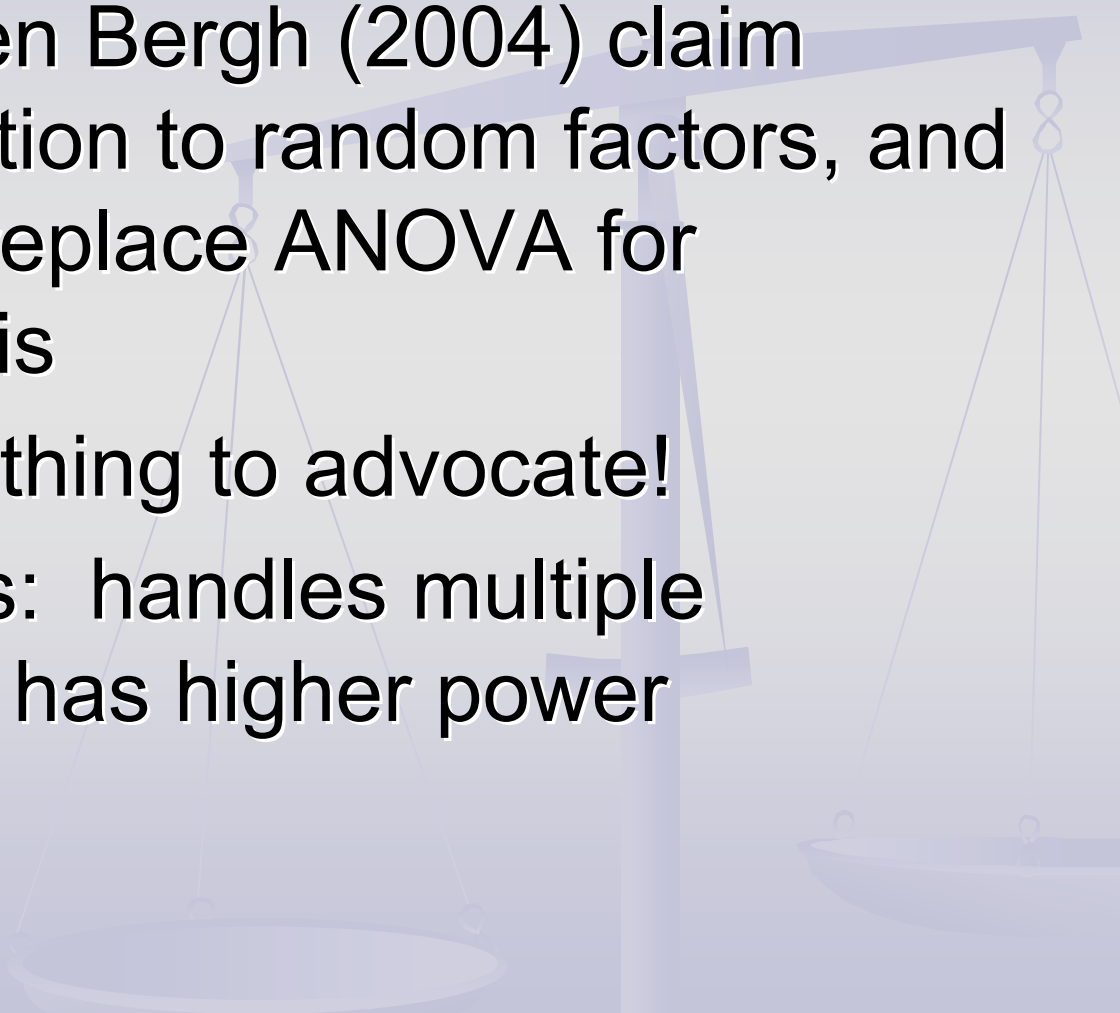
Interaction: $F(4,44)=3.47, p=0.015$

But what are you going to do with that interaction??

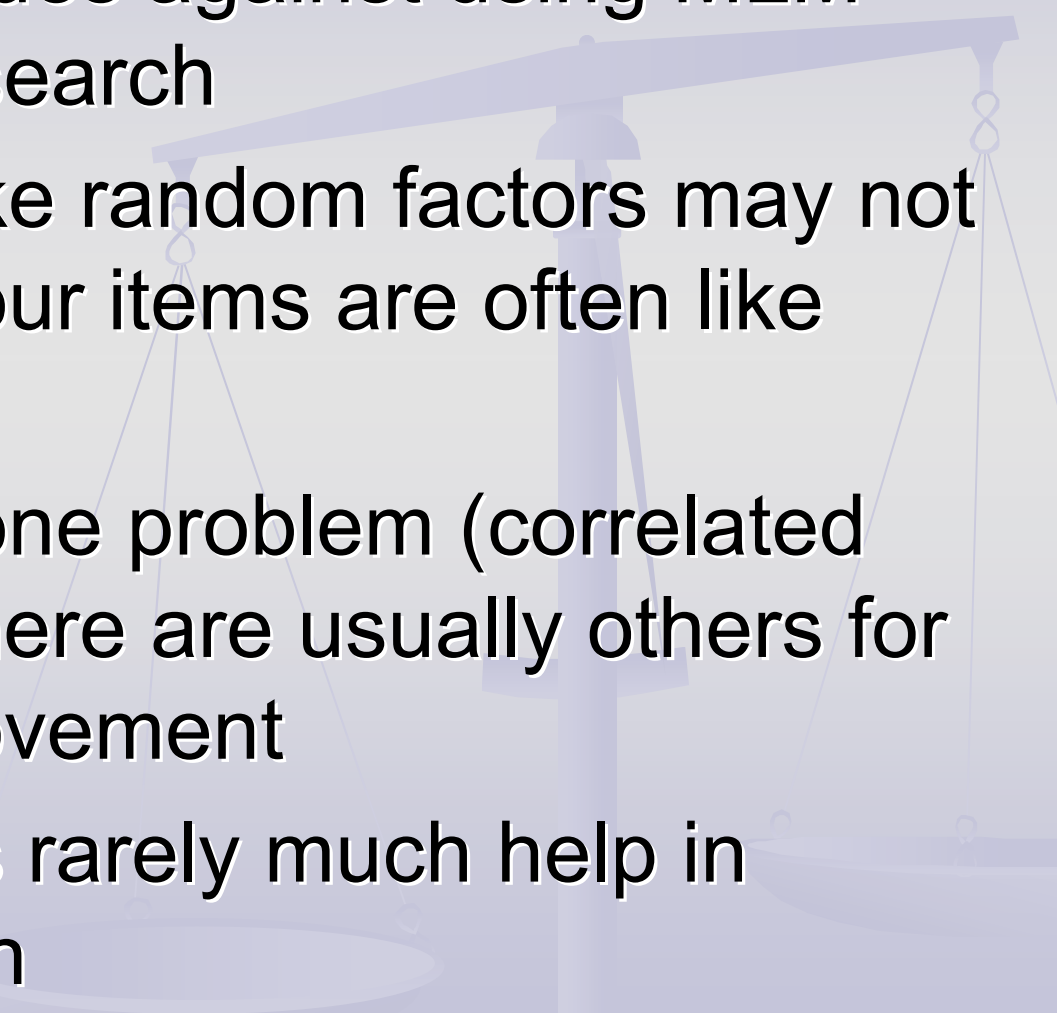
4: Multi-level modelling

- Recent suggestion: multi-level modelling (MLM)
 - Related to multiple regression
 - Can handle more than 1 random factor at once, in the same analysis
 - No need to average over anything
 - Same data: **significant with MLM**, no need to correct for sphericity
- 

Claims in favor of MLM

- Quené & van den Bergh (2004) claim MLM is the solution to random factors, and should entirely replace ANOVA for linguistic analysis
 - That's a drastic thing to advocate!
 - Their arguments: handles multiple random factors, has higher power
- 

Is MLM the solution?

- Gorard (2003) argues against using MLM for educational research
 - Things that look like random factors may not always really be (our items are often like this)
 - MLM only solves one problem (correlated clumps of data), there are usually others for which it's no improvement
 - Concludes MLM is rarely much help in education research
- 

What should we do?

- Think about sources of variability in language data
- Think about which are fixed and which random, and what information we can get from them
- Keep an eye on methods for dealing with random factors
- Let's not dispose of ANOVA quite yet!