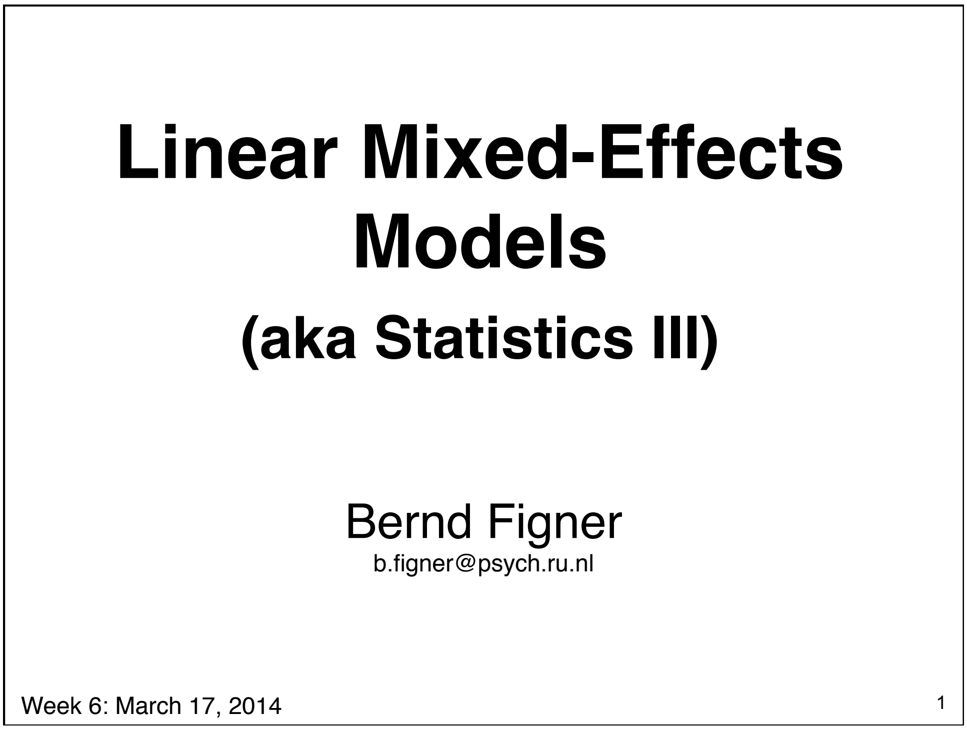


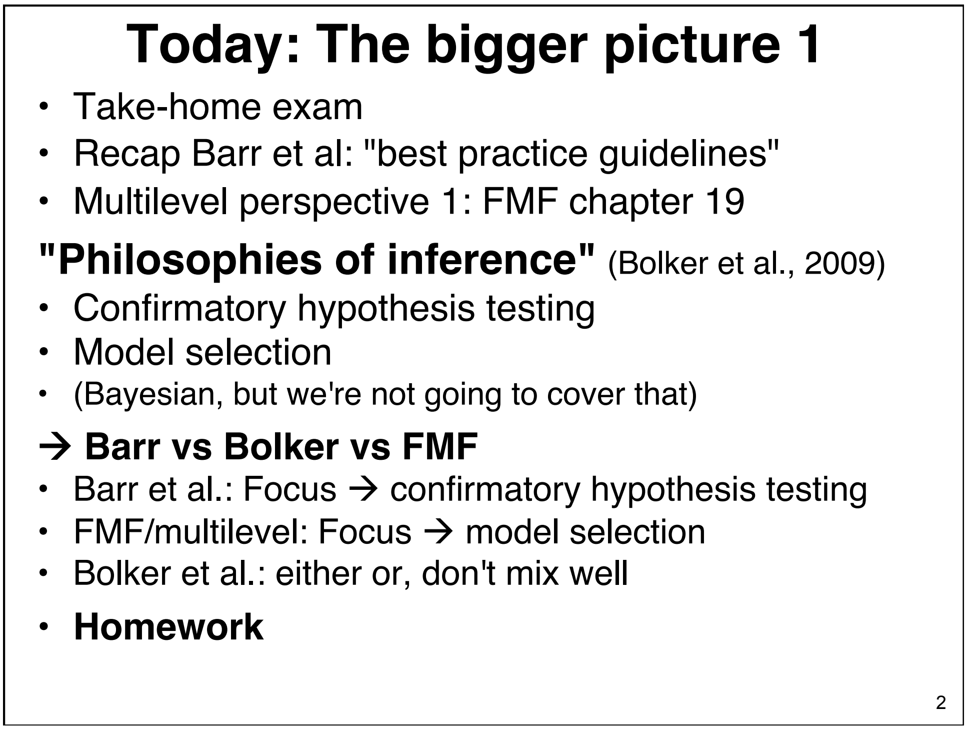
Linear Mixed-Effects Models

(aka Statistics III)

Bernd Figner

b.figner@psych.ru.nl

Week 6: March 17, 2014



Today: The bigger picture 1

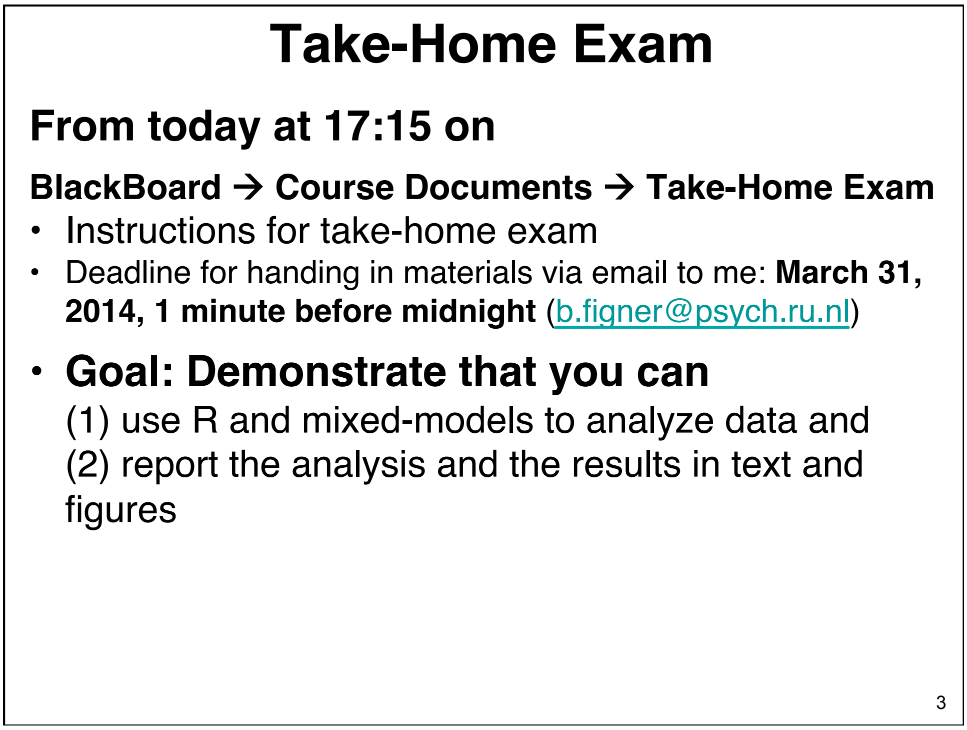
- Take-home exam
- Recap Barr et al: "best practice guidelines"
- Multilevel perspective 1: FMF chapter 19

"Philosophies of inference" (Bolker et al., 2009)

- Confirmatory hypothesis testing
- Model selection
- (Bayesian, but we're not going to cover that)

→ Barr vs Bolker vs FMF

- Barr et al.: Focus → confirmatory hypothesis testing
- FMF/multilevel: Focus → model selection
- Bolker et al.: either or, don't mix well
- **Homework**

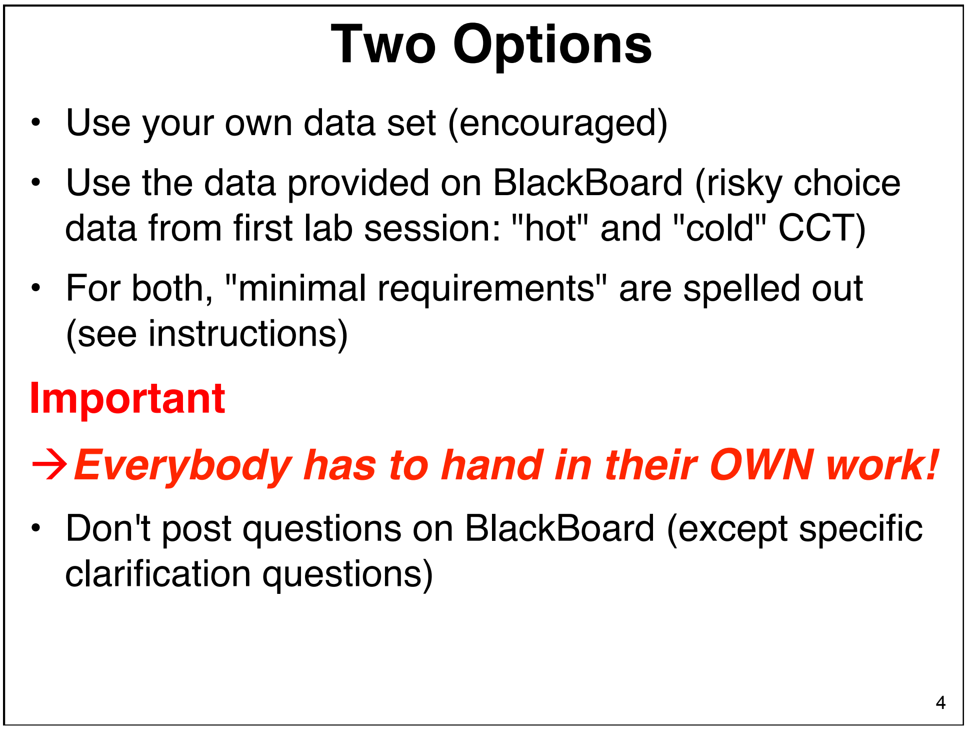


Take-Home Exam

From today at 17:15 on

BlackBoard → Course Documents → Take-Home Exam

- Instructions for take-home exam
- Deadline for handing in materials via email to me: **March 31, 2014, 1 minute before midnight** (b.figner@psych.ru.nl)
- **Goal: Demonstrate that you can**
  (1) use R and mixed-models to analyze data and
  (2) report the analysis and the results in text and figures



Two Options

- Use your own data set (encouraged)
- Use the data provided on BlackBoard (risky choice data from first lab session: "hot" and "cold" CCT)
- For both, "minimal requirements" are spelled out (see instructions)

Important

→Everybody has to hand in their OWN work!

- Don't post questions on BlackBoard (except specific clarification questions)

What you have to hand in

Word or pdf document

- Like **results** section of a journal article
- Your own data: brief description of study/measures
- Describe whole model set-up and process, the results, and briefly describe/interpret the results
- Include figures (including figure captions)
- Max. 6 A4 pages (not including title page and references)

The used data set (in csv format)

The used R script

- Must run without adjusting anything except `setwd()`
- Sufficiently commented so that others can understand what‘s being done

Questions regarding

Take-Home Exam?

Best Practice Guidelines

Barr et al. (2013) and Barr (2013)

plus some additional advice
from Dr. Bolker and yours truly

Guidelines for confirmatory hpothesis testing
("best practices" from Barr et al., 2013)

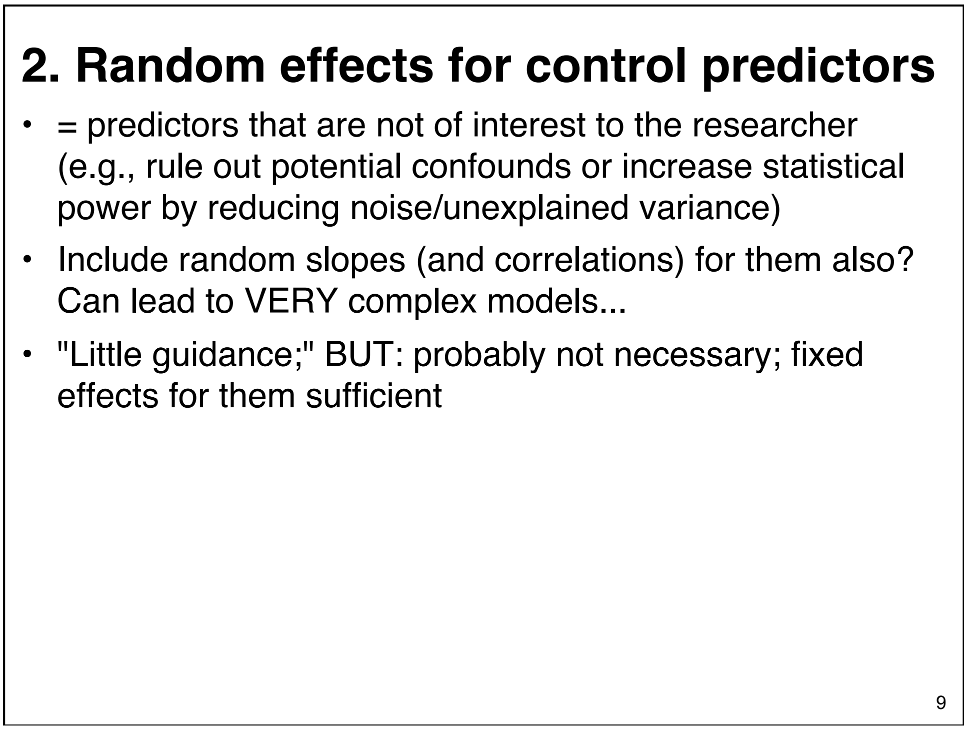
1. Identify the max random effects structure

- Which predictors are between, which within "unit?"
- Within → random slope in addition to fixed slope
- Applies also to interactions
- Also: include all possible random covariance terms

Exception: if there is only 1 observation per "cell" per unit

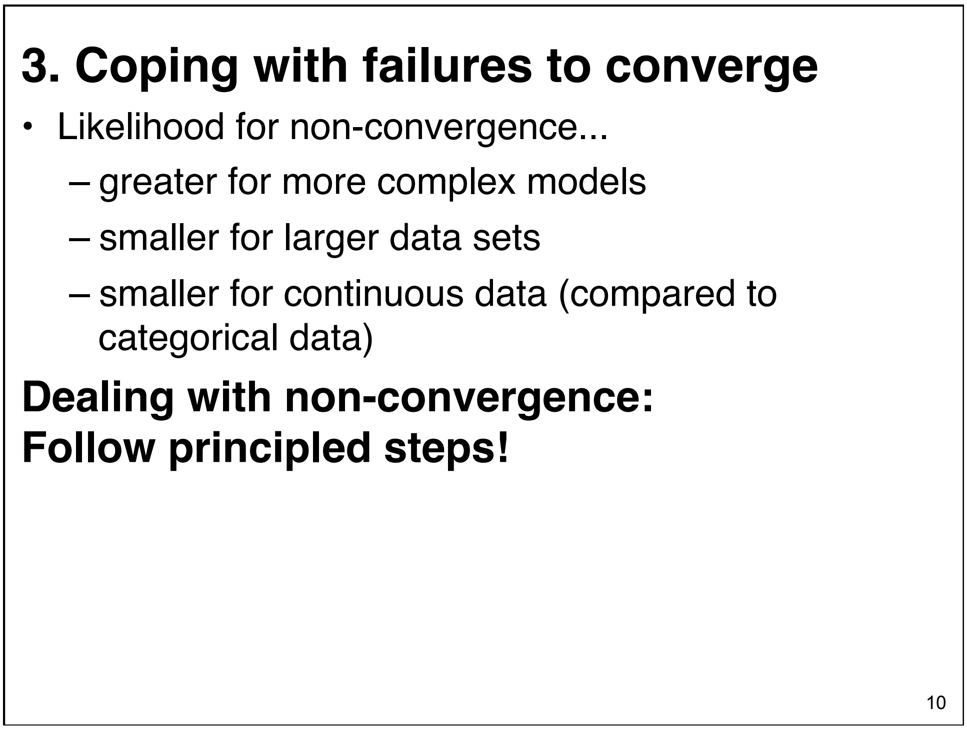
→ not enough data to estimate random slope

- random slope variation fully confounded with trial-level error
- can always fit a perfect line through 2 points



2. Random effects for control predictors

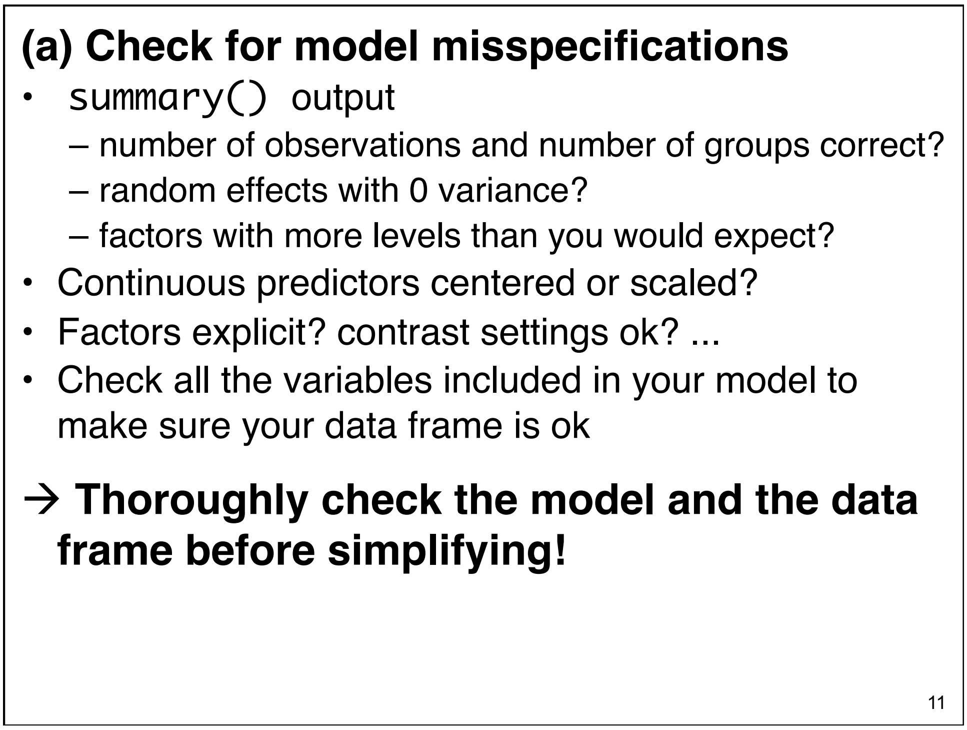
- = predictors that are not of interest to the researcher (e.g., rule out potential confounds or increase statistical power by reducing noise/unexplained variance)
- Include random slopes (and correlations) for them also? Can lead to VERY complex models...
- "Little guidance;" BUT: probably not necessary; fixed effects for them sufficient



3. Coping with failures to converge

- Likelihood for non-convergence...
  - greater for more complex models
  - smaller for larger data sets
  - smaller for continuous data (compared to categorical data)

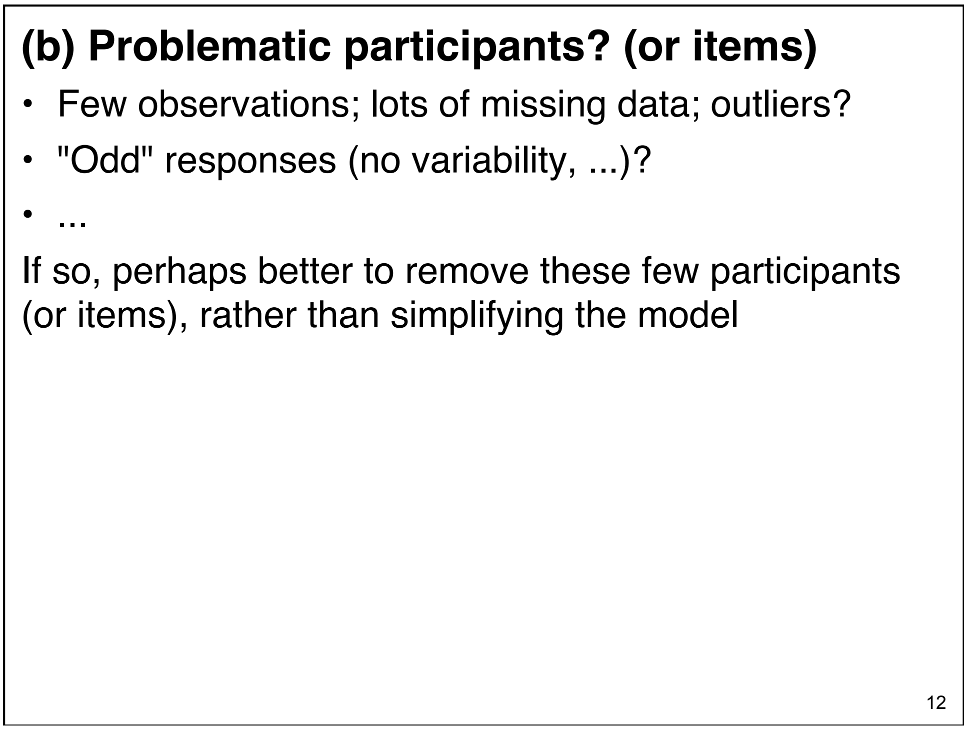
Dealing with non-convergence:
Follow principled steps!



(a) Check for model misspecifications

- `summary()` output
  - – number of observations and number of groups correct?
  - – random effects with 0 variance?
  - – factors with more levels than you would expect?
- Continuous predictors centered or scaled?
- Factors explicit? contrast settings ok? ...
- Check all the variables included in your model to make sure your data frame is ok

→ Thoroughly check the model and the data frame before simplifying!



(b) Problematic participants? (or items)

- Few observations; lots of missing data; outliers?
- "Odd" responses (no variability, ...)?
- ...

If so, perhaps better to remove these few participants (or items), rather than simplifying the model

(c) Bernd's advice

- Increase number of iterations
- Increase some more
- Scale instead of center (or vice versa)
- Try different contrast settings for factors
- Different optimizer (if possible)
- More recent package versions (particularly lme4)?

If none of these things (a to c) help: Simplify! But how? Back to Barr...

(d) "Rule of thumb"

- For the fixed effects of interest, keep the corresponding random effects in the model
  - Remove first random covariance terms and/or even random intercepts
  - Remove random slopes last
- If there are several effects of interest and model doesn't converge with all corresponding random effects, try separate analyses...

Several separate analyses

- For severe cases...
- For example effects A and B of interest, both within

Analysis 1: test significance of A

- A fixed and random slope
- B only fixed slope

Analysis 2: test significance of B

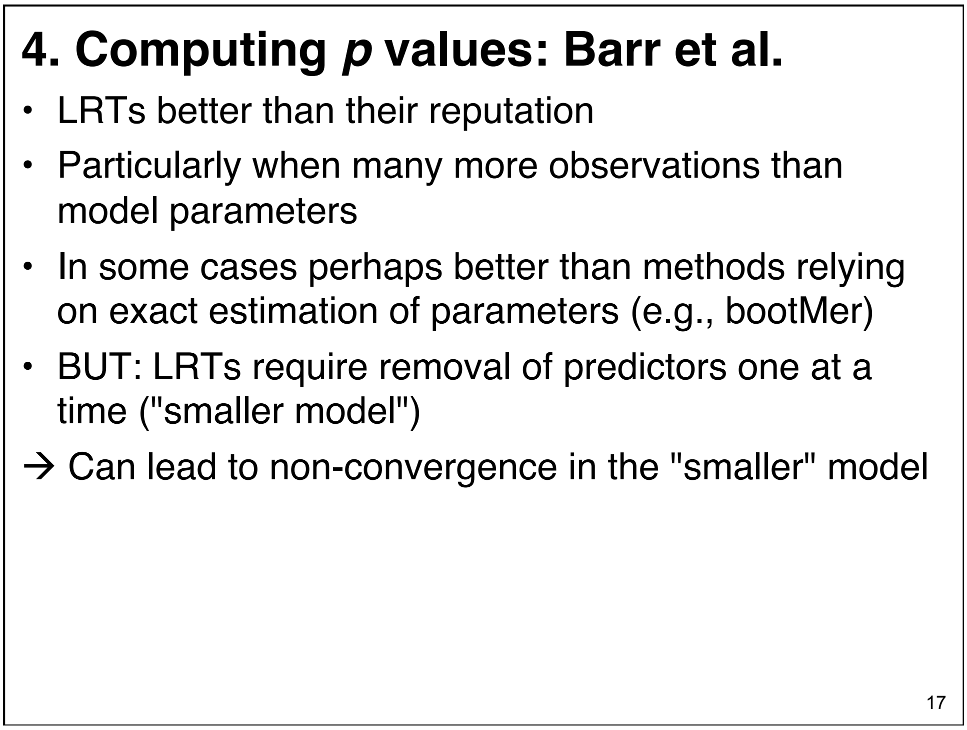
- B fixed and random slope
- A only fixed slope

If that still doesn't help...
(e) "Fallback Strategy"

→ Data driven approach (aka "model selection")

- **Barr et al** → forward strategy
  - start with simple model, test which random effects to add
  - include all that pass liberal criterion (e.g., LRT $p < .20$)
- **Bolker et al** → Information criterion approach
  - Avoid using p values for inclusion/exclusion decisions
  - Choose model that is best on AIC (or BIC or DIC)
  - Do confirmatory testing on that model

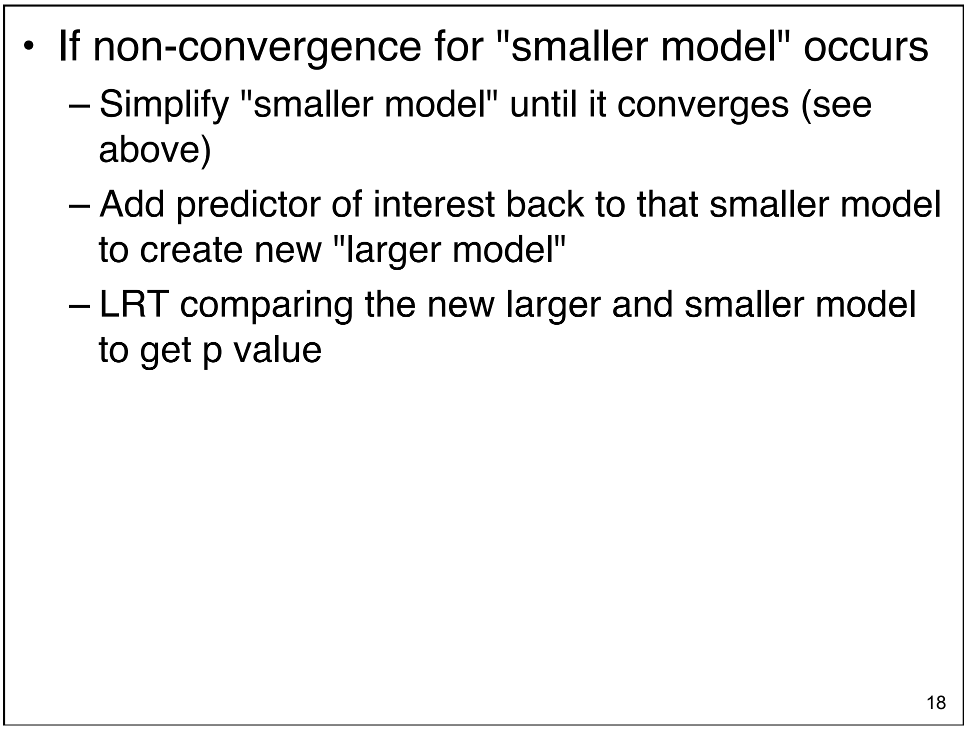
Important in all cases: Full disclosure!
Explain all the steps that you went through and on what criteria you based your modeling decisions



4. Computing *p* values: Barr et al.

- LRTs better than their reputation
- Particularly when many more observations than model parameters
- In some cases perhaps better than methods relying on exact estimation of parameters (e.g., bootMer)
- BUT: LRTs require removal of predictors one at a time ("smaller model")

→ Can lead to non-convergence in the "smaller" model



- If non-convergence for "smaller model" occurs
  - Simplify "smaller model" until it converges (see above)
  - Add predictor of interest back to that smaller model to create new "larger model"
  - LRT comparing the new larger and smaller model to get p value

4. Computing *p* values: Bolker (& lme4 team)

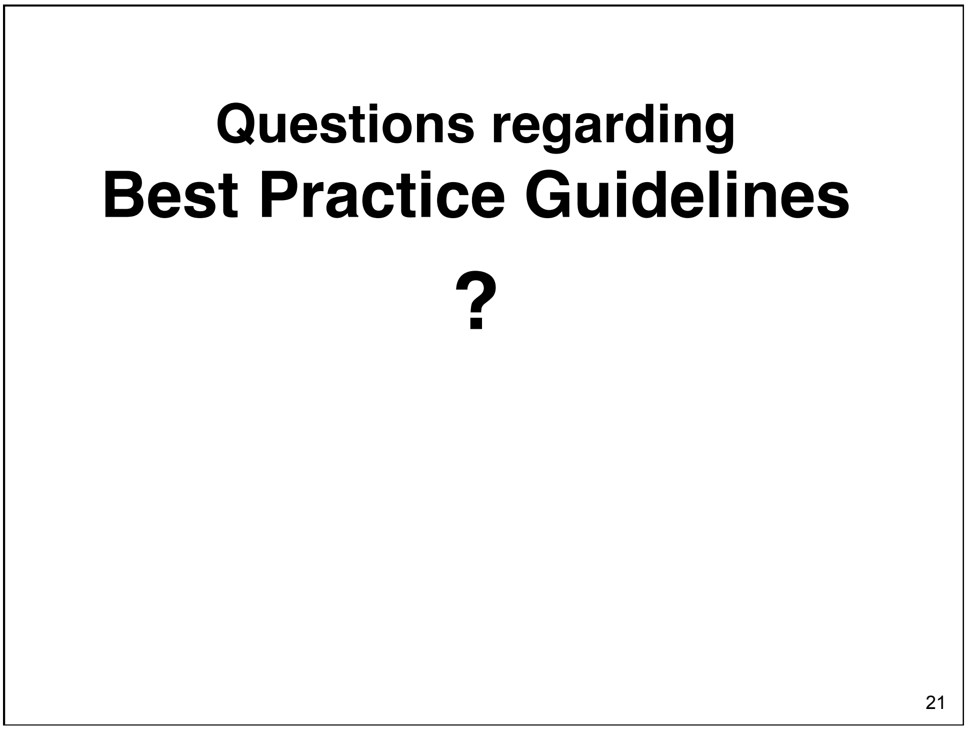
- More computationally intensive approaches typically more reliable
- Bootstrapping typically most trustworthy (PBmodcomp or bootMer)
- Simpler approaches often fine as well
  - Conditional F tests with df correction
  - `drop1()` or `anova()` (i.e., LRTs)

5. Reporting Results (Barr et al.)

- Complete model description with sufficient detail
- Which fixed and random effects, incl random correls
- If procedure included **several steps** (model selection or dealing with non-convergence): → describe procedure and your modeling decisions

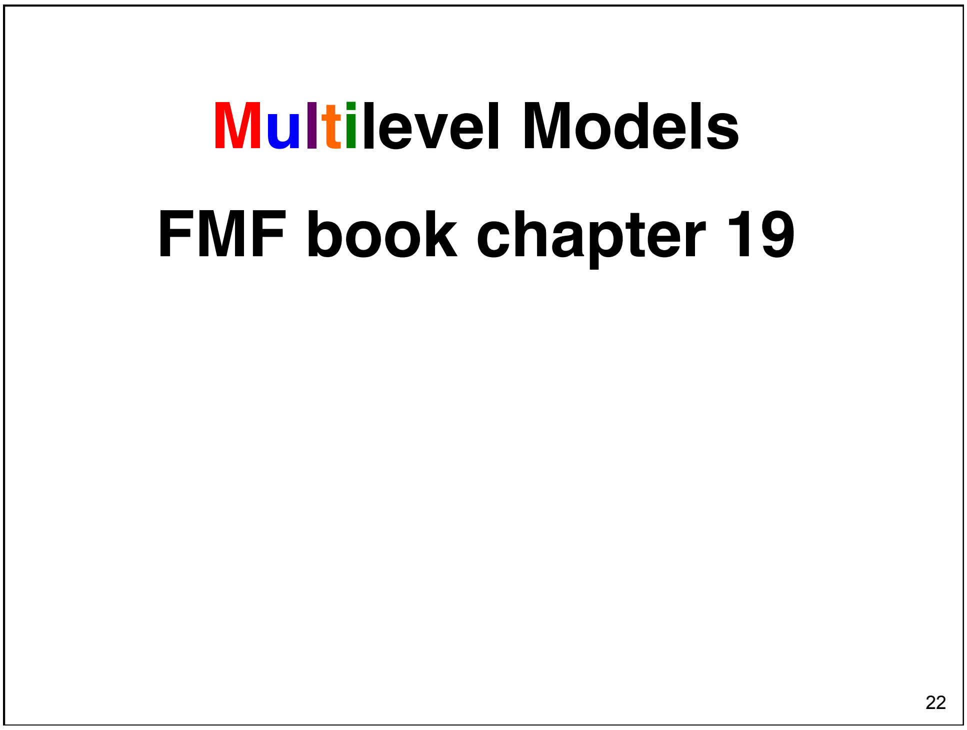
Barr et al suggestions

- Include information from `summary()` (rather atypical)
- Simpler
  - "I attempted to use a maximal random effects structure"
  - "Predictors A and B only fixed; predictors C and D fixed and random; random correlations between .. and ..." etc etc
- My example from last class: Good compromise (I hope)



Questions regarding
Best Practice Guidelines

?



Multilevel Models

FMF book chapter 19

Many things should be familiar

- Models to account for non-independence in data
- Advantages compared to, e.g., ANOVA
- ...

→ Good recap of things we discussed

New things (→ multilevel perspective)

- Nested/hierarchical multi-level data
- Quantify non-independence: baseline model, ICC
- Grand-mean vs. group-mean centering
- Covariance structures; growth-curve models; ...

The Multilevel Perspective

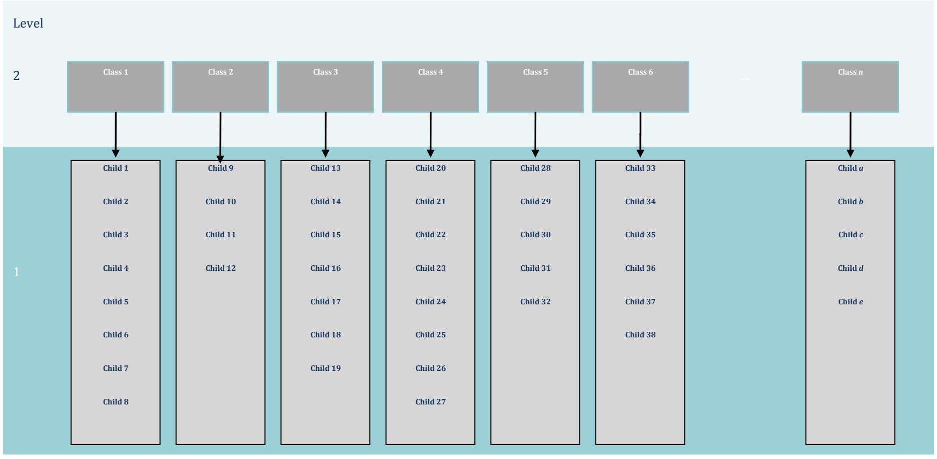
Focus on hierarchical structures

- Children nested within classes, nested within schools, ...
- Patients nested within doctors, nested within hospitals
- Employees nested within organizations
- Repeated measures nested in participants, nested in experimental conditions, ...

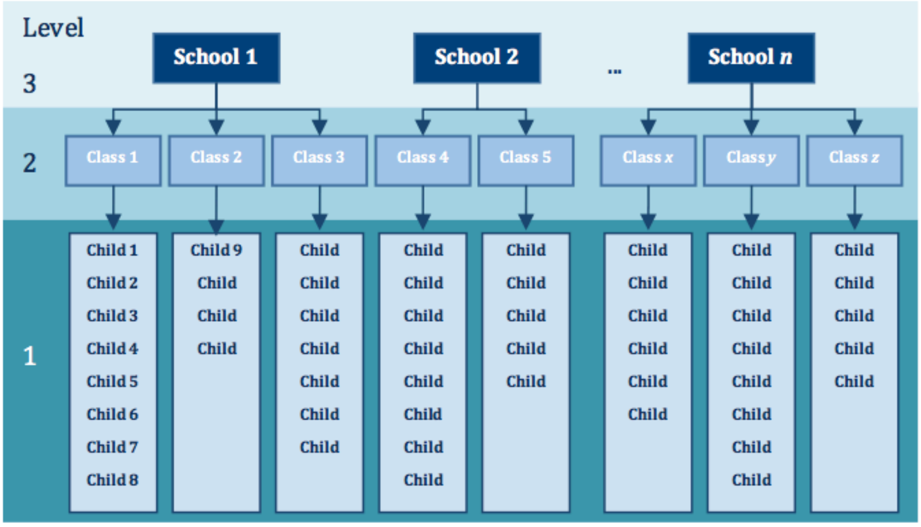
FMF: "contextual factors" (e.g., same teacher)

→ non-independence of observations
→ correlated residuals

A Two-Level Hierarchy



A Three-Level Hierarchy



Common Multilevel Procedure

First Step

- Determine dependency of observations in the data
  - "Baseline" model versus "null" model comparison
  - Intraclass correlation coefficient (ICC)
- Determine appropriate random effects structure
  - LRTs for random effects

→ Data-driven model-selection stage

→ FMF: start with simple model, increase complexity

Second Step

- Use resulting model for inference about fixed effects

Intraclass Correlation Coefficient

Quantifies dependency within units ("similarity")

- Example: children in class A are more similar to each other (compared to children in class B)
- **Large ICC** → strong non-independence of observations (children within same class similar)
- **Small ICC** → little similarity within units; observations are relatively independent

ICC → strength of effect of contextual variable
(does it matter whether a child is in class A or B?)

How to compute the ICC

In lme4

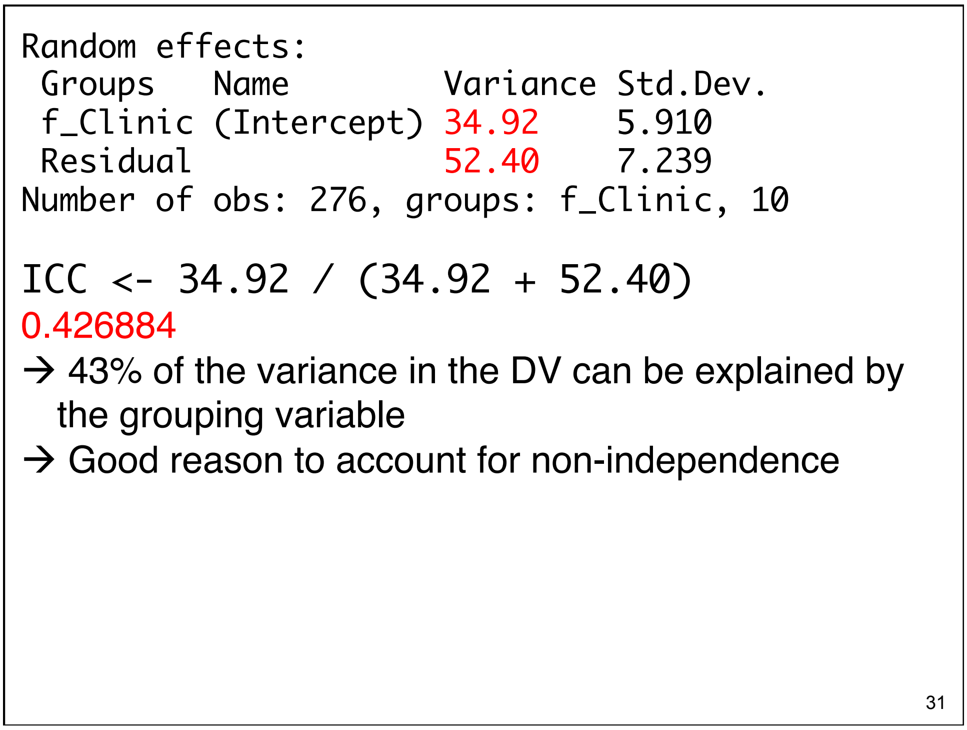
(1) Run "null model" (aka "empty model")

- Consists only of fixed intercept plus random intercept for grouping variable

```
m_0 <- lmer(DV + (1 | group), data =...)
m_0 <- lmer(Post_QoL ~ (1 | f_Clinic),
data = surgeryData)
```

(2) Divide the variance explained by the random intercept by the sum of that variance plus the residual variance

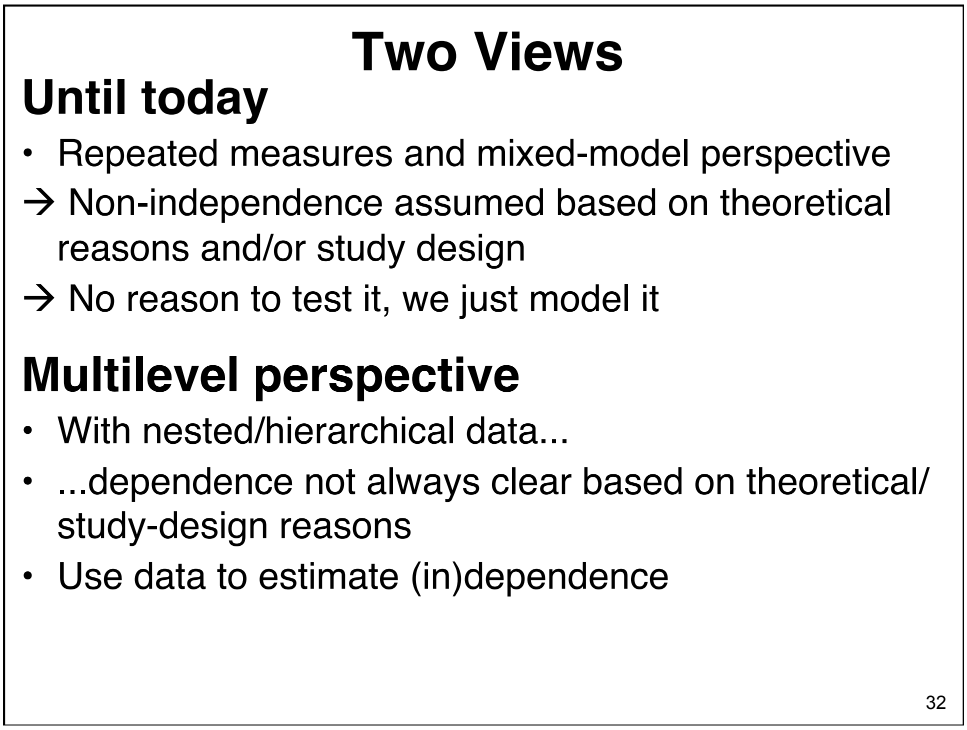
```
> summary(m0)
Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: Post_QoL ~ (1 | f_Clinic)
   Data: surgeryData

     AIC      BIC   logLik deviance df.resid
  1911.5   1922.3   -952.7   1905.5      273

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8828 -0.7607 -0.1379  0.7075  2.8608

Random effects:
 Groups   Name        Variance Std.Dev.
 f_Clinic (Intercept) 34.92    5.910
 Residual             52.40    7.239
Number of obs: 276, groups: f_Clinic, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)    60.08       1.92    31.3
```



```
Random effects:
 Groups   Name        Variance Std.Dev.
 f_Clinic (Intercept) 34.92    5.910
 Residual             52.40    7.239
Number of obs: 276, groups: f_Clinic, 10

ICC <- 34.92 / (34.92 + 52.40)
0.426884
```

→ 43% of the variance in the DV can be explained by the grouping variable

→ Good reason to account for non-independence



Two Views

Until today

- Repeated measures and mixed-model perspective

→ Non-independence assumed based on theoretical reasons and/or study design

→ No reason to test it, we just model it

Multilevel perspective

- With nested/hierarchical data...
- ...dependence not always clear based on theoretical/ study-design reasons
- Use data to estimate (in)dependence

Centering: Two common ways

For predictor variables (not the DV)

- Grand-mean centering
- Group-mean centering

Grand-mean centering

- More common
- From each value in the predictor, the overall mean is subtracted
- To reduce multi-collinearity
- Intercept → for **average** predictor value (often easier to interpret)

Group-mean centering

- From each value in the predictor, the mean of the corresponding "unit" is subtracted
  - How? → Ron's slides on `plyr` and `ddply` (slides 10-19)

Example: Repeated-measures RT task

- Predict performance (correct/incorr) by trial's RT
- Participant A: mean RT of 513 msec
  - For all her observations, I subtract 513 for the new group-centered RT predictor
- Participant B: mean RT of 298 msec
  - For all his observations, I subtract 298 for the new group-centered RT predictor

- Typically two predictors in model:
  - New group-centered RT predictor
  - Each participants' average RT

RT task example

- Grand-mean centered: How is RT related to correct vs. incorrect responding?
- Group-mean centered: How are participants' **atypically long/short RTs** related to correct/incorrect responding? → Individual "reference-level" (shorter/longer than usual)

When which centering approach?

- Default: grand-mean centering
- Group-mean centering: if specific research question

→ **more information Enders & Tofighi (2007)**

lme4 versus nlme

nlme is predecessor of lme4 (same developers)

- **lme4:** faster, more flexible, bootMer makes bootstrapping easy, ...
- **nlme:** `summary()` gives p values; different covariance structures to choose from

Recommendation: lme4, unless good reasons to use nlme

FMF Example: Cosmetic Surgery

Is quality of life related to cosmetic surgery?

BlackBoard → Course Documents → Week 6 → "Cosmetic Surgery.dat"

Data set

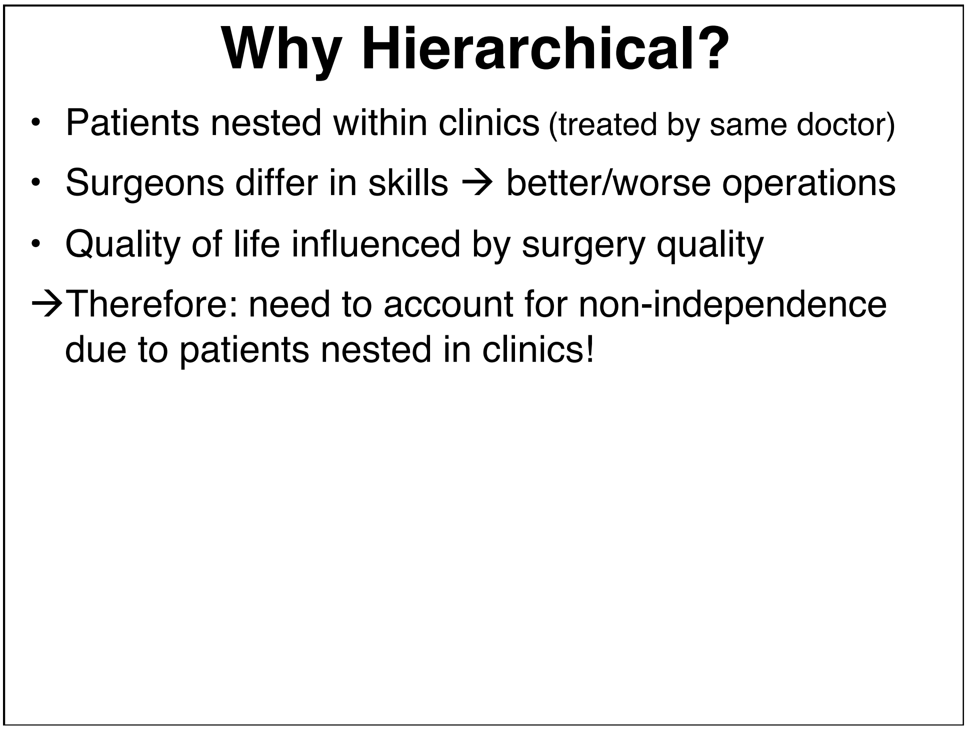
- Total of 276 patients (= level 1): each only 1 data point!
- 10 clinics participated (= grouping unit!)

Patients characteristics

- In which clinic?
- Already undergone surgery or waiting for it
- Quality of life after surgery
- Quality of life before surgery
- Medical or purely aesthetic reasons for surgery
- Age
- Depression (BDI)
- Gender

Variable Names

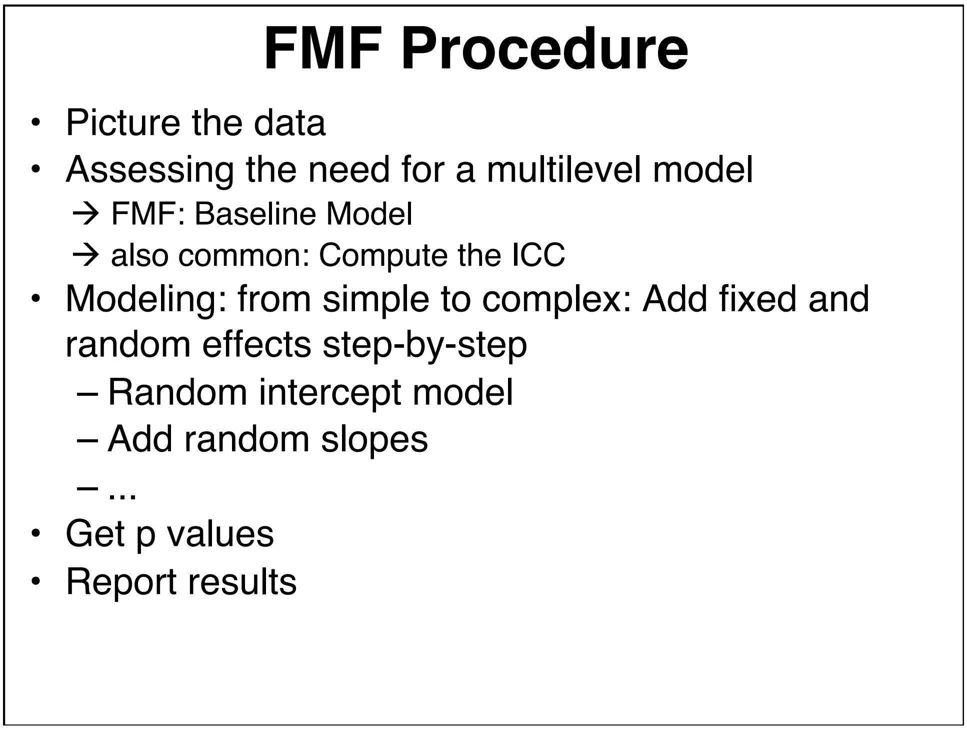
- Post_QoL: measure of quality of life **after** the cosmetic surgery.
- Base_QoL: Quality of life **before** the surgery.
- Surgery: A dummy variable that specifies whether the person has **undergone cosmetic surgery (=1)** or whether they are on the **waiting list (=0)**.
- Clinic: Which of 10 clinics the person attended to have their surgery.
- Age: The person’s age in years.
- BDI: Natural levels of depression measured using the Beck Depression Inventory (BDI).
- Reason: This dummy variable specifies whether the person had/is waiting to have surgery purely to **change their appearance (=0)**, or because of a **physical reason (=1)**.
- Gender: Whether the person was a **man (=1)** or a **woman (=0)**.



Why Hierarchical?

- Patients nested within clinics (treated by same doctor)
- Surgeons differ in skills → better/worse operations
- Quality of life influenced by surgery quality

→Therefore: need to account for non-independence due to patients nested in clinics!



FMF Procedure

- Picture the data
- Assessing the need for a multilevel model
  → FMF: Baseline Model
  → also common: Compute the ICC
- Modeling: from simple to complex: Add fixed and random effects step-by-step
  - Random intercept model
  - Add random slopes
  - ...
- Get p values
- Report results

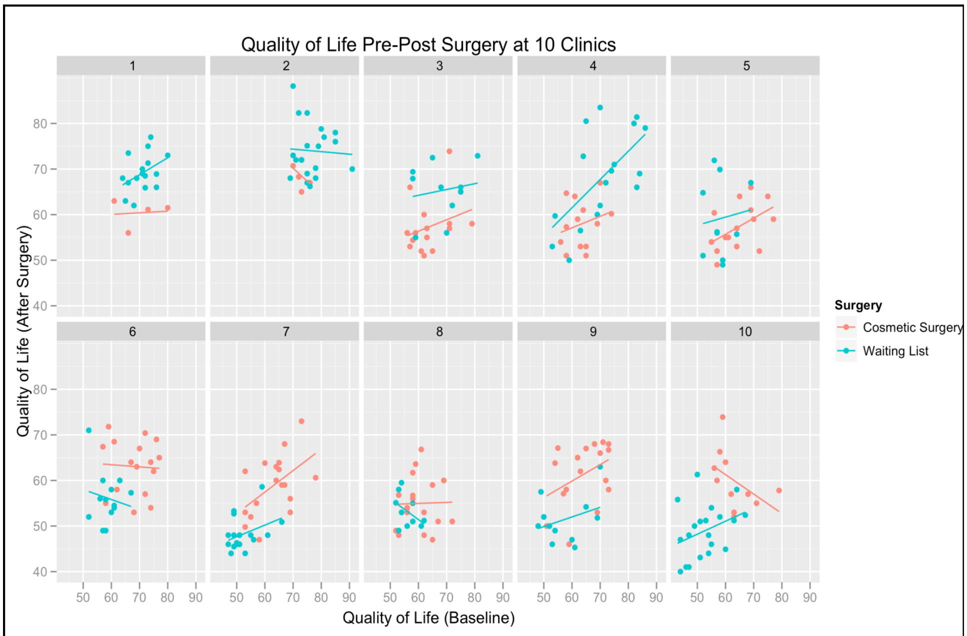
Picture The Data

Relationship between pre- and post-surgery quality of life as a function of:

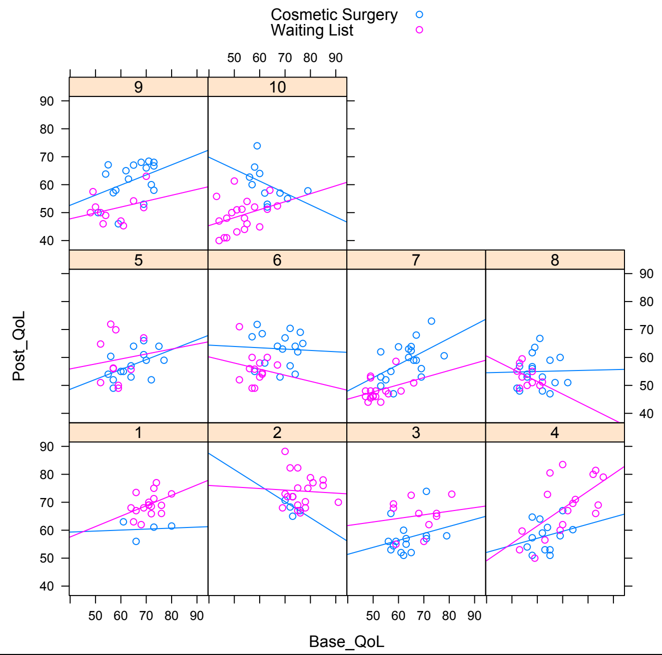
- Clinic
- before vs. after surgery

Code for graph in book (ggplot2)

```
pgrid <- ggplot(surgeryData, aes(Base_QoL, Post_QoL)) +
opts(title="Quality of Life Pre-Post Surgery at 10 Clinics")

pgrid + geom_point(aes(colour = Surgery_Text)) +
geom_smooth(aes(colour = Surgery_Text), method = "lm", se = F)
+ facet_wrap(~Clinic, ncol = 5) + labs(x = "Quality of Life
(Baseline)", y = "Quality of Life (After Surgery)")
```



```
xyplot(Post_QoL ~ Base_QoL | as.factor(Clinic), group = Surgery_Text,
data = surgeryData, type = c('p', 'r'), auto.key = TRUE)
```



Assess Need for Multilevel Model

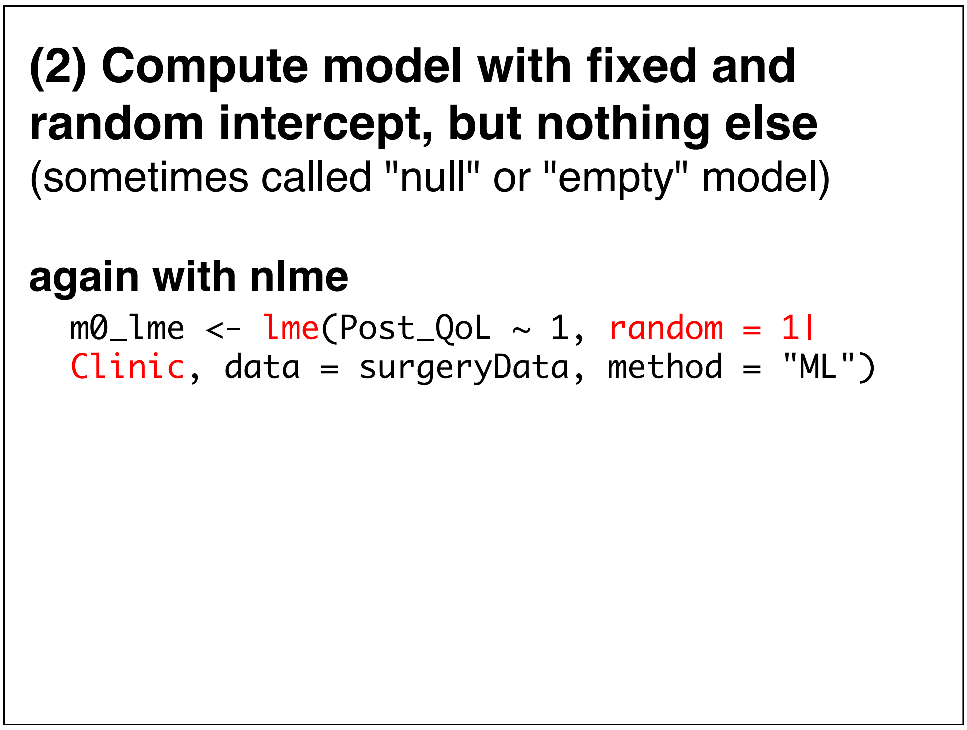
(1) Compute baseline model

- Contains only **fixed** intercept, nothing else

→FMF use `gls()` from `nlme`

`gls()` = generalized least squares

```
m00 <- gls(Post_QoL ~ 1,
data=surgeryData, method="ML")
```

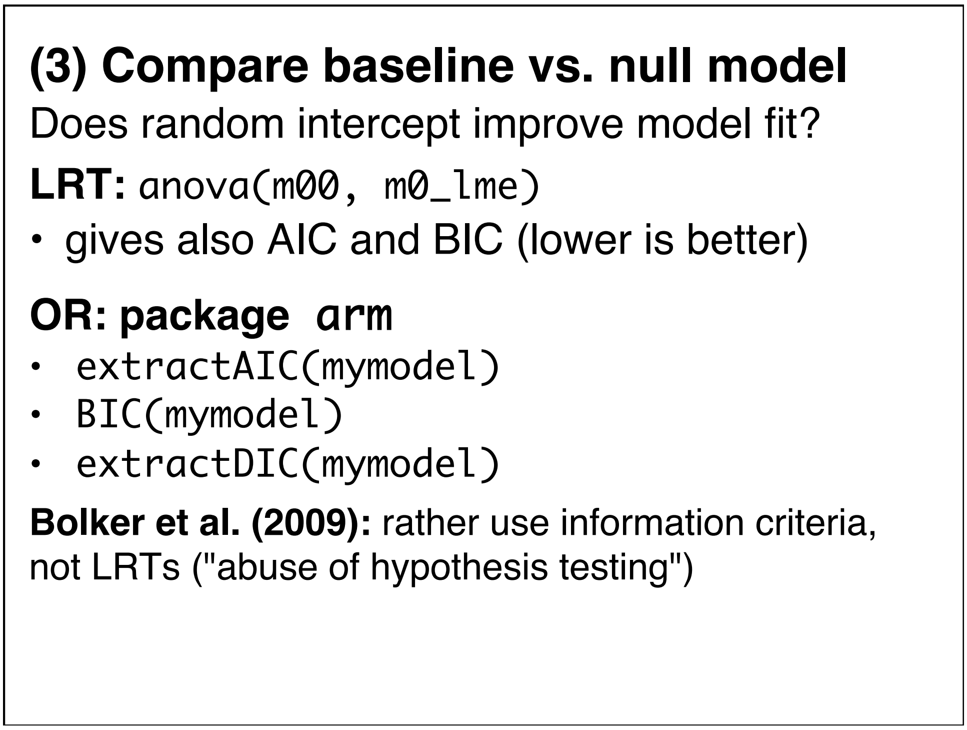


(2) Compute model with fixed and random intercept, but nothing else

(sometimes called "null" or "empty" model)

again with nlme

```
m0_lme <- lme(Post_QoL ~ 1, random = 1|
Clinic, data = surgeryData, method = "ML")
```



(3) Compare baseline vs. null model

Does random intercept improve model fit?

LRT: `anova(m00, m0_lme)`

- gives also AIC and BIC (lower is better)

OR: package `arm`

- `extractAIC(mymodel)`
- `BIC(mymodel)`
- `extractDIC(mymodel)`

Bolker et al. (2009): rather use information criteria, not LRTs ("abuse of hypothesis testing")

(4) Increasing complexity

FMF add then fixed effects and random slopes in stepwise matter, always checking LRT and AIC/BIC with anova()

In the end

```
final_model <- lme(Post_QoL ~ Surgery +
Base_QoL + Reason + Reason:Surgery, data =
surgeryData, random = ~Surgery|Clinic, method =
"ML")
summary(final_model)
```

summary(final_model) → **What do you notice?**

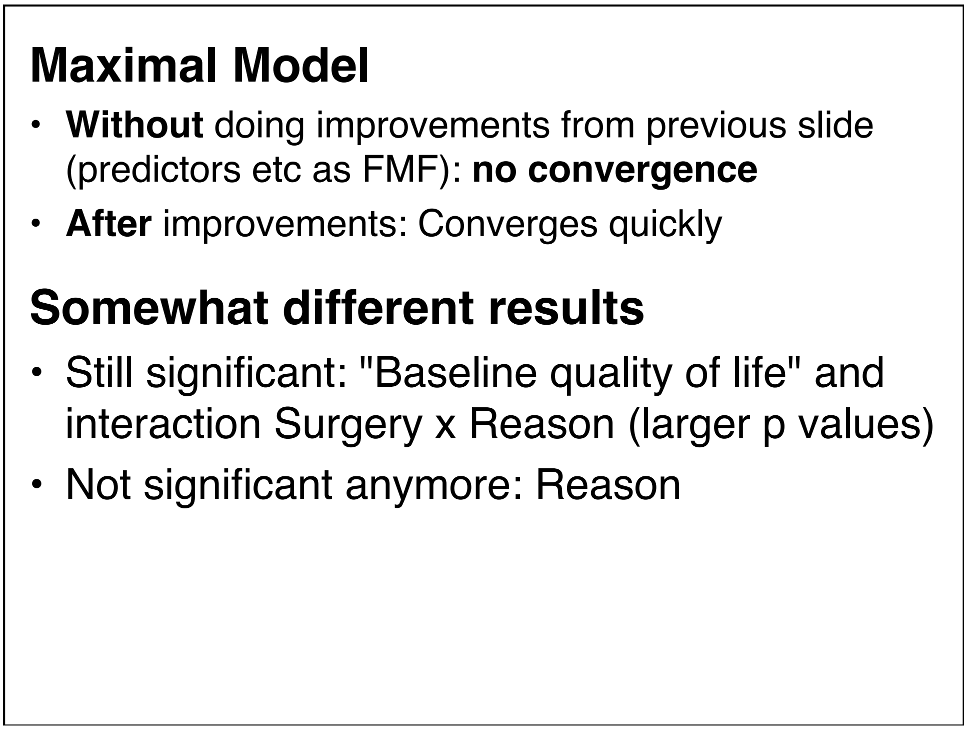
```
             StdDev   Corr
(Intercept) 5.482366 (Intr)
Surgery     5.417501 -0.946
Residual    5.818910

                   Value Std.Error  DF   t-value p-value
(Intercept)     42.51782  3.875318 262 10.971440  0.0000
Surgery         -3.18768  2.185369 262 -1.458645  0.1459
Base_QoL         0.30536  0.053125 262  5.747833  0.0000
Reason          -3.51515  1.140934 262 -3.080938  0.0023
Surgery:Reason   4.22129  1.700269 262  2.482717  0.0137
```

→What do you notice?

- *p* values!
- Only 2 random effects
  - Random intercept varying over Clinic
  - Surgery: random slope varying over Clinic
  - What about other predictors??!! no random slopes?
- 0/1 coded binary variables; some NOT explicity factors; dummy coding
- non-centered, non-scaled continuous predictors

What would we do differently?

- Not stepwise approach
- use lme4
- *p* values via one of the discussed approaches (e.g., Conditional F tests with K-R df adj)
- Sum-to-zero contrasts
- center or scale continuous predictors
- Maximal random effects structure

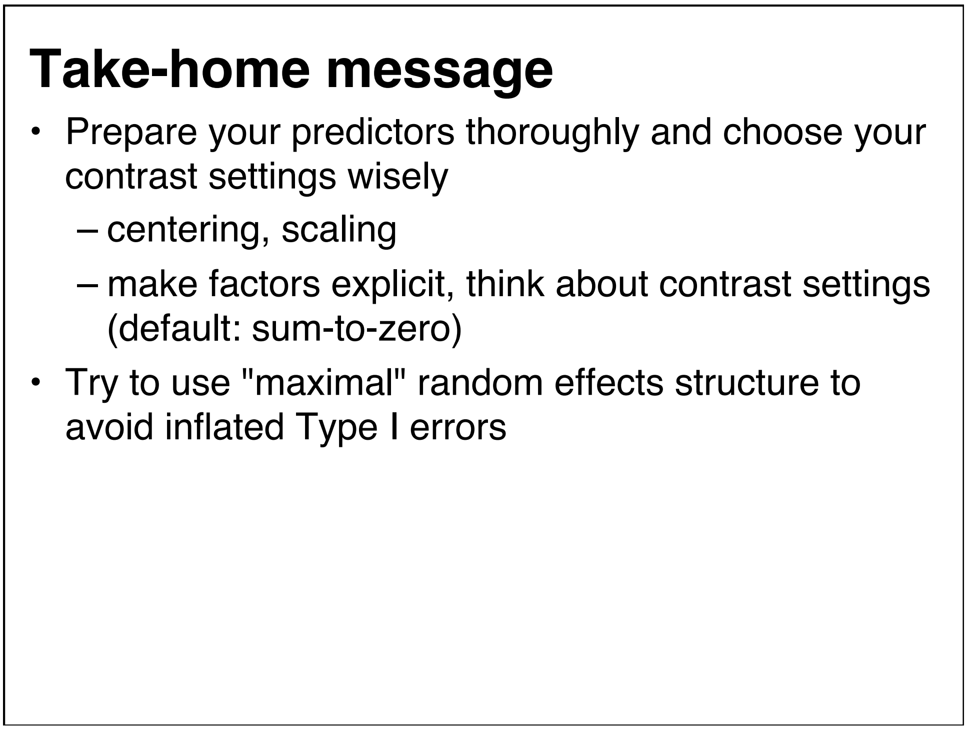


Maximal Model

- **Without** doing improvements from previous slide (predictors etc as FMF): **no convergence**
- **After** improvements: Converges quickly

Somewhat different results

- Still significant: "Baseline quality of life" and interaction Surgery x Reason (larger p values)
- Not significant anymore: Reason



Take-home message

- Prepare your predictors thoroughly and choose your contrast settings wisely
  - centering, scaling
  - make factors explicit, think about contrast settings (default: sum-to-zero)
- Try to use "maximal" random effects structure to avoid inflated Type I errors

Stay tuned for...

...next week:

- Real data from (PhD) student projects!
- More on multilevel models
  - Covariance structures
  - Growth curve models
- **Generalized** linear mixed models (GLMMs)

Questions? Comments?

Homework

(1) Use Cosmetic Surgery data (on BlackBoard)

- Run FMF "final_model" using nlme (as in book)
- Run same model in lme4 (with FMF predictors)
- Fix the problems
  - Sum-to-zero contrasts
  - Turn categorical predictors into explicit factors
  - center or scale continuous predictors, ...
- Run it again in lme4, with the problems fixed
- Create and run a "maximal" version of the model
- Compare the results of the 4 models: explained variance of random effects; p values of fixed effects

Homework cont.

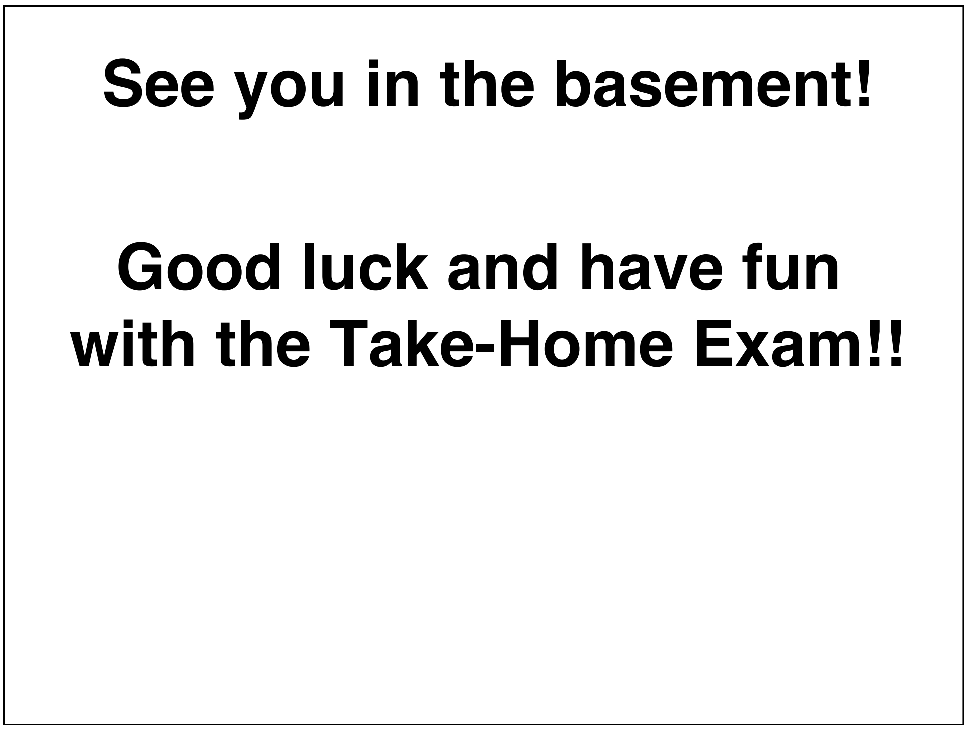
(2) Read the following paper

(much easier than Bolker et al., 2009!)

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

What you have to hand in

- Sworn statement that you did (1) and (2)
- For (1): R script with your observations/thoughts added as comments
- Deadline: March 24, 15:30



See you in the basement!

Good luck and have fun
with the Take-Home Exam!!