# Week 4: Three-level Nesting

- This week we will focus on more complex random effect structures. In this lecture, I'll show you how to handle three levels of nesting
- So far, the data we've analysed had two levels of nesting: there was a group level and an individual participant level, with multiple observations for each participant
    - It's also possible to have more levels of nesting or other nesting structures, all of which can be captured by specifying the random effect structure
- Let's consider three-levels of nesting: this is a hypothetical treatment study where each therapist administers a control or treatment condition to multiple subjects who are tested at multiple time points.
    - So, working from the bottom up, we have observations at different time points nested within subjects, and subjects nested within therapists. The group (treatment vs. control) is between-subjects (and between-therapists)
- Here's another example (with simulated data): imagine we have a new computer-based active learning method and we want to know if it improves scores on a math test
    - We measured students' math scores and proportion of time spent using the computer-based active learning software
    - This program was implemented in 3 schools, 8-12 classrooms per school, and 12-24 students per class
    - So students are nested within a classroom (those student data are not independent) and classrooms are nested within school (those classroom data are not independent)
- The data are stored as a CSV file and here I'm doing a little data wrangling when I read them in:
    - Creating a unique class ID (so classes don't get confused across schools)
    - Centering time spent on the computer relative to the overall mean
    - Centering class size relative to the overall mean and the school mean
    - Recall: the point of the centering is so that we'll be estimating meaningful intercept parameters rather than estimating for class sizes of 0 or no time spent on the computer
- Once the data are read in, we can make some exploratory plots to get a sense of what is going on
    - Looks like math scores are better in smaller classes (left)
    - And for students with a higher active learning proportion (right)
- Two more things we need to consider:
    - Left: Classes differ in size, and this is unequal across schools: school 1 tends to have smaller classes and school 3 tends to have larger classes
    - Right: Students differ in proportion of time spent on active learning, but each class has a big range and the ranges seem fairly similar across classes
- Let's start with a really simple model: math score as function of active learning time and class size, with random intercepts by class (that is, classrooms have vary in their average math score)
    - Both main effects are significant and the interaction is marginal
    - Keep an eye on the effect of class size and its interaction with active learning time
- If we look at the model fits, we can see offset parallel lines for the classrooms – they had random intercepts but not random slopes for the effect of active learning
    - To allow that variability, we need to add random slopes
- In this model, the fixed effects are the same, but we've added by-class random slopes of active learning time

        - Adding those random slopes substantially increased the standard errors and reduced estimated degrees of freedom for Active Time fixed effects. The main effect was very strong, and it still is, but the interaction is not significant any more
    - I said before that omitting random slopes makes the corresponding fixed effects anti-conservative: this is an example of that pattern
- When we plot the model fits we can now see random class-level variation in the relationship between math scores and active learning (that is, the lines have different slopes)
- This 2-level model assumes that classrooms are independent, but this is not quite true. Classrooms are nested within schools and there may be school-level differences in math scores, effects of active learning, class size, etc.
    - If we split up the data by school, we can see that there really might be school-level differences and those aren't anywhere in the 2-level model
- There's a specific issue here to be concerned about: school and class size are confounded. What appears to be a class size effect, might actually be differences between schools.
    - A class size effect would be interesting – that's something we might try to generalise to other classes.
    - But it's hard to know what differences between schools mean, especially when we only have 3 schools and don't know anything else about them
    - One way to deal with this is to use school-mean centered class sizes and control for overall school-level differences
- This confound is related to an important statistical paradox called Simpson's Paradox: you can get an overall effect even when each of the sub-groups shows *the opposite* effect. This happens when there is a confounding variable. A famous example is a study of gender bias in admissions at UC Berkeley in the 1970's
    - In multilevel modelling contexts we need to be particularly careful about sub-group (that is, cluster-level) differences that might be confounded with the variables of interest
- Coming back to our model, here's how we can capture three levels of nesting: we use the school-mean-centered class size and we have both school-level and class-level random effects
- And now the class size effect is no longer significant (not even close). Looks like school differences were masquerading as class size differences – a Simpson's paradox type of effect
- There is one more thing to worry about: there were only 3 schools, so how is the model going to estimate those school-level random effects
    - Looks like it had some trouble: the by-school intercept-slope correlation is -1.00, which is almost certainly untrue
    - We can use the double-pipe notation to remove that mis-estimated correlation: you can see we still have school-level random slopes and random intercepts, but not the correlation between them
- That didn't make a big difference for the fixed effects – class size still not significant, interaction still marginal. If it *had* made a big difference, I'd be worried about the overall model, but this way I'm feeling pretty confident about these results
- To sum up the key points from this lecture
    - The nested structures we've talked about can extend to three (or more) levels and this can be captured by the random effects structure
    - This is particularly important when sub-group (cluster-level) differences might be confounded with variables of interest (Simpson's Paradox).

- We saw a concrete example of how omitting random slopes tends to make models anti-conservative (inflates rates of false positives). My general recommendation is to start with a "full" or "maximal" random effect structure and reduce as needed
- We saw an example of that when we had only a few observational units (only 3 schools) and estimating the random effects is very difficult. This can produce unreliable fixed effect estimates. Over-parameterised random effect structures should be simplified and we saw how to identify a (probably) mis-estimated random effect and remove it from the model.

# Week 4: Crossed Random Effects

- This week we are focusing on more advanced nesting structures and how to capture them with random effects. In this lecture I will cover crossed random effects
- In pretty much all of the examples you've considered so far, there was variability between subjects and you wanted to evaluate the reliability of some effect in the context of that variability
    - We can ask an analogous question about variability between test items. This often comes up in laboratory studies where participants solve some problems or answer some questions.
        - We could average across those test items to get a subject mean, but that inter-item variability might be important
        - The inferences we want to make are usually intended to generalise to other items of the same general type so we have the same problem of quantifying the effect relative to inter-item variability
    - Historically, this was done by conducting separate by-subjects and by-items analyses, which were called "F1" and "F2" tests. And journals sometimes even required this (a long time ago, I had a paper rejected because a key result was statistically significant by subjects but only marginal by items).
    - Multilevel models offer a better solution to this problem
- Here's an example using simulated data for an intervention intended to improve problem solving ability. 120 participants were randomly assigned to either the Treatment or Control condition, then solved 30 problems (16 hard ones and 14 easy ones)
        - The outcome we care about is the response time (RT) for a correct solution. Note that there are missing data because if the participant failed to solve the problem, there is no RT
- If we were going to do a traditional kind of analysis, we'd take these data and calculate by-subject means for each problem type
    - Looks like everyone solves Easy problems faster than Hard ones, and
        - The Treatment group seems faster at problem solving, esp. for Hard problems
- Then we'd run a by-subject repeated measures ANOVA, which we can do with the afex package
    - Looks like there is an overall problem difficulty effect, and no effect(s) of the intervention.
    - But hang on: not all word problems are the same and we're going to make inferences about solving problems of this type, not just about solving these particular problems
- So we can do the analogous by-items thing: calculate item means
- And run a by-items repeated measures ANOVA – in this analysis all of the effects are significant. Why the difference?
- Problem type (difficulty) had a large effect, and that effect was significant in both analyses
        - The effect of Condition (the intervention) and its interaction with problem type were small.
    - Condition is between-subjects but within-items, so the between-subjects variability is strong in the by-subjects analysis but gets averaged away in the by-items analysis.
        - This makes the by-items analysis look overly strong (subject variability is missing) and the by-subjects analysis look overly weak (items consistency is missing).
        - This kind of thing comes up a lot because we can try to improve power by designing studies to be within-subjects or within-items, but it's often impossible to do both.
    - Also, the idea that effects should be significant in separate by-items and by-subjects analyses (aka F1 and F2) is generally thought to be overly conservative.
- Multilevel models offer an alternative approach because they can simultaneously model random variability at subject and item levels, as well as the group-level effects that we are interested in.

        - The key observation is that data are nested (or clustered) both within-subjects (each subject solved a set of problems) and within-items (each problem was solved by a set of subjects). These are called “crossed random effects”
        - In the model specification we have by-subject random effects (with random slopes of problem type, since that was within-subjects) and by-item random effects (with random slopes for Condition, since that was within-items)
- When we do the analysis this way, each of the effects is significant
        - Though notice that they are not as strong as in the by-items ANOVA (which was overly strong) nor as weak as in the by-subjects ANOVA (which was overly weak)
        - This analysis is capturing both the variability and the consistency across these different levels of nesting.
- It can be tricky to plot the results of a crossed random effects analysis because either by-subject or by-item averaging can mis-represent the data. This is a case where the effects package is very useful – you can get the model-estimated means and confidence intervals
- and plot those: here I’ve constructed something like a box plot with a marker for the mean, thick lines for +/- 1 standard error, and thin lines for the 95% confidence intervals
- This week we covered some more complex random effect structures: 3-level nesting and crossed random effects. Although the random effects are more complicated, the other aspects of multilevel models that you learned in prior lectures also hold here:
        - p-value estimation was done using Satterthwaite method; model comparisons would've been a good alternative
        - start with a full or “maximal” random effect structure and can simplify if model doesn't converge by removing correlations and random "slopes"
        - be aware of how your categorical variables as coded; can conduct pairwise comparisons using a single model
        - use logistic regression for binary outcomes