Introduction to Mixed-Effects Regression using R.

Keith Lohse | PhD | Pstat | (he/him) |
Associate Professor | Physical Therapy and Neurology |
Washington University School of Medicine in Saint Louis |


lohse@wustl.edu

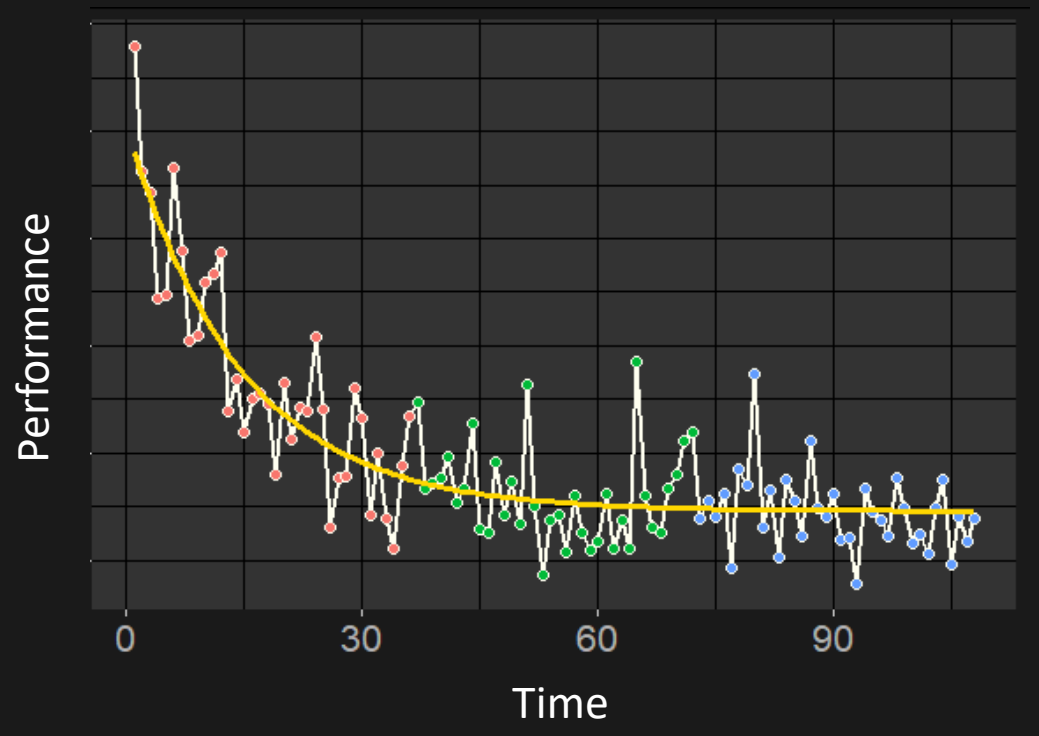

keithlohse

My Background

- In my research, I study the **psychological** processes of human learning, the **physiological** mechanisms of recovery, and how they interact during **rehabilitation**.

- I specialize in statistical modeling and methodology:
  - Reliability, validity, and conceptual challenges in studying recovery. [Lohse et al, *NNR*, 2021]
  - Methods for modeling longitudinal data. [Lohse et al., *JMLD*, 2020]
  - Statistical best practices. [Borg, Lohse, Sainani, *PMR*, 2020; Lohse et al., *JMLD*, 2016]
  - Pedagogy (HSE section of ASA) and continuing education. [Kozlowski & Lohse at ACRM; recent R25 submission with Lei Liew at USC]

Modeling Learning and Recovery



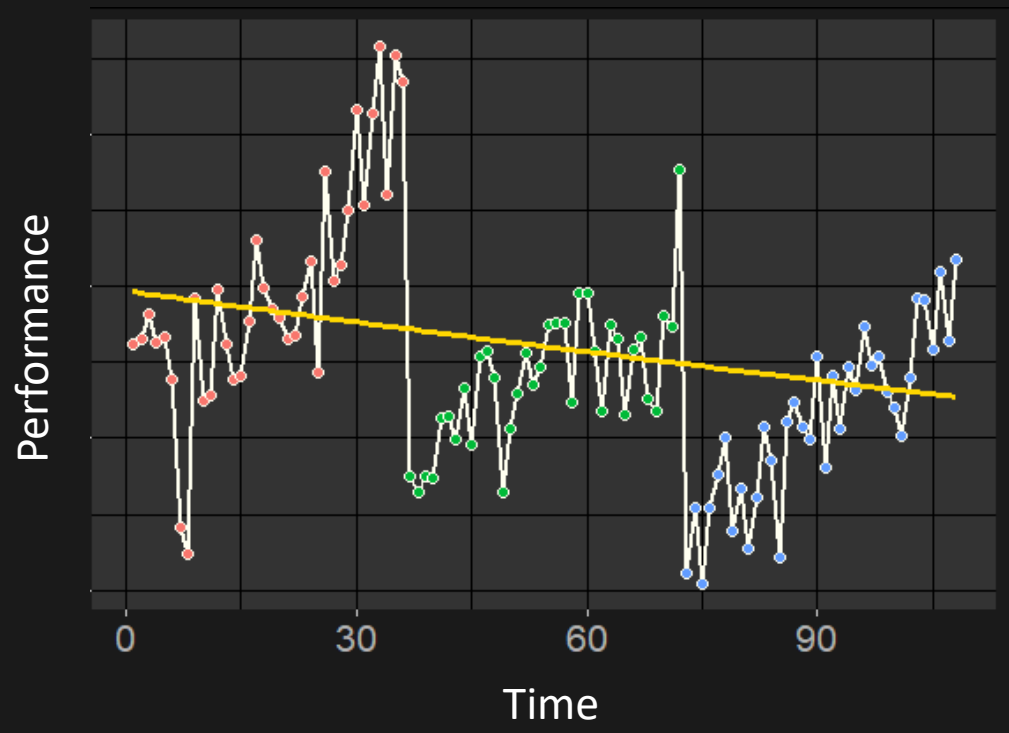
- Rehabilitation science fundamentally requires tracking some variable of interest over time.
- We can explain these data with an **exponential decay** function.

[Olivier, Paul, Lohse, et al., JNPT 2019]

Modeling Learning and Recovery

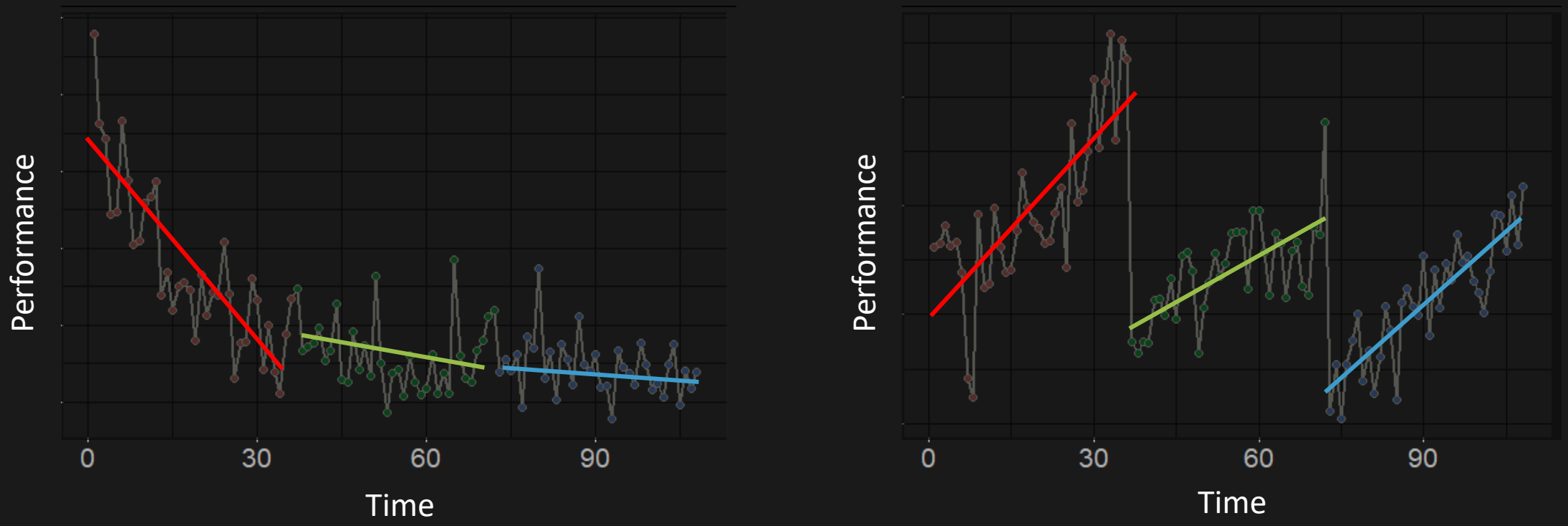
- But this person's data do not follow an exponential decay function!!!
- This person gets worse during any given day, but also improves from day to day.

[see also Park & Schweighofer, 2017]



[Olivier, Paul, Lohse, et al., JNPT, 2019]

Flexible Statistical Models



Using these models, we can answer important basic and applied questions in rehabilitation science. [Lohse et al., *APMR*, 2016; Lohse et al., *Bio Psyc*, 2019; Oliver, Lohse et al., *JNPT*, 2019]

Why mixed-effect regression?

Mixed-effect regression

- Many of us are probably comfortable with mixed-factorial *ANOVA* where we have between-subjects and within-subjects *factors*.
- Many of us are probably familiar *ordinary least squares* regression using the *general linear model*.
- Some of us probably recognize that these analyses are in fact related (i.e., factorial ANOVA is a special case of OLS GLM).
- Fewer of us are probably familiar mixed-effect regression as an analytical technique.

Mixed-effect regression is new(er)

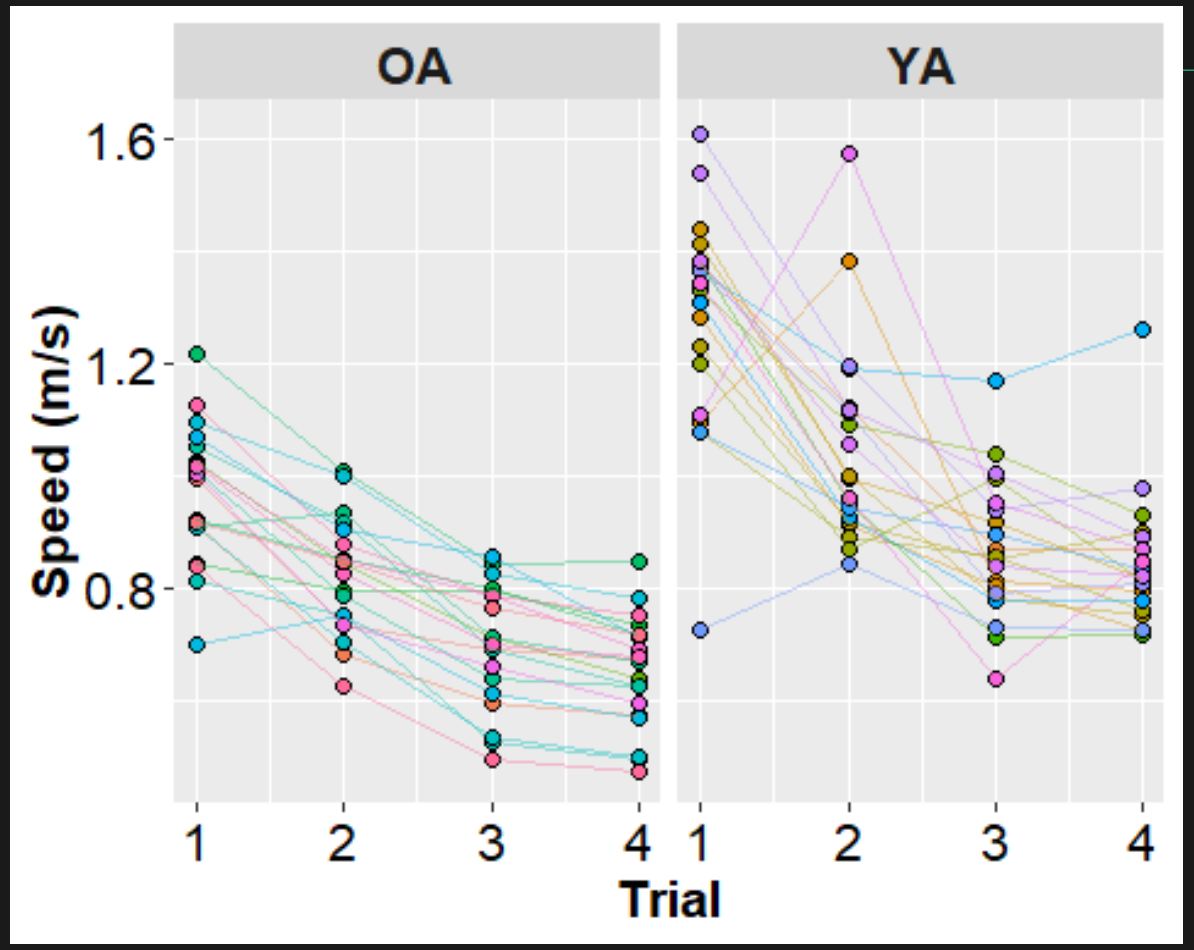
- If you have heard of it, you have probably heard of some of the advantages that mixed-effects regression has over ANOVA (e.g. its ability to handle missing data).

- However, because mixed-effects regression is relatively new (compared to ANOVA),
  - It is not taught in a lot of applied statistics classes,
  - It has less documentation for a non-specialist audience,
  - It is mostly applied in specialty fields,
  - It is often poorly reported in published literature ("*mixed-muddles*" – S. Senn)

In this workshop study, I want to…

- Provide an introduction to linear mixed-effect regression.
- Discuss its strengths and weaknesses, especially relative to factorial ANOVA.
- Give you data files and code to implement these models in the open programming language R.
- Leave you in a good position to further your education and apply these analyses to your own data.

What is Mixed-Effect Regression?



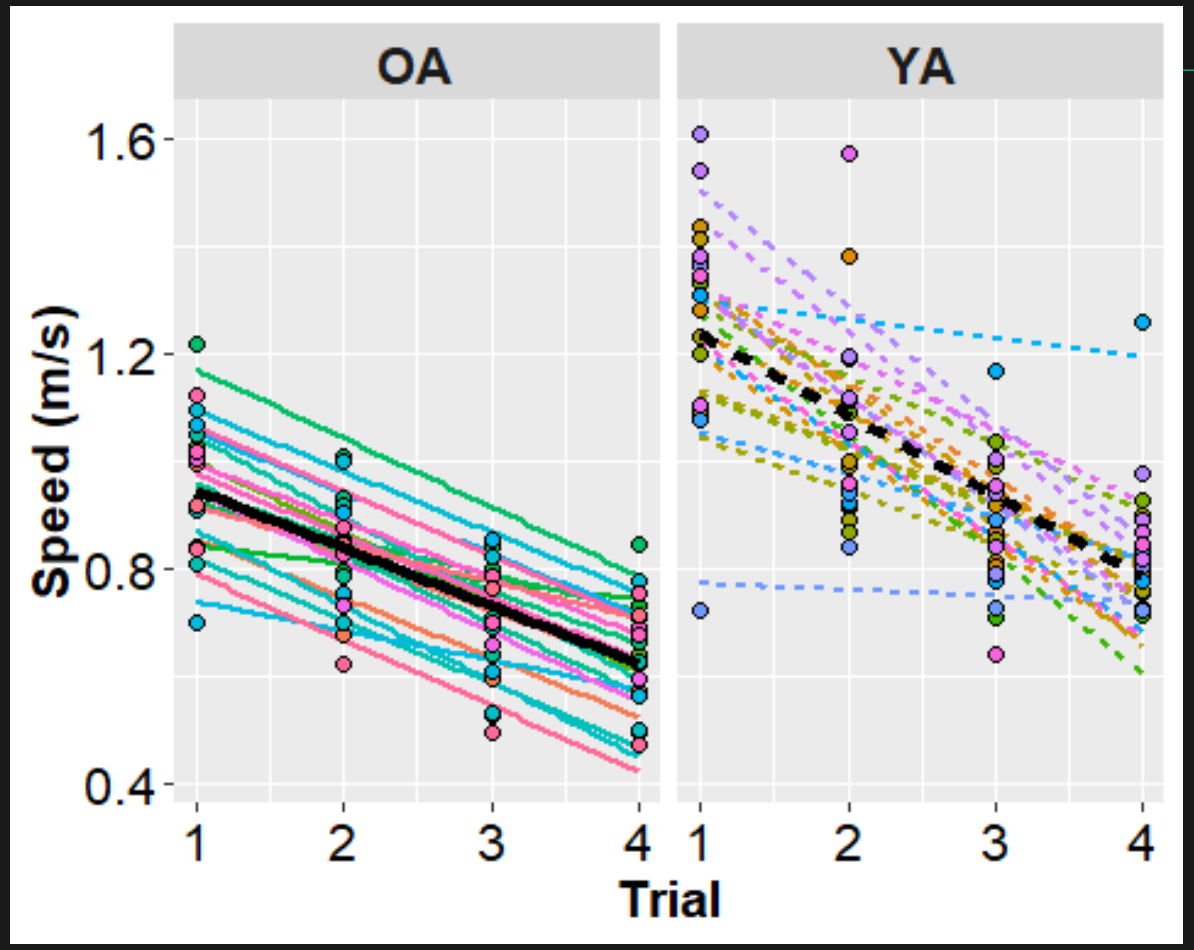
OLS - Regression

$$y_i = \beta_0 + \beta_1(Time_i) + \epsilon_i$$

LME - Regression

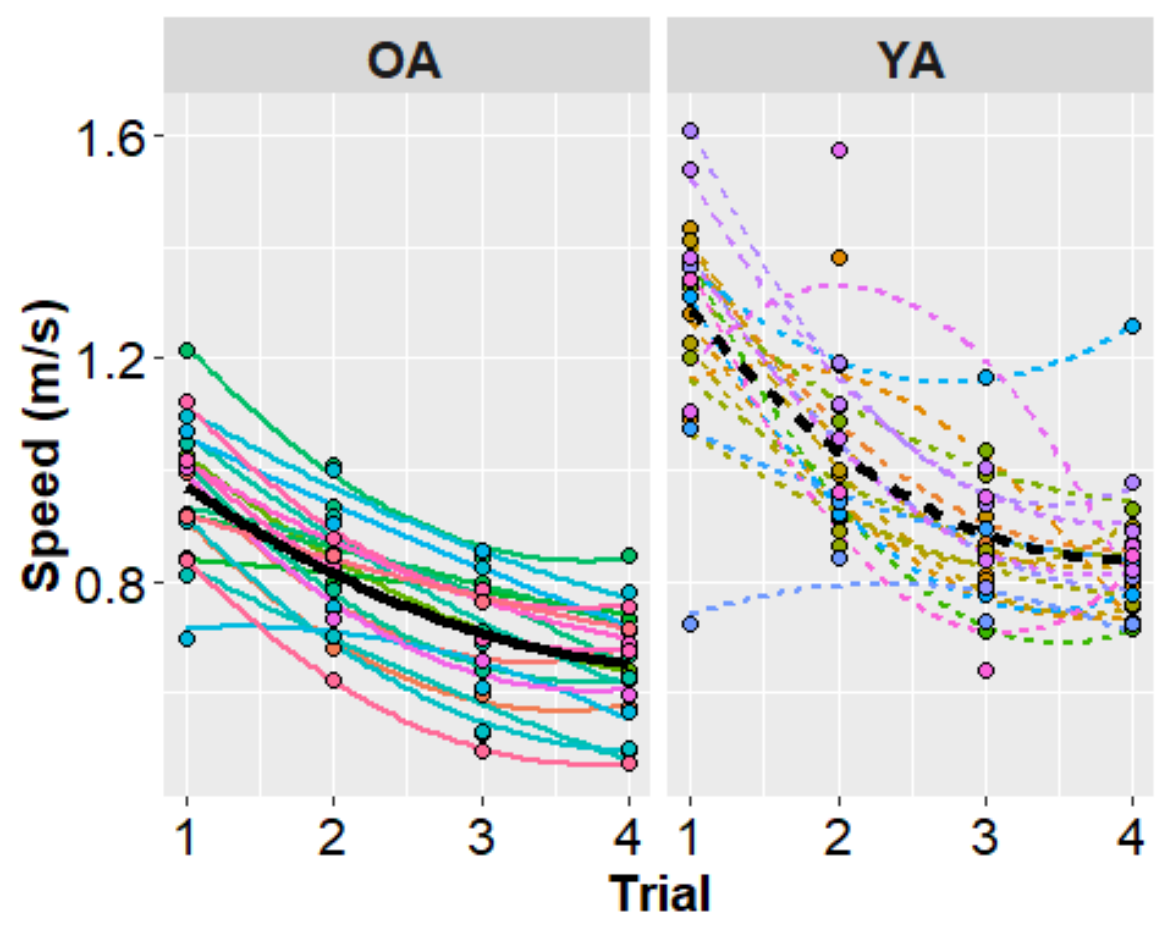
$$y_{ij} = \beta_0 + \beta_1(Time_{ij}) + U_{0j} + U_{1j}(Time_{ij}) + \epsilon_{ij}$$

$$y_{ij} = (\beta_0 + U_{0j}) + (\beta_1 + U_{1j})(Time_{ij}) + \epsilon_{ij}$$



LME - Regression

$$y_{ij} = (\beta_0 + U_{0j}) + (\beta_1 + U_{1j})(Time_{ij}) + \epsilon_{ij}$$



LME - Regression

~~$y_{ij} = (\beta_0 + U_{0j}) + (\beta_1 + U_{1j})(Time_{ij}) + \epsilon_{ij}$~~

$$y_{ij} = (\beta_0 + U_{0j}) + (\beta_1 + U_{1j})(Time_{ij}) + (\beta_2 + U_{2j})(Time_{ij}^2) + \epsilon_{ij}$$

The Mixed-Effects Model:

$$DATA = MODEL + ERROR$$

$$y_{ij} = (\beta_0 + U_{0j}) + (\beta_1 + U_{1j})(Time_{ij}) + \epsilon_{ij}$$

- The ***MODEL*** includes fixed effects and random effects.
- Fixed-Effects are the group-level β's, these effects parallel the traditional main-effects and interactions that you have probably encountered in other statistical analyses.
- Random-Effects are the participant-level U_j's that remove statistical dependency from our data. (This is bit of a simplification, but you can think of not including the appropriate random-effects like running a between-subjects ANOVA when you should be running a repeated-measures ANOVA.)
- The *ERRORS*, or more specifically Random Errors, are the difference between our *MODEL*'s predictions and the actual *DATA*, ϵ_{ij}'s.

Contrasting Mixed-Factorial ANOVA and Mixed-Effect Regression.

1. Modeling Outcomes Over Time

RM ANOVA

- Addresses questions about mean-differences, after between-subject variance is removed.
- Discrete timepoints are treated as categorical, with only the mean at each timepoint formally considered.

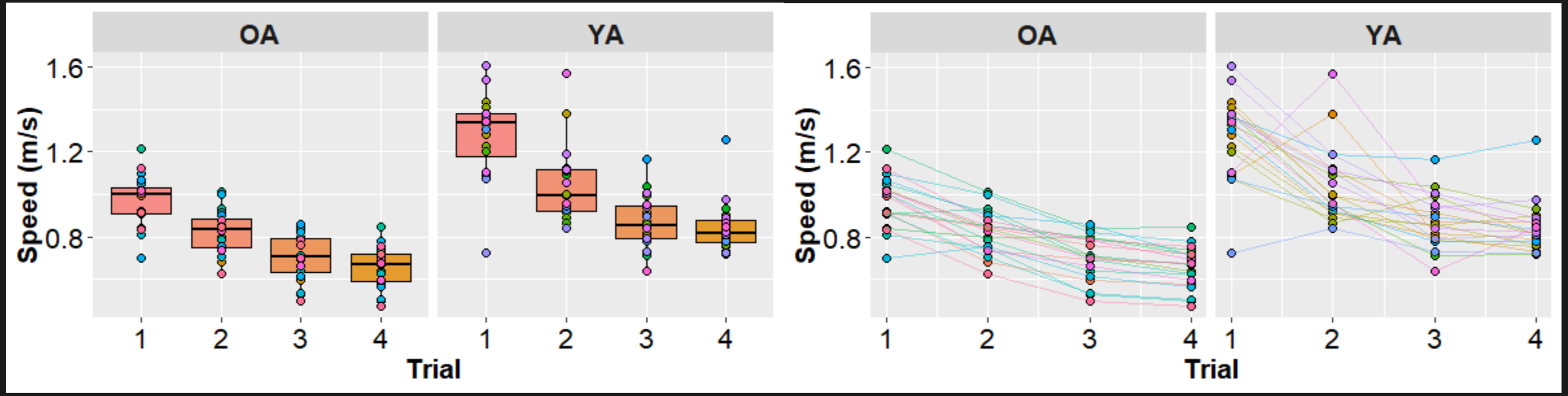
Mixed-Effect Regression

- Time is modeled explicitly as a trajectory for each individual.
- The shape of the trajectory is determined by fitting progressively more complex mathematical functions.

1. Modeling Outcomes Over Time

RM ANOVA

Mixed-Effect Regression



2. Variability in Time

RM ANOVA

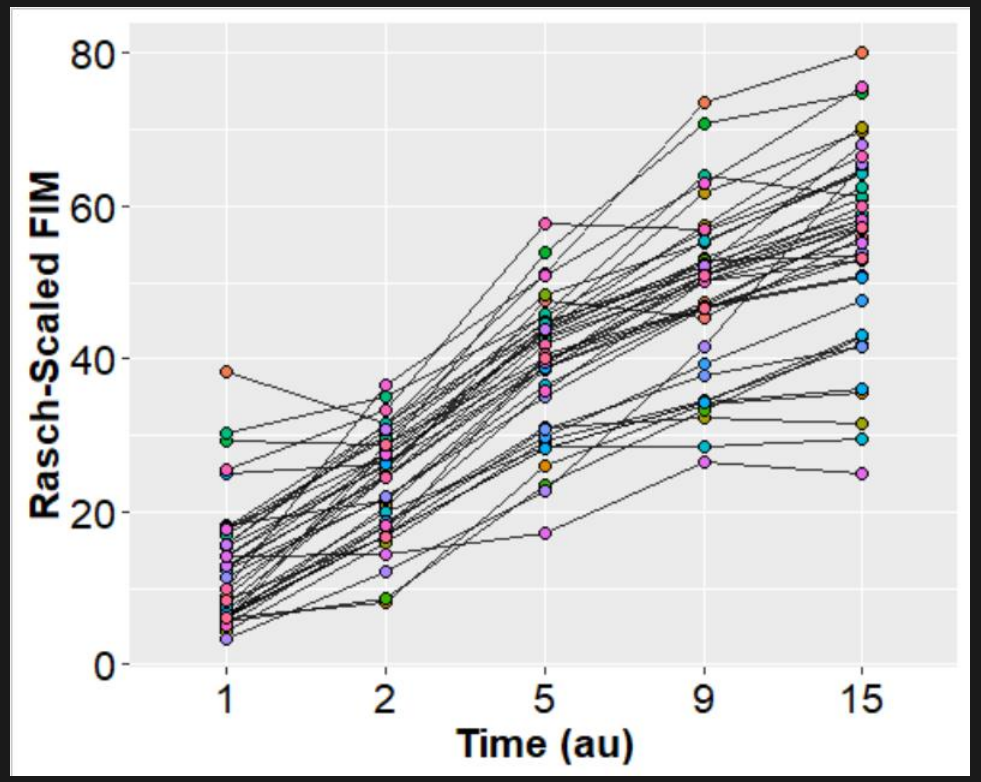
- Assumes common, identically timed data collections.
- This can lead to increased variability in the discrete time-points that is really due to variation in when data were collected.

Mixed-Effect Regression

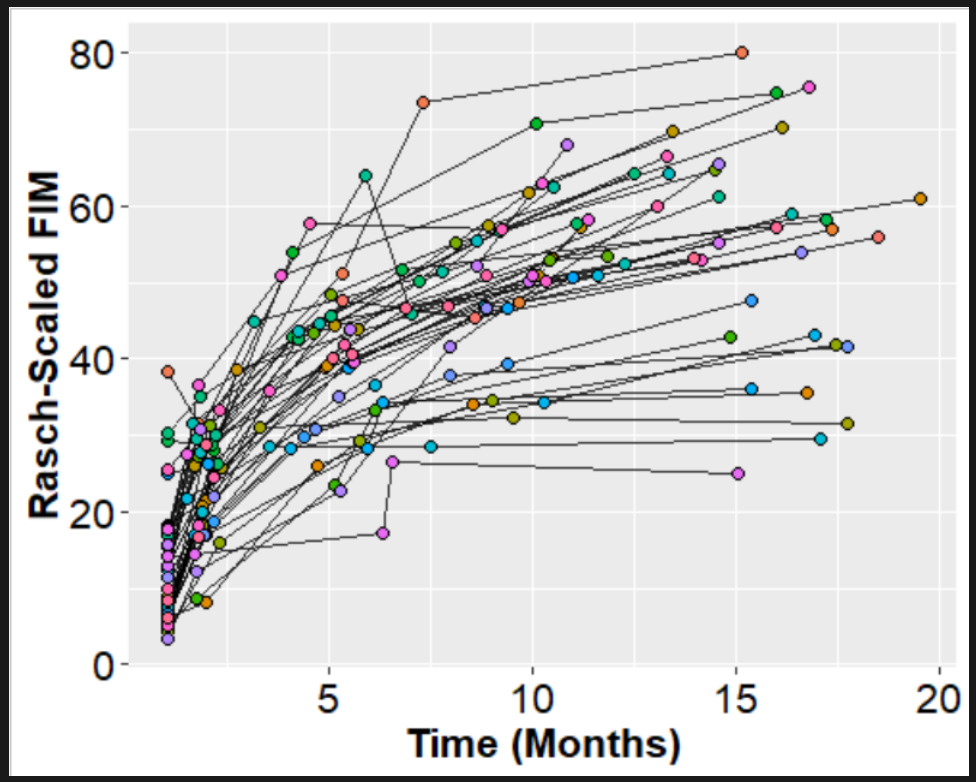
- Can accommodate variability in spacing of time points and in the actual timing of individual data collection.
- The model can also account for increased heterogeneity of the data over time, but residuals still need to be homogeneous.

2. Variability in Time

RM ANOVA



Mixed-Effect Regression



3. Data Missing on the Outcome

RM ANOVA

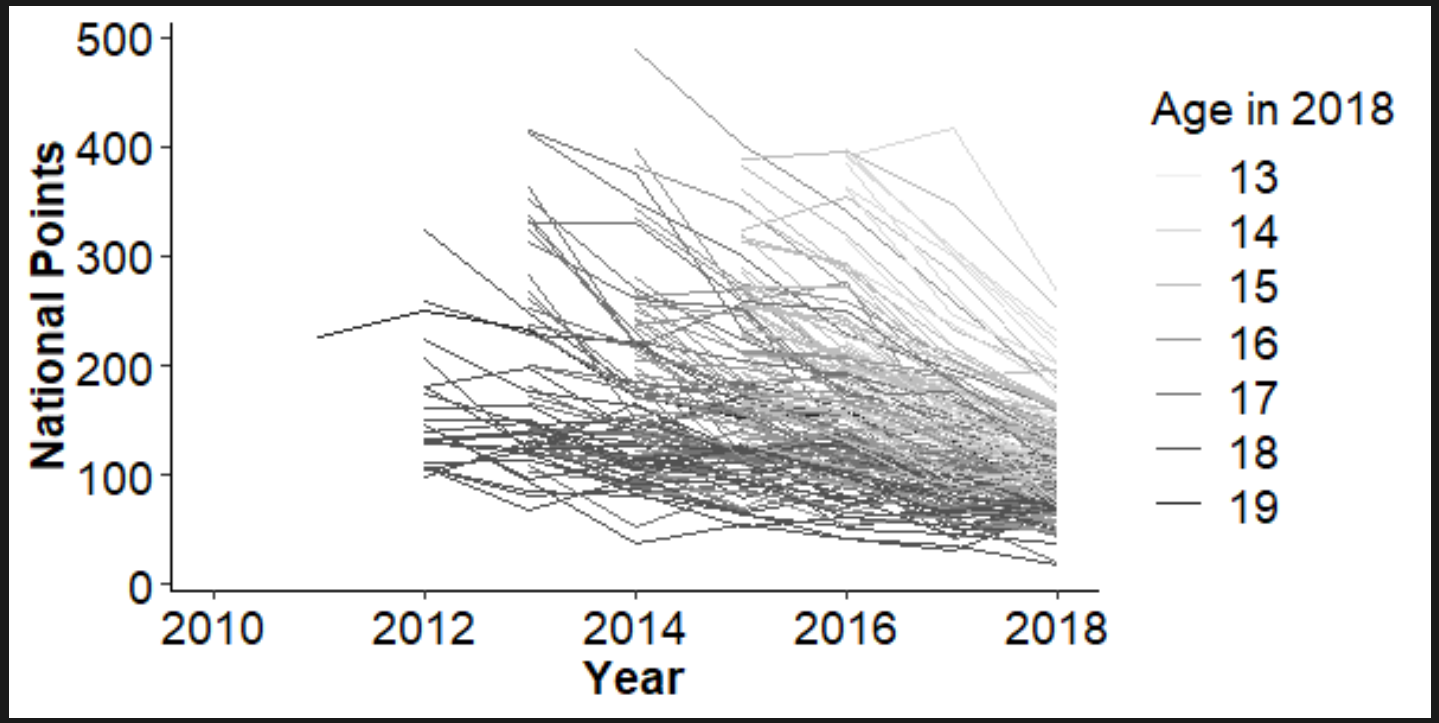
- Missing data generally cannot be accommodated.
- If data are *missing at random* (MAR), multiple imputation can be used for estimation.
- If data are *missing not at random* (MNAR), listwise deletion will reduce statistical power and potentially introduce bias.

Mixed-Effect Regression

- Data that are MAR can be accommodated without exclusion or imputation.
- Data that are MNAR can be fit, but factors associated with missingness need to identified and included in the model.
- If data are missing around key-moments, this may lead to poor model fit/selection.

3. Data Missing on the Outcome

Mixed-Effect Regression



[Lohse, Chen, & Kozlowski, 2020]

4. Data Missing on Explanatory Variables

RM ANOVA

- Missing between-person data in explanatory variables/covariates cannot be accommodated.
- Cases need to either dropped from the model or imputed.

Mixed-Effect Regression

- Missing between-person data in explanatory variables/covariates cannot be accommodated.
- Cases need to either dropped from the model or imputed.

4. Data Missing on Explanatory Variables

RM ANOVA

Mixed-Effect Regression

5. Including Covariates that Change over Time

RM ANOVA

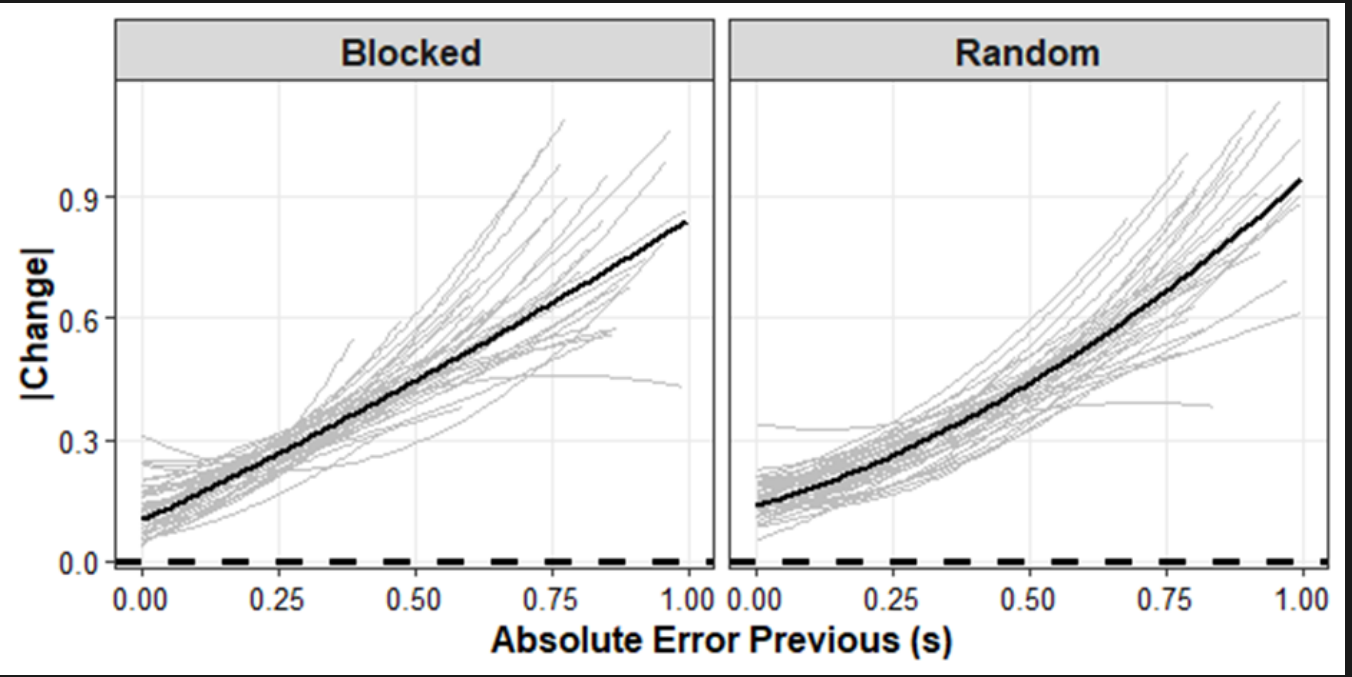
- Time-varying covariates cannot be included in an RM ANOVA model.

Mixed-Effect Regression

- Time varying covariates can be included, but you need to careful about collinearity and variance at both the between- and within-subject levels.

5. Including Covariates that Change over Time

Mixed-Effect Regression



- The correction on the current trial as a function of error on the previous trial, ***controlling for time in practice*** (i.e., trial number).

[Taylor, Thomas, Fawver, Williams, Lohse, in prep]

Conclusions

- Ultimately, there are many statistical reasons to prefer MER to RM ANOVA when it comes to analyzing longitudinal data.

- However, MER is computationally more complex and building these models requires a lot of judgment and care.
  - Often there are similar to ways to analyze the same data. Neither way might be "wrong" but they could both be different and we need to appreciate these subtle distinctions.

- As a guide, I highly recommend Jeff Long's *Longitudinal Data Analysis for the Behavioral Sciences Using R*.
  - Code and examples are provided in each chapter.
  - We can then adapt this code to our own problems.

Using R?

Open Source Software

- ***R*** is a programming language developed by Ross Ihaka and Robert Gentleman, based on the earlier programming language ***S***.
  - R is available at no cost to users through the GNU license and is maintained by an active community of programmers, statisticians, and researchers.

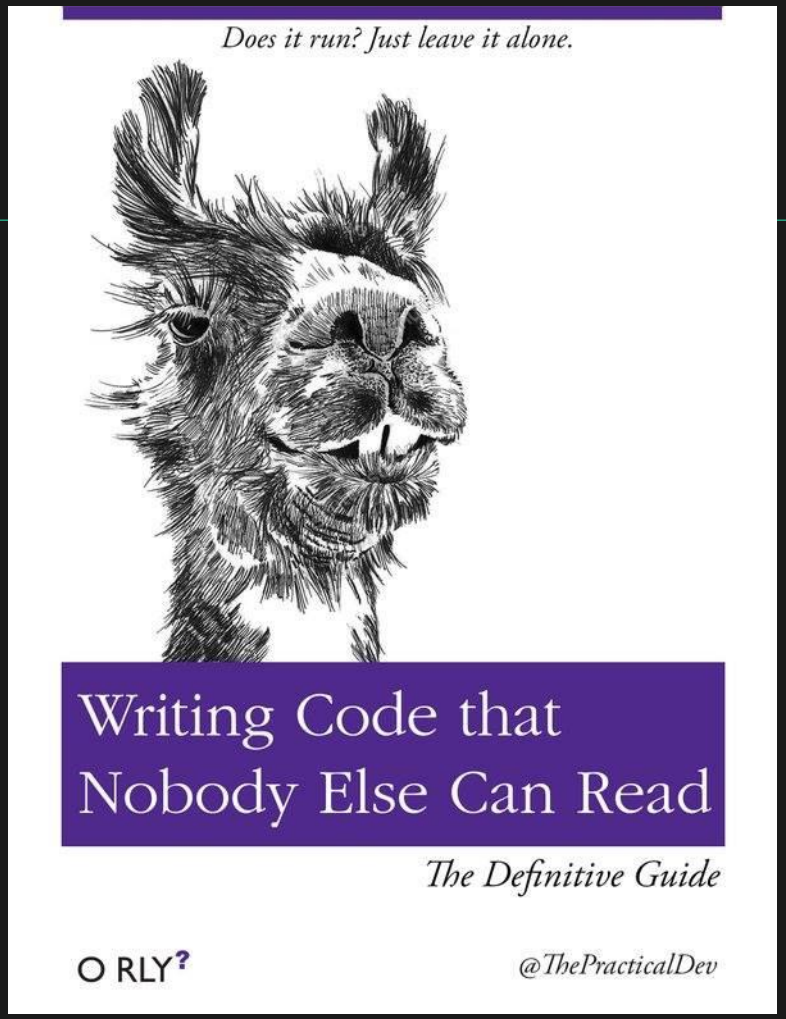
- **R Studio** is an integrated design environment (IDE) in which R can run.
  - R Studio makes it a lot easier to run R on your computer in an organized way.
  - If you are familiar with MatLab or Python IDEs (like Spyder), then the environment will feel familiar.

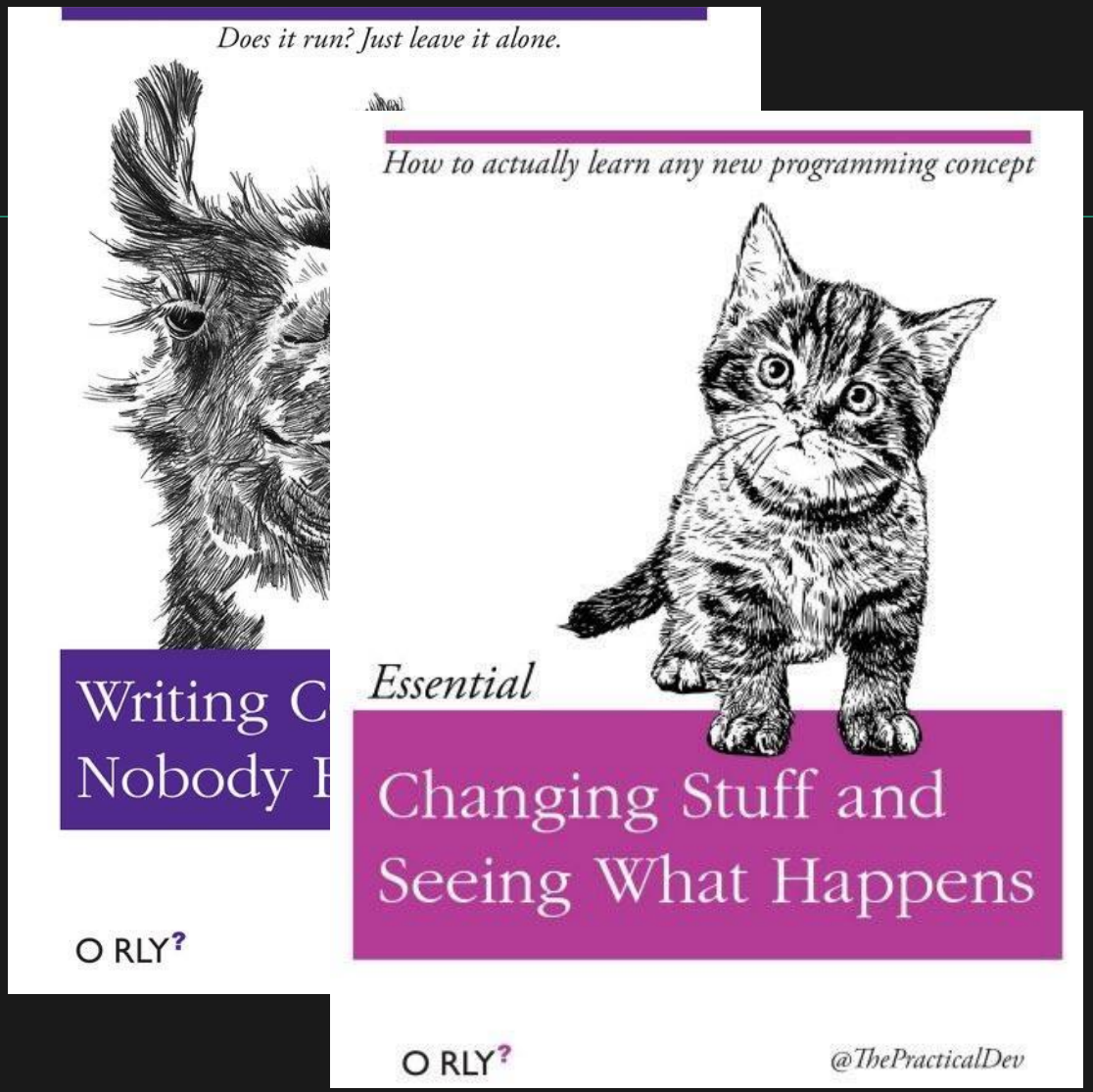
[The Legend of Zelda, 1986]

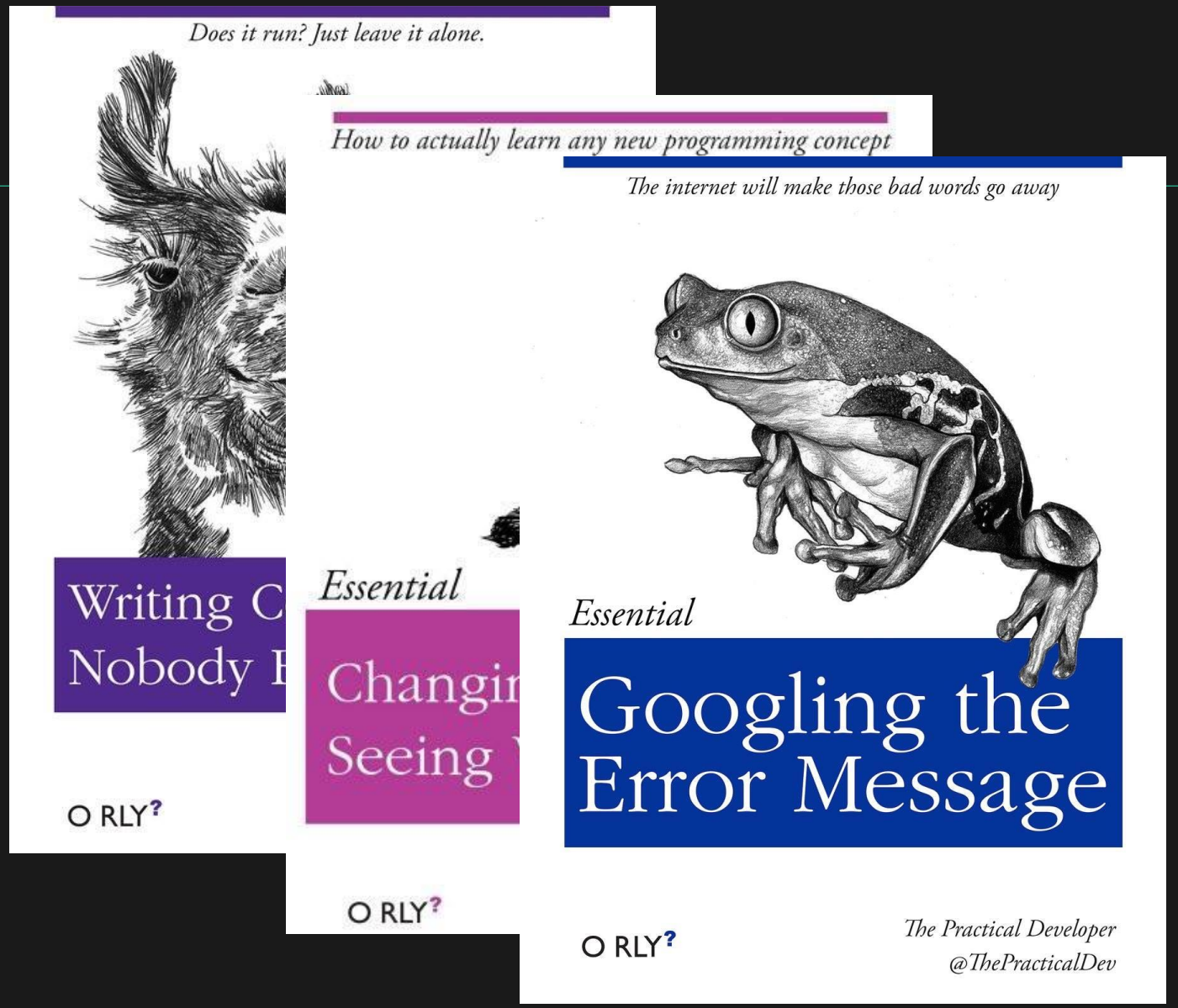
But programming is long journey…

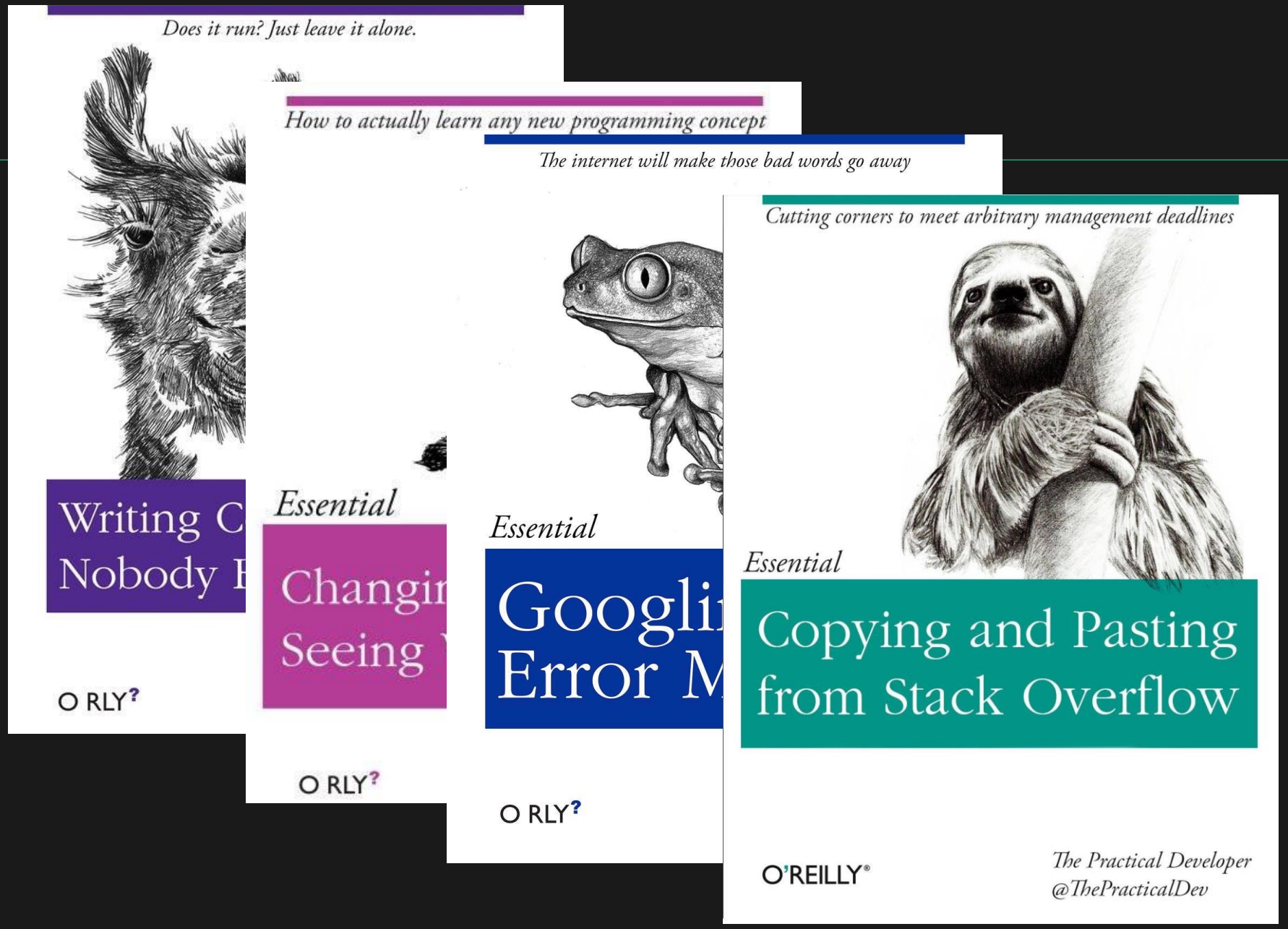
- Keep in mind that the increased power and flexibility of programming comes at a cost.
  - It takes a lot time and effort (that I think is well worth it in the end).

- It is also totally natural to feel lost and confused at times, that's a sign that you are learning.
  - Fortunately there are a lot of resources, classes, and active online communities to help you.
  - I've been doing this since 2007, but still…


Does it run? Just leave it alone.
Writing Code that Nobody Else Can Read
The Definitive Guide
O RLY?
@ThePracticalDev


Does it run? Just leave it alone.
Writing C
Nobody
O RLY?
How to actually learn any new programming concept
Essential
Changing Stuff and
Seeing What Happens
O RLY?
@ThePracticalDev


Does it run? Just leave it alone.
Writing C
Nobody
O RLY?
How to actually learn any new programming concept
Essential
Changi
Seeing
O RLY?
The internet will make those bad words go away
Essential
Googling the Error Message
O RLY?
The Practical Developer
@ThePracticalDev


Does it run? Just leave it alone.
Writing C
Nobody
O RLY?
How to actually learn any new programming concept
Essential
Changi
Seeing
O RLY?
The internet will make those bad words go away
Essential
Googli
Error M
O RLY?
Cutting corners to meet arbitrary management deadlines
Essential
Copying and Pasting from Stack Overflow
O'REILLY
The Practical Developer
@ThePracticalDev