# Managing Active Data: Take Good Notes

Jessica Logan, Ph.D.

Associate Director of Research

Crane Center for Early Childhood Research and Policy

# My job…

- Working on active data collection on 5 different projects right now

- Statistician - Analyzing data for publication from several projects where data collection is complete

# Exciting statistical methods?

Exploratory/Confirmatory Factor Analysis

Structural Equation Modeling

Latent Class Analysis

Multilevel Modeling

Growth Modeling

Quantile Regression

Behavior Genetics Models

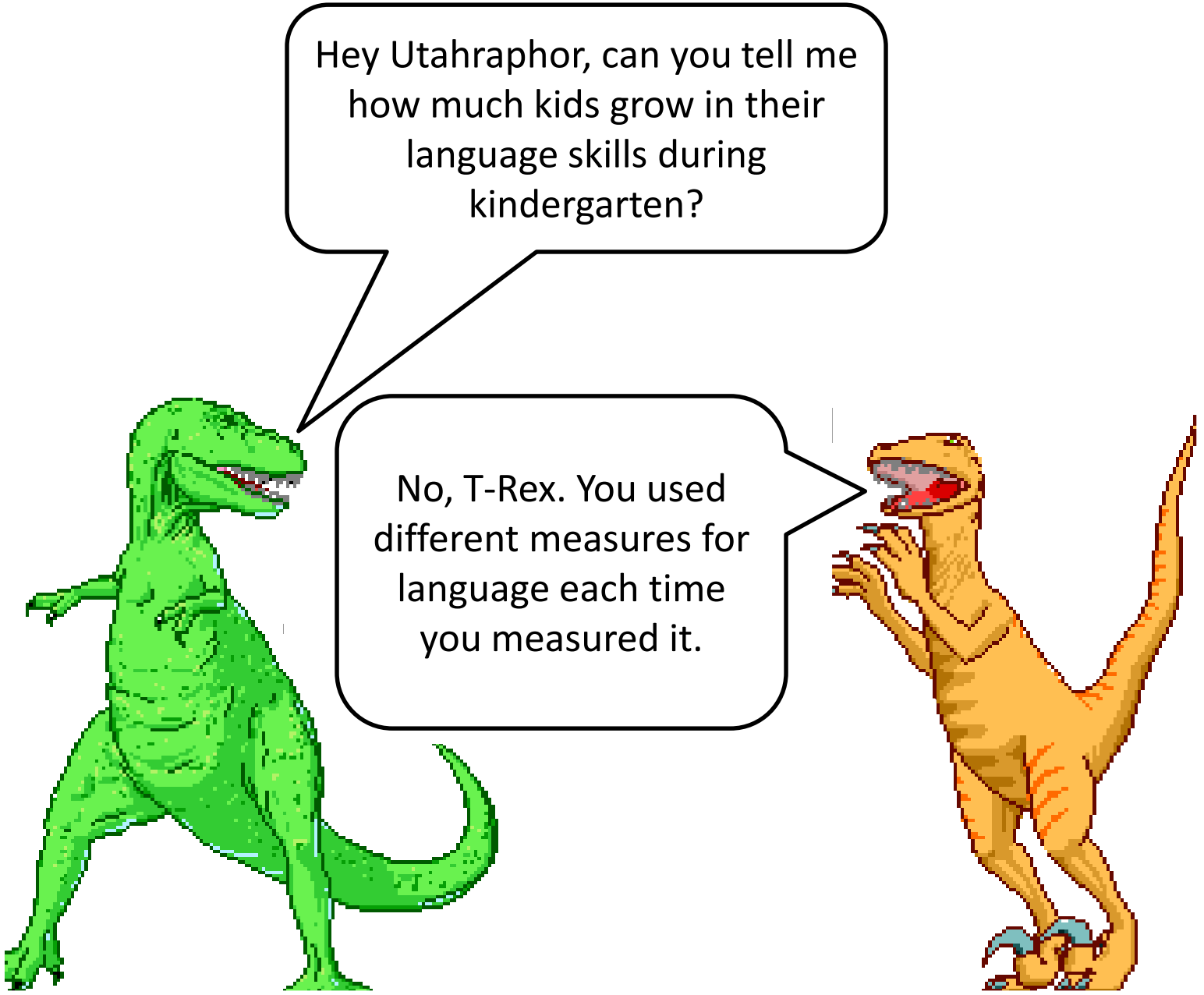
Non Linear Growth Models

Analysis of logistic data

Cross Classified Models

Latent Transition Analysis

# The Patient
# Is Already Dead




Respectfully parodied from "Dinosaur Comics" by Ryan North

# General Phases of *Active* Research

- Planning
  - Begin at the Beginning!
  - Storage *starts* here
- Active Data collection
  - Document everything
  - Clean your data
  - Share with collaborators
- Writing
  - Dataset naming
  - Calculating Variables

# Planning: Start at the Beginning

- “Know what you have” is much easier when you plan for what you’re going to get.
  - Decisions about how to *store* data start before I even submit my IRB

- We make protocols for everything:
  - ID number generation
  - Variable Naming
  - Value labels
- Code Books

# Junk in – Junk out.

Use established measures to assess the constructs of interest.

**WHAT ARE THE HEALTHIEST SITES ON THE INTERNET?**

| SITE/PLATFORM | CALORIES BURNED PER HOUR |
|---|---|
| Google | 0 |
| Yahoo! | 0 |
| Rotten Tomatoes | 0 |
| Twitter | 0 |
| Facebook | 0 |
| eBay | 0 |
| YouTube | 0 |
| Amazon.com | 0 |
| Pinterest | 0 |
| PayPal | 0 |
| ESPN.com | 0 |
| Instagram | 0 |
| Netflix | 0 |
| The Weather Channel | 0 |
| Zillow.com | 0 |
| Buzzfeed | 0 |

# Protocol Excerpt

## APPLE:Ohio Protocol

**RE**: **Protocol for following children into Kindergarten (collecting assessment data, obtaining state student identification numbers, and KRA-L data)**

**Authored Date**: February 22, 2011

**Latest Revision Date**: ~~March 29, 2011~~, ~~April 6, 2012~~, ~~November 30, 2012~~, December 27, 2012

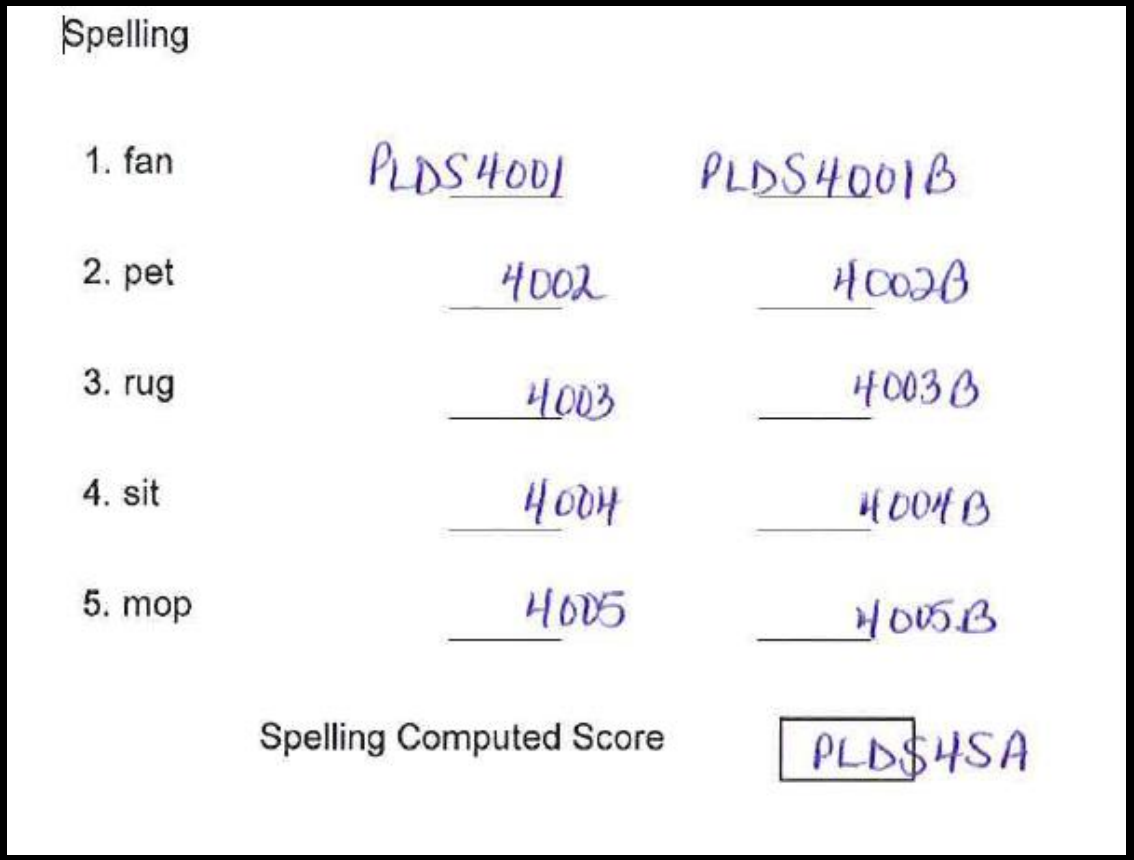
**Authored By**: Kristin Henkaline; Katie Mlod; Heather Doyle Fraser

**Related Protocol(s):**

**Decision**

The systematic approach used to follow children from preschool into Kindergarten and initiate contact with administrators and teachers will be uniform among all regions and across cohorts to the greatest extent possible.

# Good Codebook Excerpt



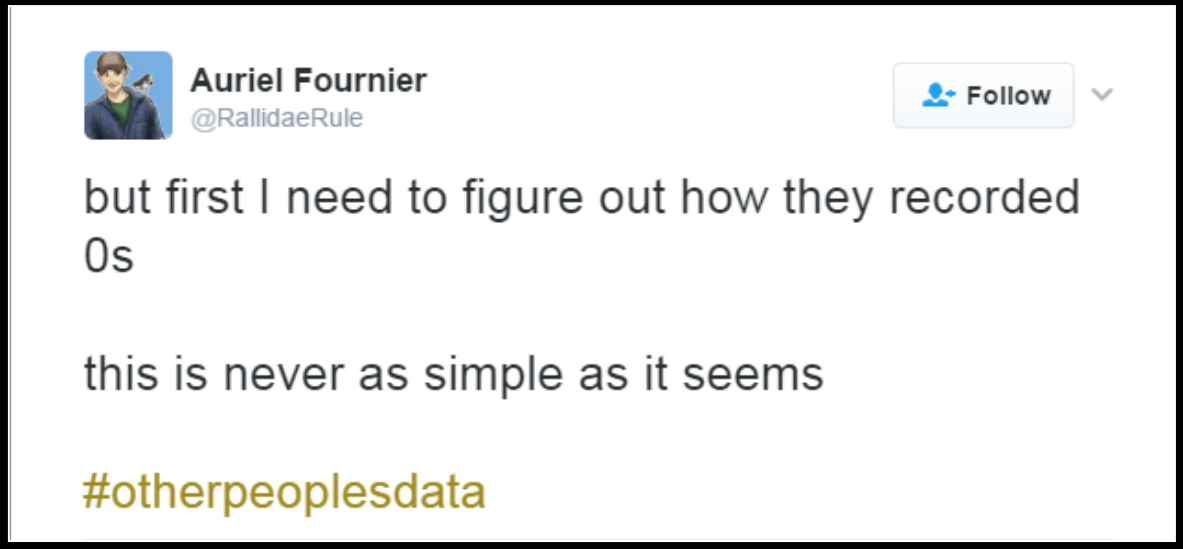
Spelling

| | | |
|---|---|---|
| 1. fan | PLDS4001 | PLDS4001B |
| 2. pet | 4002 | 4002B |
| 3. rug | 4003 | 4003B |
| 4. sit | 4004 | 4004B |
| 5. mop | 4005 | 4005B |

Spelling Computed Score PLDS4SA

# Other People's Data

- Think about someone else using your data later, and what they will need to know.



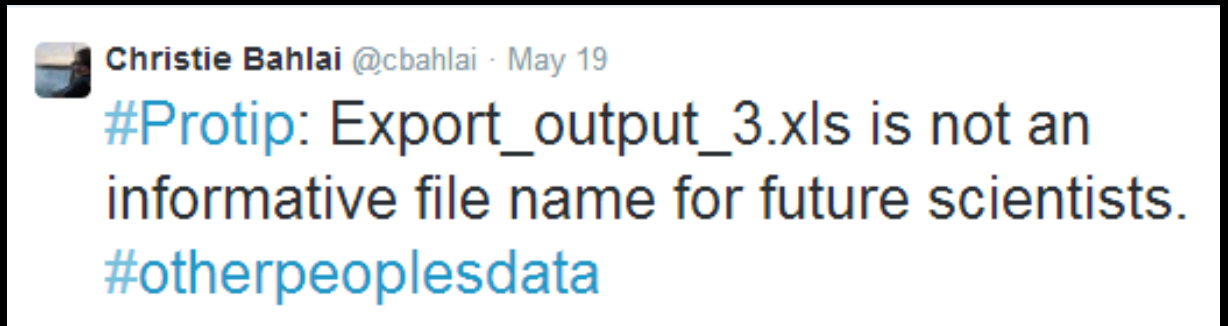
# Other People's Data

- Think about someone else using your data later, and what they will need to know.
- Every project has at least two data analysists – you and future you.

# Document. Everything.

It's hard to remember what you measured, when you measured it, or why you measured it.

Your best defense if a really good code book, explaining the study design, and what all of your variable names mean.



# Cleanliness is next to Godliness

Every data set has some problems. If you don’t think you have data problem, then you haven’t looked at it yet.

Data needs to be cleaned to be used meaningfully. Be flexible because data cleaning is hard.

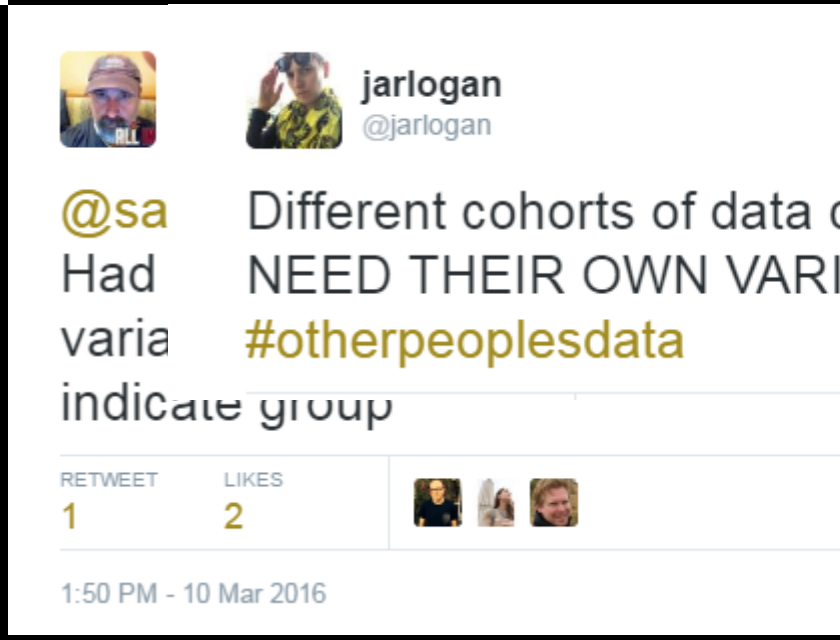
# Avoid manually-entered questions

**Noah Veltman** @veltman · Apr 25

Don't let anybody tell you that manually entered data is inconsistent.
pic.twitter.com/N3KtqQ1AEy

| | | | |
|---|---|---|---|
| H ISPANIC | HAIPANIC | HAISPANIC | HAPANIC |
| HASIAN | HASPANIC | HESPAINC | HHISP |
| HHISPANIC | HI | HIAP | HIAPANIC |
| HIAPNIC | HIAPSNIC | HIASPANIC | HICPANIC |
| HIDP | HIDPANIC | HIDPSNIC | HIDSPANIC |
| HIHSP | HIIIIIISPANIC | HIIIISP | HIIISP |
| HIISP | HIISPANIC | HIKSPANIC | HINDI |
| HINDU | HIOSP | HIOSPANIC | HIOSP[ANIC |
| HIP | HIPANIC | HIPANICX | HIPS |
| HIPSANIC | HIPSNAIC | HIPSNIC | HIPSPANIC |
| HIS | HIS0P | HIS;ANIC | HISANIC |
| HISANIC. | HISANID | HISAPANIC | HISAPNIC |
| HISDP | HISDPANIC | HISIPANIC | HISJPANIC |
| HISO | HISOANIC | HISOPANIC | HISP |
| HISP, | HISP- | HISP-ANIC | HISP. |
| HISP.\ | HISP3 | HISP; | HISPA |
| HISPA NIC | HISPAANIC | HISPABIC | HISPABNI |
| HISPABNIC | HISPACIC | HISPACNI | HISPAIC |

# Don’t make these mistakes!

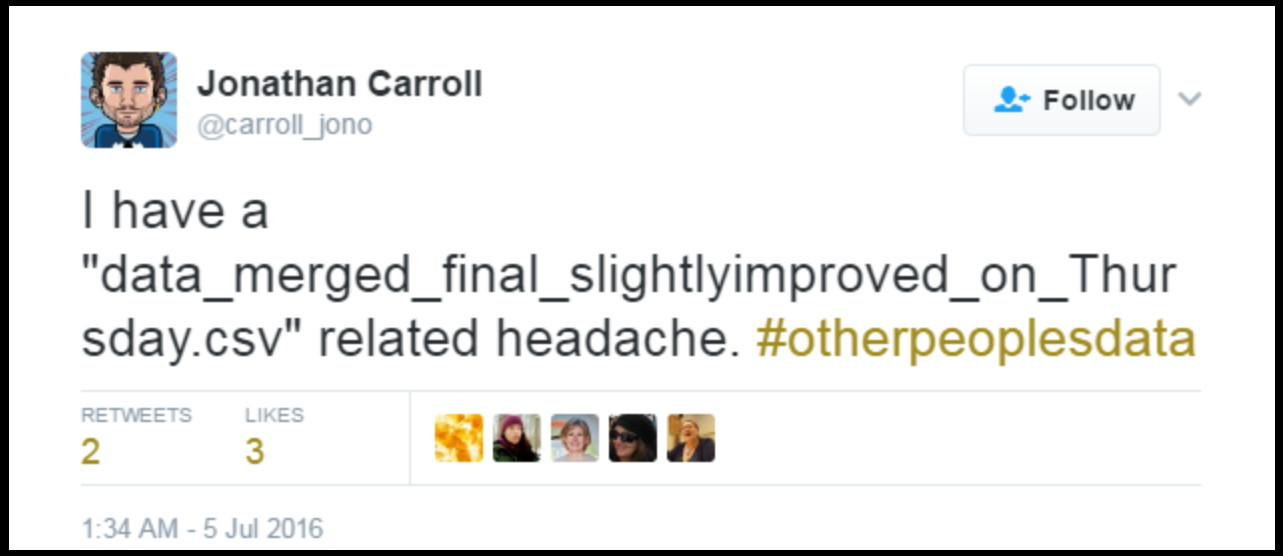
# Tidy Data

1) Make a tidy data set for sharing.
   - Make a codebook describing each variable and its values in the tidy data set.
   - Tidy data usually only contains most essential variables.

2) Also share the raw (unprocessed) data
   - Document how you went from the raw data to the tidy data.

# Ready to write

- You’re ready to write, now what?
- Versions are important:



# Ready to Write

- My process:
  - I always work from a *temporary copy* of the file created from code.
  - Code is annotated with what I do
  - After the paper is accepted, archive a copy of the final dataset.
  - Example:

# Sample code:

```
data a; set bb.mydatacp;
  if ageyears < 3.5 then AgeGroup = 0;
  if ageyears > 3.5 then Agegroup = 1;
  if gender = 1 then male = 1;
  if gender = 0 then male = 0;
  run;
```

```
*compute change scores for:
1) Sound Discrimination (single item composite, z-scored. Given to 3yo only)
2) Language composite (vocab, comprehension, communication)
4) Pre-literacy composite (4yo only - Letter, rhyme, SG)
;

data b; set a;
Lang_pre = mean(zvocpre, zcppre, zccpre);
Lit_pre = mean(zltpre, zrhpre, zsgnewpre);
Lang_post = mean(zvocpost, zcppost, zccpost);
Lit_post = mean(zltpost, zrhpost, zsgnewpost);
Lang_change = lang_post - lang_pre;
Lit_change = lit_post-lit_pre;
vocab_change = zvocpost - zvocpre;
run;
```

# Lab-wide data rules

- Data must be deleted off of shared drive as soon as active data collection ends.
  - Currently saved on external hard drives
- Funding:
  - Data processing is hard
  - Write in funding for data processing on grants!
  - Invest in a data manager.
- Why subjects were withdrawn:
  - Documented by project

# Data Management / Sharing Resources

- White et al., (2013). Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution 6(2): 1–10, 2013

- Tenopir, C. et al., (2011). Data sharing by scientists: practices and perceptions. *PloS one*, *6*(6).

- Hadley Wickham (@Hadleywickham) had.co.nz

- ICPSR's *Guide to Social Science Data Preparation and Archiving: Best practice throughout the data lifecycle.*

- Jeff Leek (@jtleek) jtleek.com

- Caitlin Rivers (@cmyeaton) Caitlinrivers.com

**Aly Baumgartner**
@kyrietree

Follow

My kingdom for some metadata!
#otherpeoplesdata

2:01 PM - 20 Jul 2016