Data visualization: From quality assurance to final publication.

Keith Lohse, PhD

Department of Health, Kinesiology, and Recreation

Department of Physical Therapy and Athletic Training

University of Utah

rehabinformatics@gmail.com

@keith_lohse

https://github.com/keithlohse/ASNR

HEALTH
UNIVERSITY OF UTAH

Road Map

- General principles of data visualization.
  - Save yourself a lot of time with reproducible, code based graphics.
- Visualizing discrete data.
  - Special considerations for repeated measures data.
- Visualizing continuous data.
  - Special considerations for time-series data.

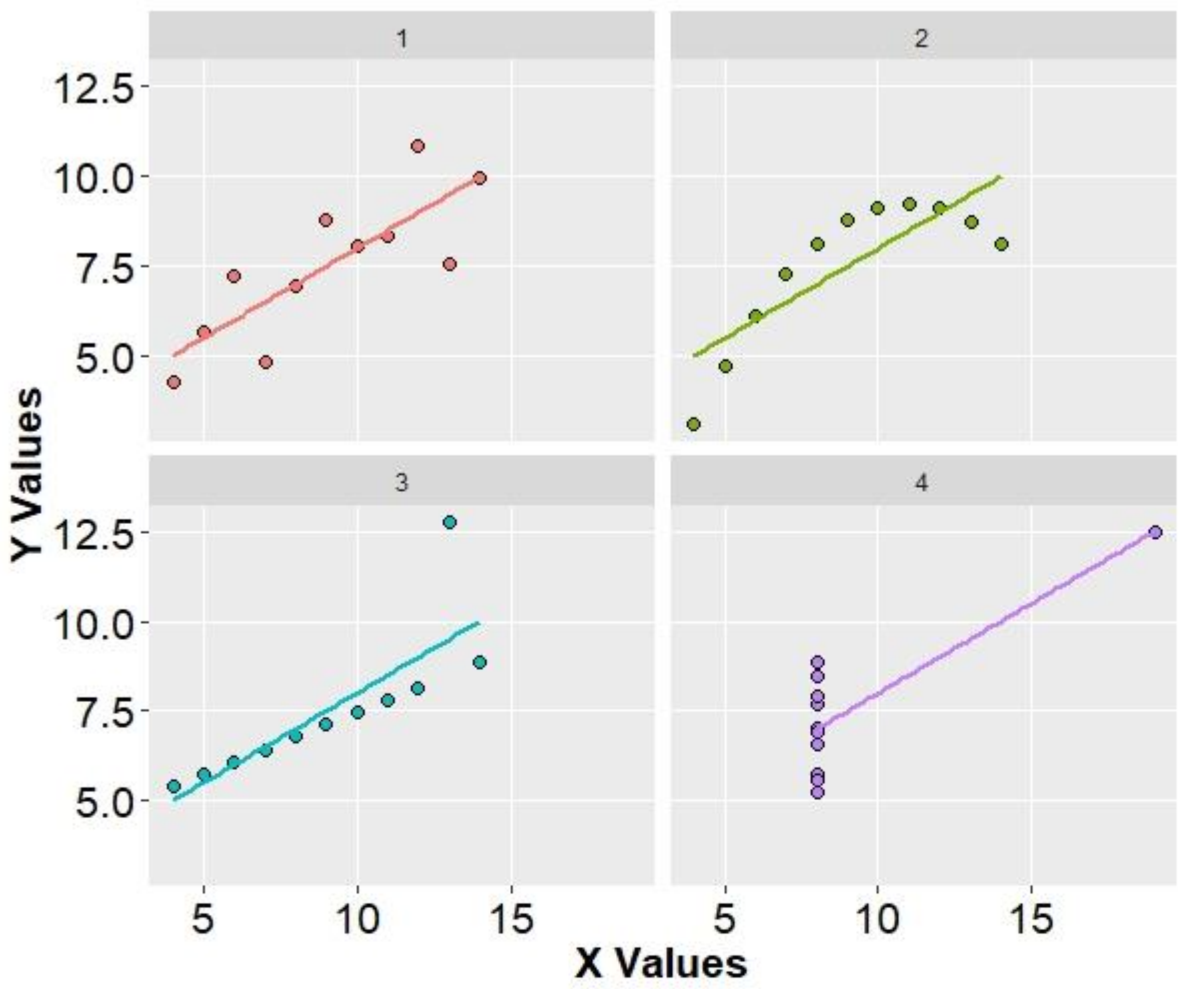
General Principles of Data Visualization

Why is visualizing data so important?

- Let’s say I run an analysis in my stats program regressing Y onto X.
  - The **Intercept** is 3.00 and statistically significant.
  - The **Slope** is 0.50 and statistically significant.

- Is it fair to assume that a 1-unit change in X leads to a 0.5-unit change in Y in an approximately linear relationship?

Why is visualizing data so important?



- All of these datasets have identical slopes, intercepts, and p-values.
- **<u>Model 1</u>** is the only one that meets all the assumptions of linear regression.
  - 2 = nonlinear.
  - 3 = non-constant error variance.
  - 4 = extreme leverage (Cook's distance).

[Anscombe, 1974]

IMO, Good Visualizations Should…

1. **Put your data on the table.** Show "person-level" data and "group-level" statistics to paint the complete picture.

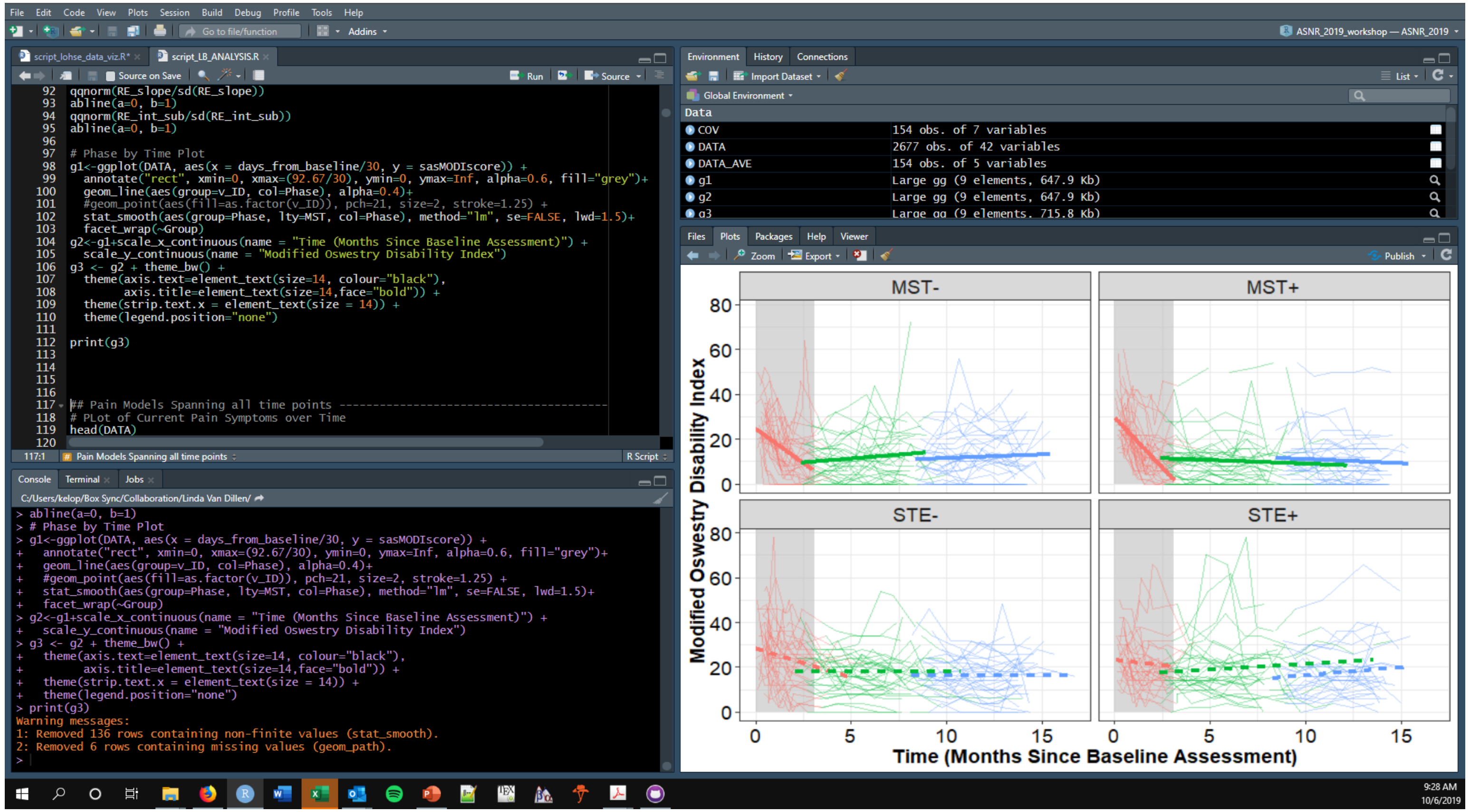
2. **Reduce unnecessary complexity**. The question motivating a visualization should be clear, as should the answer.

3. **Have correspondence to your analysis**. I can "see" your result without understanding the finer points of your analysis. The inferential statistics are just there to "back it up".

4. **Accept uncertainty.** The data should speak for itself and visualizations should accurately reflect the data above all else.

[But see Healy, 2018; Tufte, 2001; Tukey, 1980; Wickham & Grolemund, 2017]

Reproducible, code based graphics.



[I create most of my graphics in ‘ggplot2’ using R. Any post-processing I do in the Gnu Image Manipulation Program.]

- Ultimately, any way you create your visuals is fine as long as your visuals are accurate and informative.
- **<u>But</u>**, code-based approaches have a lot of advantages in terms of efficiency **<u>and</u>** reproducibility.

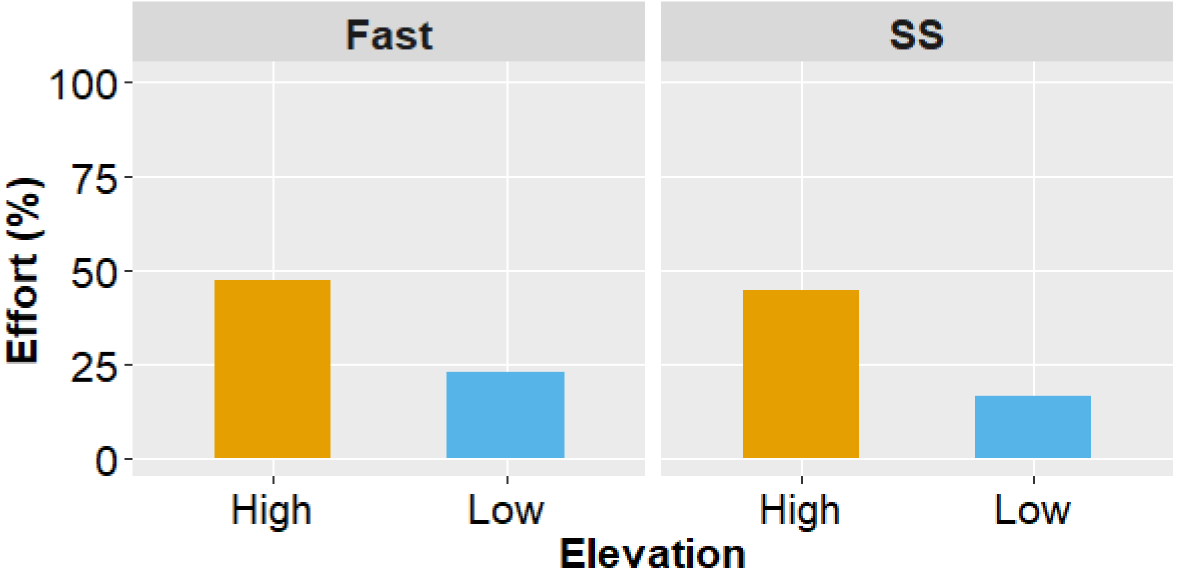
Visualizing Discrete Data

HEALTH
UNIVERSITY OF UTAH

Consider a 2 x 2 factorial design.

- Participants walked at fast or selected speeds at virtual high or low heights.

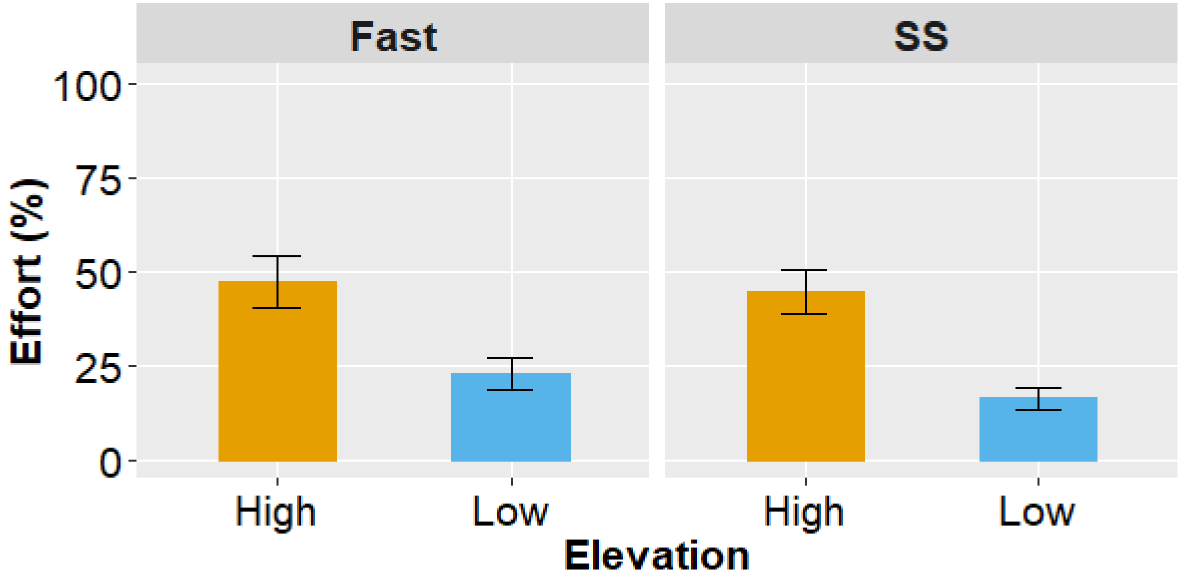
- Among other things, we collected psychological perceptions of effort across the different trials.



1. ~~No measure of central tendency.~~
2. No measures of spread.
3. No information about sample size.
4. No information about the correlation between measures.

[Raffegeau et al., *Under Review*]

Consider a 2 x 2 factorial design.



1. ~~No measure of central tendency.~~
2. ~~No measures of spread.~~
3. No information about sample size.
4. No information about the correlation between measures.

- Adding standard errors is arguably better **but…**
  - These are between subjects standard errors, and our manipulation occurred within subjects. [Loftus & Masson, 1994; Morey, 2008]
  - The standard error confounds standard deviation with sample size, $se = s/\sqrt{n}$.

[Raffegeau et al., *Under Review*]

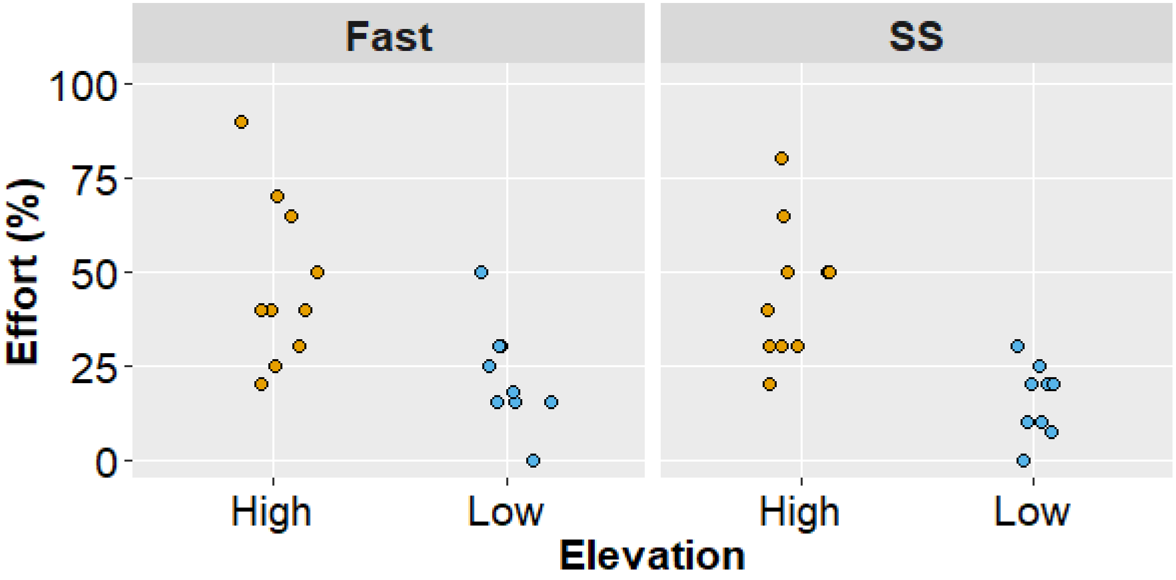
Consider a 2 x 2 factorial design.

- What if we just plot all of the data?

- To paraphrase Karl Pearson, we have now put our "data on the table", **but** something has also been lost. [Stigler, 2002]

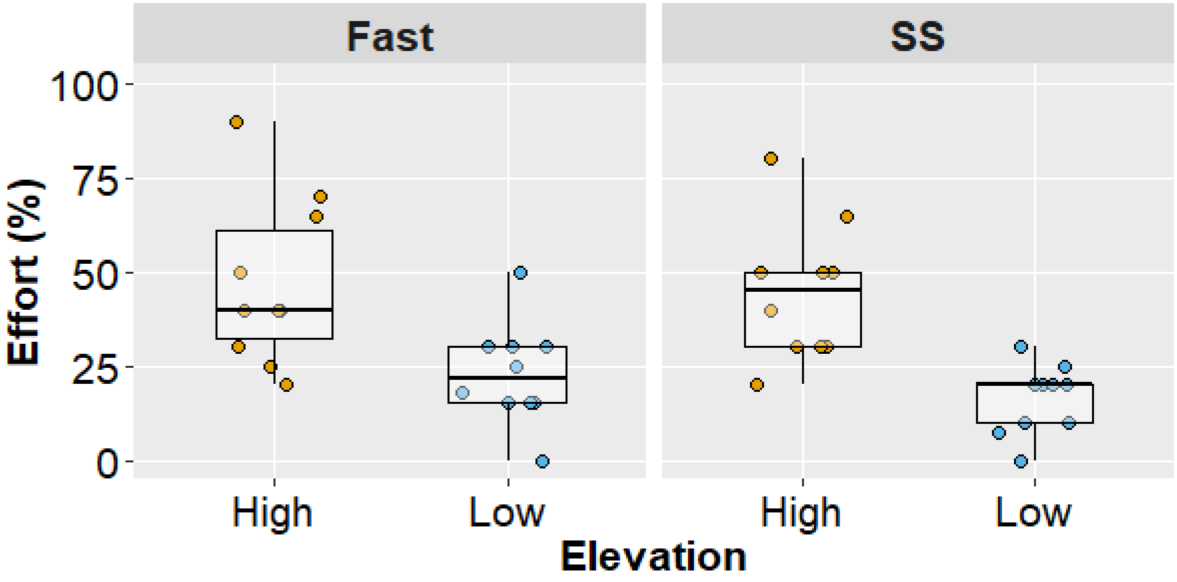
    - Measures of central tendency are critical to our **statistical inference**.

    - We have gained a **rich description** of our sample, but lost the correspondence to our analysis.



1. No measure of central tendency.
2. ~~No measures of spread.~~
3. ~~No information about sample size.~~
4. No information about the correlation between measures.

Consider a 2 x 2 factorial design.



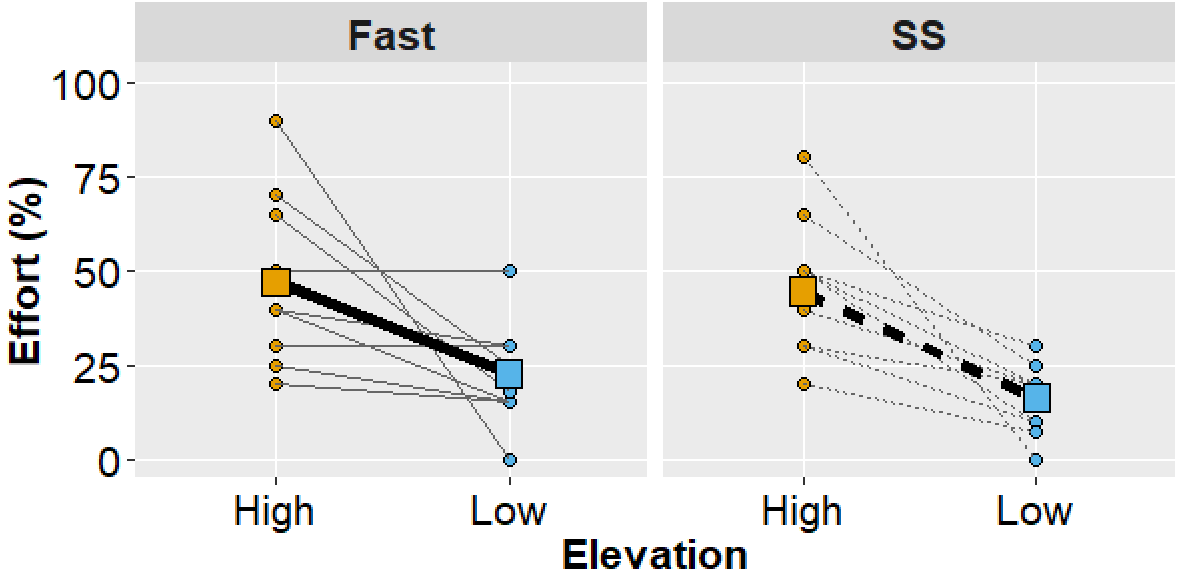
1. ~~No measure of central tendency.~~
2. ~~No measures of spread.~~
3. ~~No information about sample size.~~
4. No information about the correlation between measures.

- Now this is good! By playing with overlay and transparency, **group-level** statistics are emphasized (for inference).
- But all of the **participant-level** data are also visible (for description/assumptions).
  - One issue is that boxplots show medians, but most of our inferential statistics are based on means.
  - This isn't bad, but potentially lacks correspondence between visualization and analysis.
- In a within-subject design, we might want to know which points belong to whom.

[Raffegeau et al., *Under Review*]

Consider a 2 x 2 factorial design.

- We can overlay the means for each condition on top of the data for each condition.
- By connecting the dots, we can also provide information about the correlation between conditions.



1. ~~No measure of central tendency.~~
2. ~~No measures of spread.~~
3. ~~No information about sample size.~~
4. ~~No information about the correlation between measures.~~

[Raffegeau et al., *Under Review*]

Visualizing Continuous Data

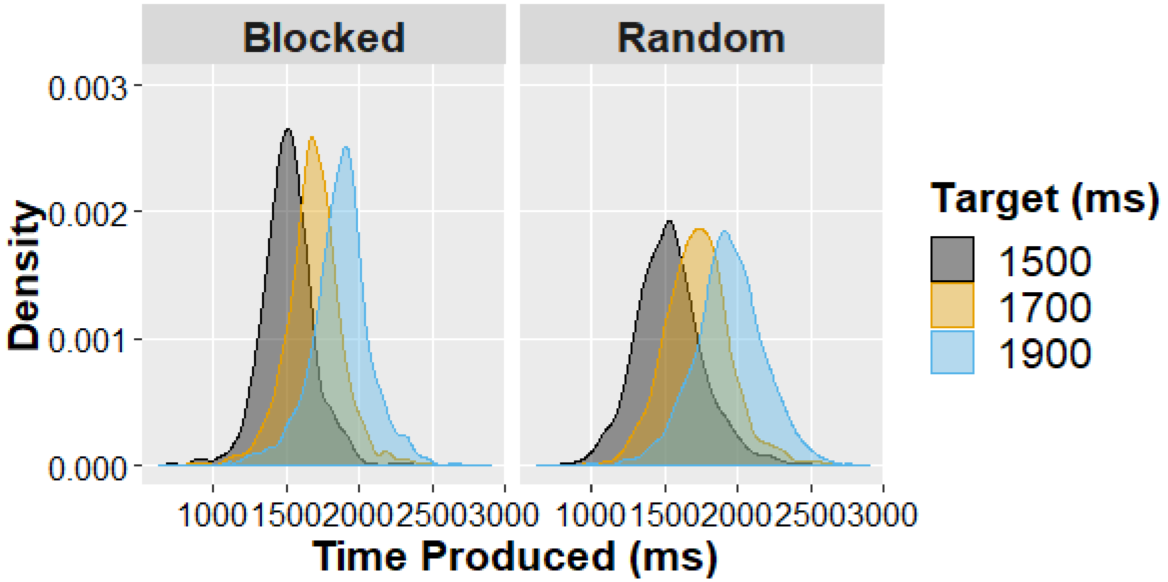
HEALTH
UNIVERSITY OF UTAH

A Two Group Longitudinal Study

- In a variable versus blocked practice experiment, participants learned to estimate different intervals of time in either a **blocked** order or a **random** order.
  - 1500 ms
  - 1700 ms
  - 1900 ms

- Focusing on response distributions to study '**confusability**'.

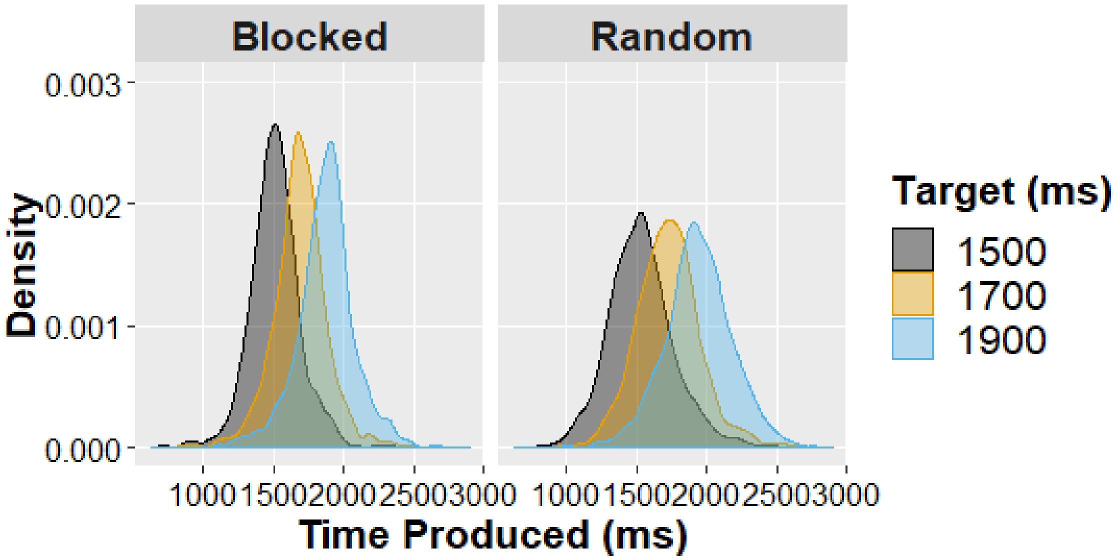
Acquisition



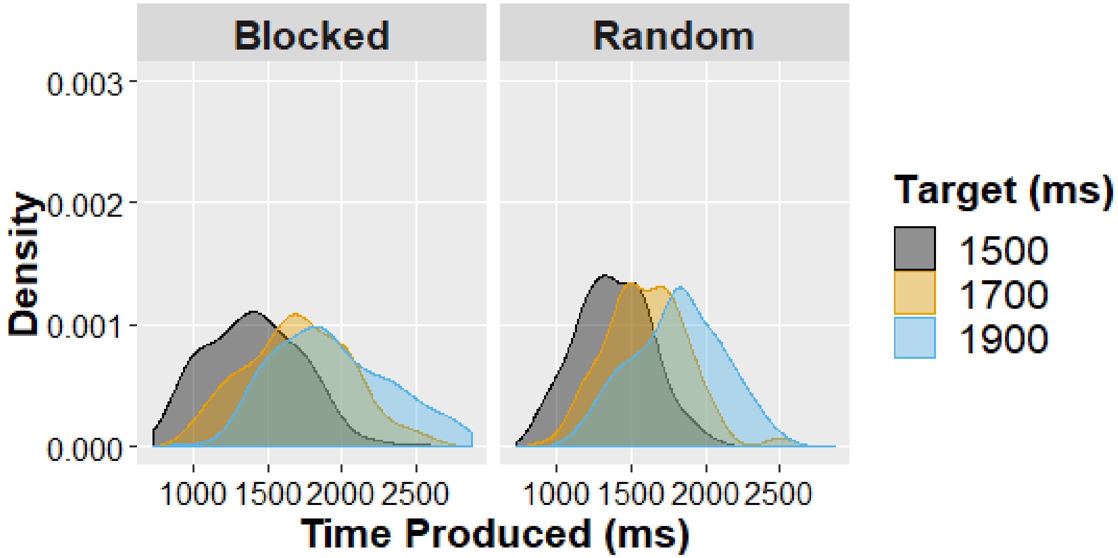
[Thomas et al., *In Preparation*]

A Two Group Longitudinal Study

Acquisition



Delayed Retention Test

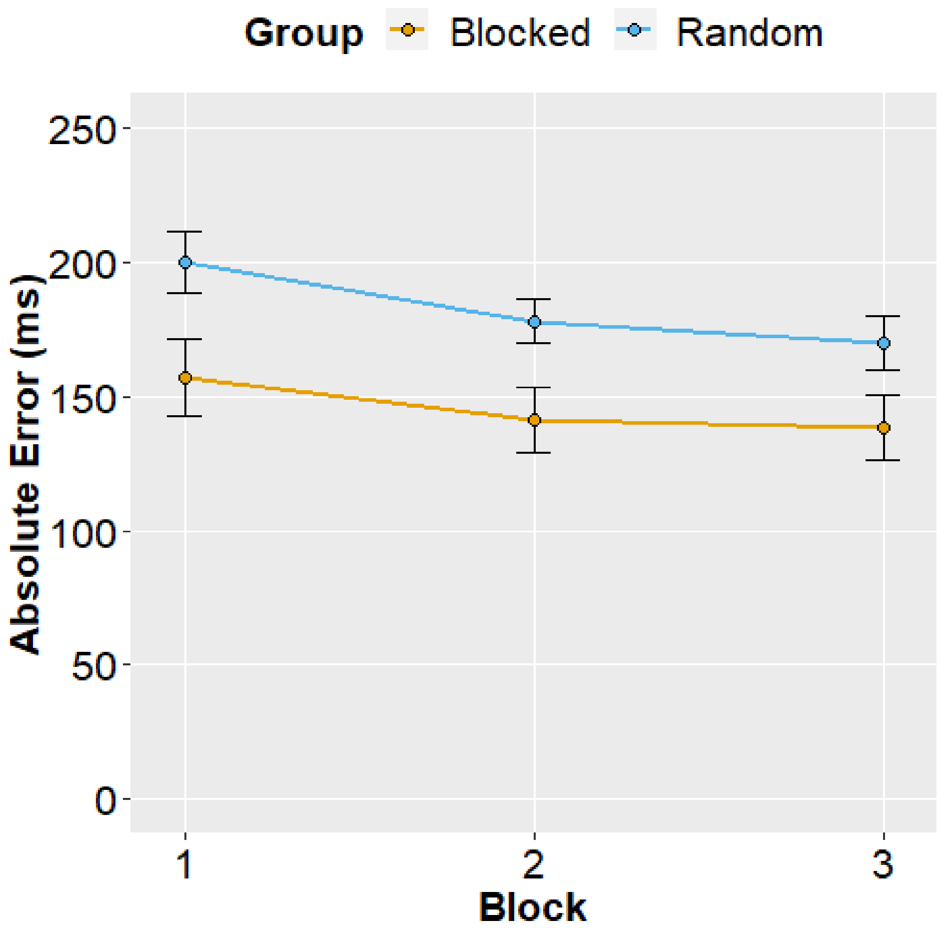


[Fall Out, InterPlay Ent.]

It's a slightly different way of looking at it, but we replicate the traditional contextual interference effect.

[Thomas et al., *In Preparation*]

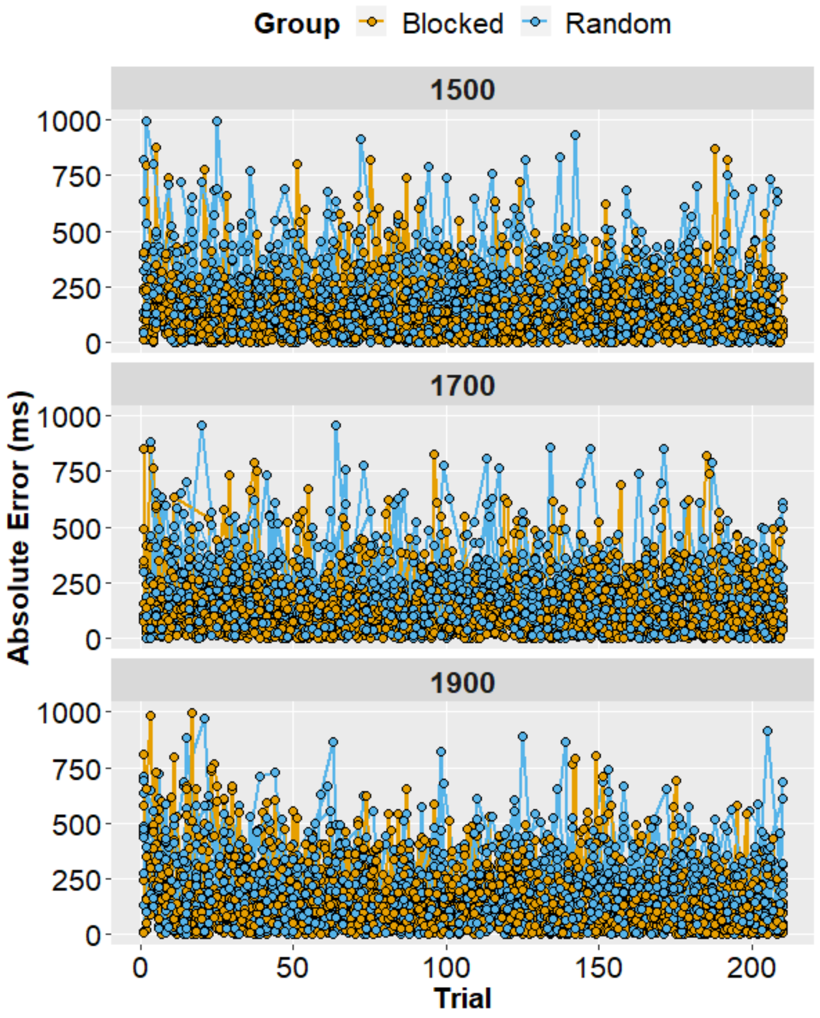
A Two Group Longitudinal Study



- But learning is a continuous process that happens over time.
- In a more “classic” plot, we might average across trials and targets to look at average error in each block of practice.

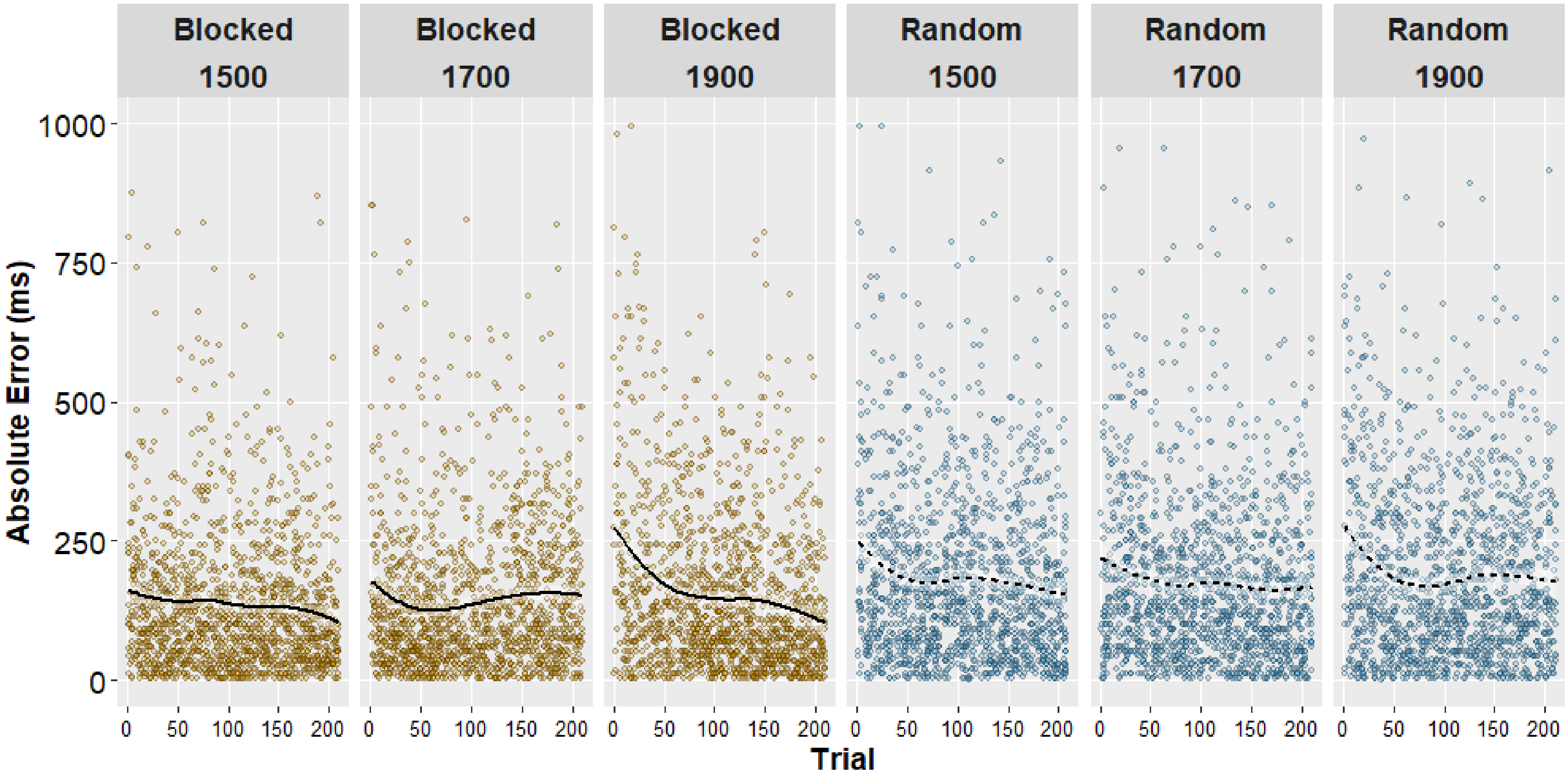
A Two Group Longitudinal Study

- What if I want to **see all of the data**?
- It's too much!
- We have so many acquisition trials that it makes identifying performance curves almost impossible.



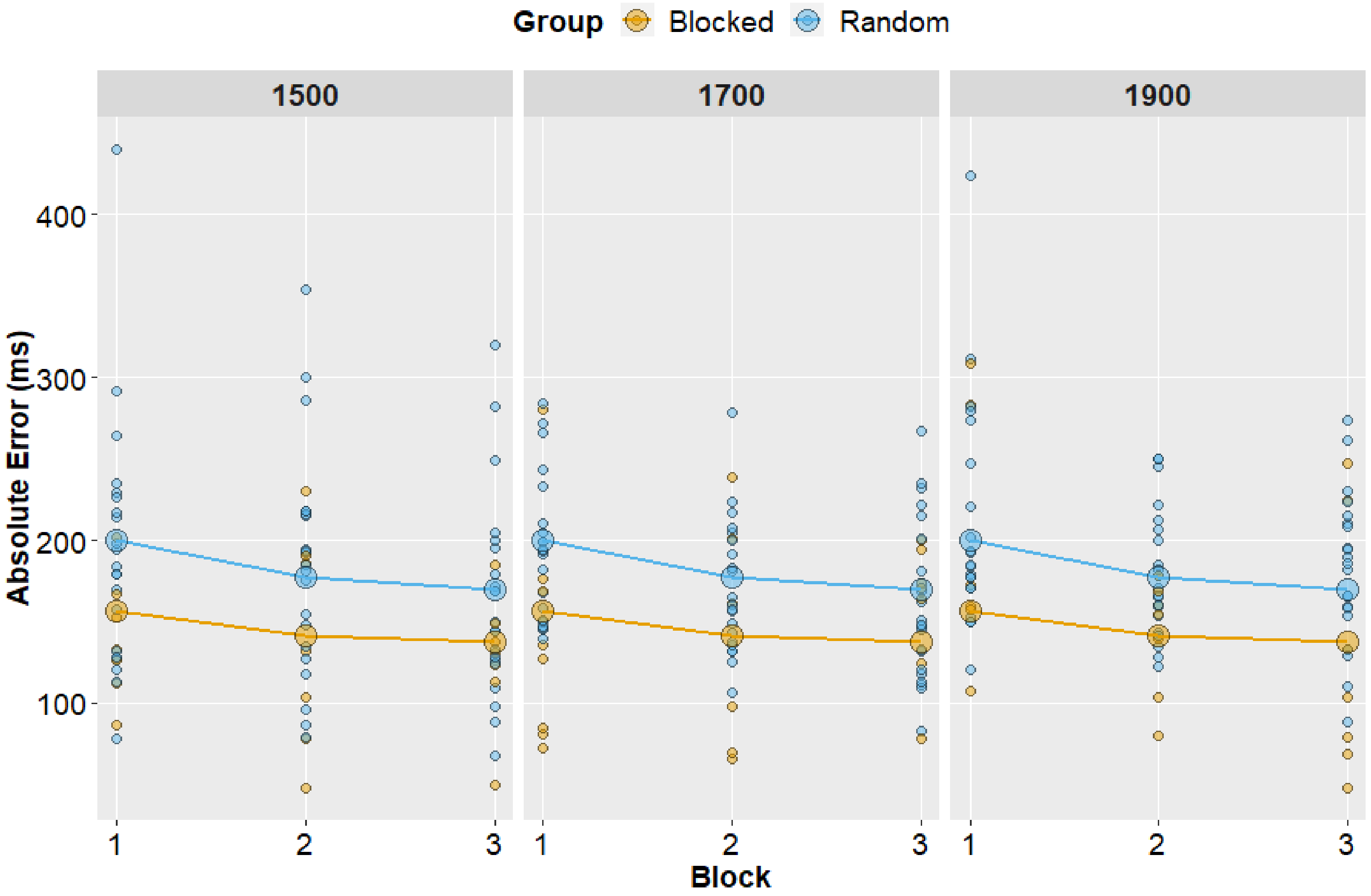
A Two Group Longitudinal Study

- However, by playing with aesthetic features such as spacing, transparency, and line type, we can make the overall pattern much more interpretable.



A Two Group Longitudinal Study

- With a little bit of aggregating, we might be able to find a happier "middle ground".



Thank You!

For more information:

rehabinformatics@gmail.com

@keith_lohse

https://github.com/keithlohse/ASNR

Special Thanks:

- Lei Liew and James Finley
- The American Society for Neurorehabilitation
- Tiphanie Raffegeau and Joe Thomas for sharing parts of their data.