Arbitrary choices, arbitrary results: A multiverse analysis of L2 reaction time data

Ryo Maie

Michigan State University

American Association for Applied Linguistics 2021

Warm-up

Suppose you have a research question:

Do L2 English speakers process lexical items more slowly than L1 speakers?

- You set up a lexical decision task where participants judged whether the word presented on a screen was a word or a nonword.
- You collected the response time as the depedent variable.

Warm-up

Prior to statistical analyses, you have to make the following decisions:

1. Should all responses be included or only those that are correct?
   - **all or only correct**
2. Should participants be excluded if they have an unacceptably low accuracy rate. What criterion must they reach?
   - **80%, 85%, or 90%**

Warm-up

Prior to statistical analyses, you have to make the following decisions:

3. Should data points be excluded if they are abnormally faster or slower than normally expected? At what point should you consider them to be unacceptable?
    - **200ms, 250ms, or 300ms, for the lower bound**
    - **1500ms, 2000ms, or 2500ms for the upper bound**
4. Should data points be excluded if they deviate from the overall trend of individual participants? What range should you adopt as an acceptable range of variation?
    - **2SD, 2.5SD, or 3SD**

Flexibility in data processing

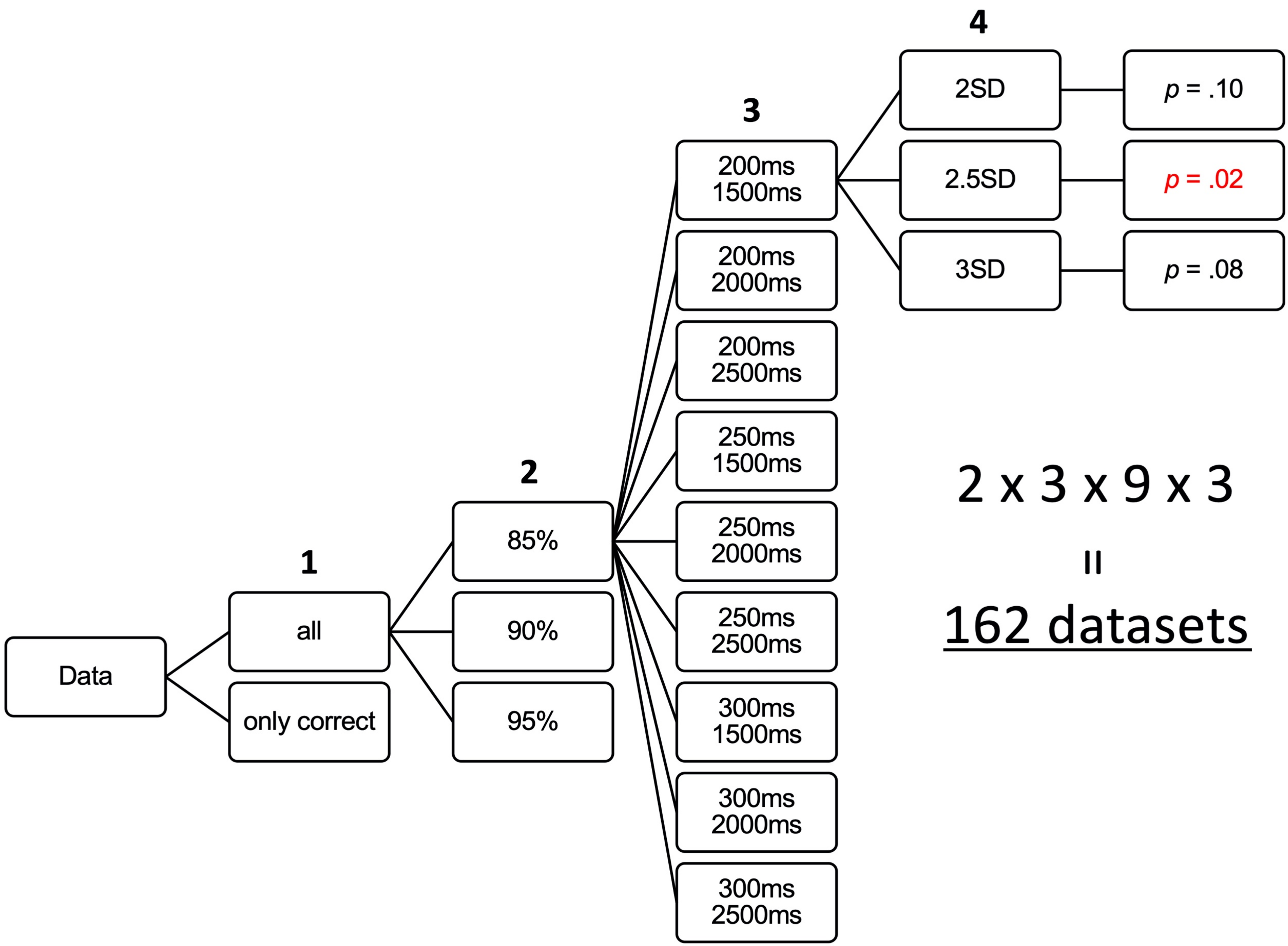
Out of these four decisions, how many were you able to justify with any substantive reasoning?

- Some may be empirically justiable: word recognition $<$ 300ms is unlikely
- Others are equally (un)justifiable and you can choose any of them
- **There is always an article to cite to defend any of your decisions!!!**

Flexibility in data processing

How do you decide on:

- 2SD, 2.5SD, or 3SD for outliers?
- 1500ms, 2000ms, or 3000ms for unusual responses?


Data
1
all
only correct
2
85%
90%
95%
3
200ms 1500ms
200ms 2000ms
200ms 2500ms
250ms 1500ms
250ms 2000ms
250ms 2500ms
300ms 1500ms
300ms 2000ms
300ms 2500ms
4
2SD
2.5SD
3SD
p = .10
p = .02
p = .08
2 x 3 x 9 x 3
=
162 datasets

What flexibility causes

- Flexibility in data processing -> Researcher degree of freedom
- When exploited, researcher *df* can lead to:
    - arbitrary variation in statistical results between studies
    - questionable research practices (such as *p*-hacking)
    - What if a researcher goes through many possibilities and only report the one he likes?

p-hacking

Simmons et al. (2011) demonstrated that:

1. even **one** binary decision inflates the rate of false-positive discoveries to the point that is unacceptable (i.e., larger than 0.05)
2. with four degrees of freedom, it can become as high as an astonishing rate of **61%**


Somewhere in the
Multiverse....

Multiverse analysis

Steegen, Tuerlinckx, Gelman and Vanpaemel (2016)

- Data used in a statistical analysis are usually not person-free. They are "to a certain extent **actively constructed**" (p. 702)
- An analyst constructs a given dataset from raw data, but a single analysis on one dataset is not sufficient because it is the only one of many equally reasonable possibilities, that is, a many worlds or **multiverse** of datasets.
- The concept of a multiverse of datasets is critical because the arbitrariness in reaching a particular dataset is inevitably inherited by the statistical result, and thus:
  - "the data multiverse directly implies a multiverse of statistical results"

Multiverse analysis

In multiverse analysis, a researcher:

1. identifies every possible method of data processing (e.g., coding, cleaning, and transforming data) that are equally justifiable
2. performs the same set of analysis across the whole data multiverse

- There is also, however, **a multiverse of statistical models** that can interact with and multiply the multiverse of datasets (we do not consider it here)

A multiverse analysis of L2 RT data: Study

Maie and DeKeyser (2020)

- Adult L1 speakers of English learned a semiartificial language called, *Japlish* (i.e., English lexicon and Japanese syntax and case-marking system), from incidental exposure
- The researchers investigated whether the exposure results in explicit and/or implicit knowledge
- Three structures:
  - Simple: O-S-V and O-S-I-V
  - Complex: O-S-[S-V]-V and O-S-[S-I-V]-V
  - Case: *-ga*, *-o*, and *-ni*

A multiverse analysis of L2 RT data: Analysis

- Variables to consider
  1. **Complexity**: Simple (-1), Complex (0 or 1), and Case (0 or 1)
     - Effect coded
  2. **Grammaticality**: Grammatical (0) and Ungrammatical (1)
     - Dummy coded
- The original study used a combination of (M)ANOVAs
- For the current demo, a mixed-effects model:
  - `RT ~ Complex*Grammaticality + Case*Grammaticality + (1|Subject) + (1|Item)`
- Here, we only consider the regression coefficient, `Case:Gramamticality`, which estimates a difference between the grammatical and ungrammatical sentences on Case
  - The original study found an effect here!

Creating data multiverse

Literature Review on:

1. Previous studies that used a word-monitoring task as a measure of implicit knowledge (e.g., Godfroid 2016; Granena, 2012; Suzuki, 2015)
2. An often-cited book that discussed technical aspects of conducting L2 RT research (Jiang, 2012)

- I identified **five steps** in processing raw RT data (from a WMT) to construct datasets

Creating data multiverse

1. Excluding participants who do not reach a certain criterion level of accuracy in comprehension questions
    - 63% (Jiang, 2004)
    - 70% (Jiang, 2012)
    - 75% (Granena, 2012; Suzuki, 2015)

Creating data multiverse

2. Excluding outlier data points that are larger or smaller than some absolute cutoff points

- Lower bound
  - 100ms (Suzuki, 2015)
  - 120ms (Jiang, Novokshanova, Masuda, & Wang, 2011)
- Upper bound
  - 1500ms (Jiang, 2012)
  - 2000ms (Jiang et al, 2011)
  - 2500ms (Suzuki, 2015)

Creating data multiverse

3. Excluding data points that are outside of the normal range of a given participant
   - 2SD (Jiang, 2012)
   - 2.5SD (Godfroid, 2016)
   - 3SD (Granena, 2012; Suzuki, 2015)

Creating data multiverse

4. Excluding participants whose mean word monitoring latency is larger or smaller than the range of the whole group
   - 2SD
   - 2.5SD
   - 3SD

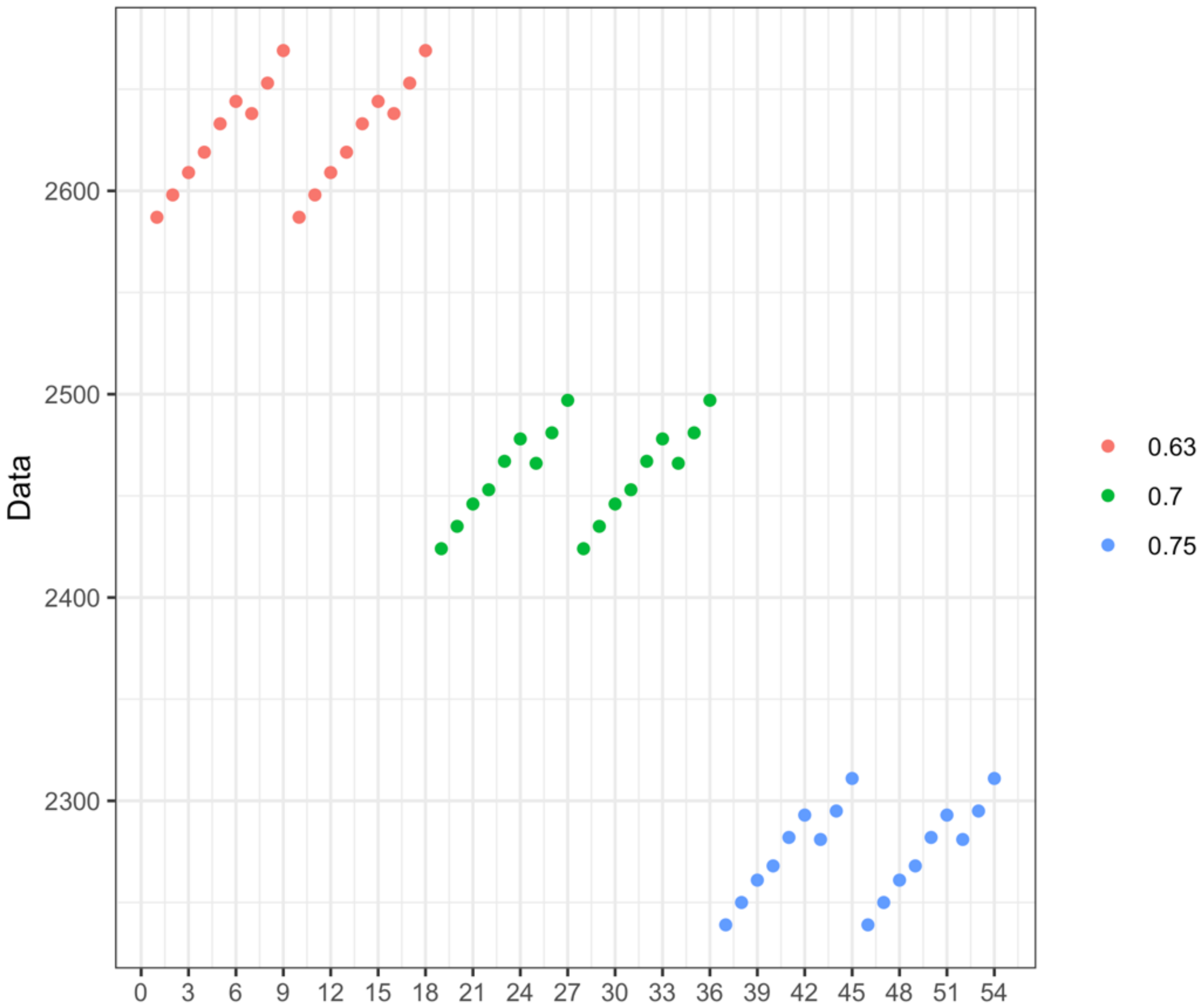
Creating data multiverse

5. Transforming data for statistical analysis
   - Logarithm (with base $= 10$)
   - Reciprocal

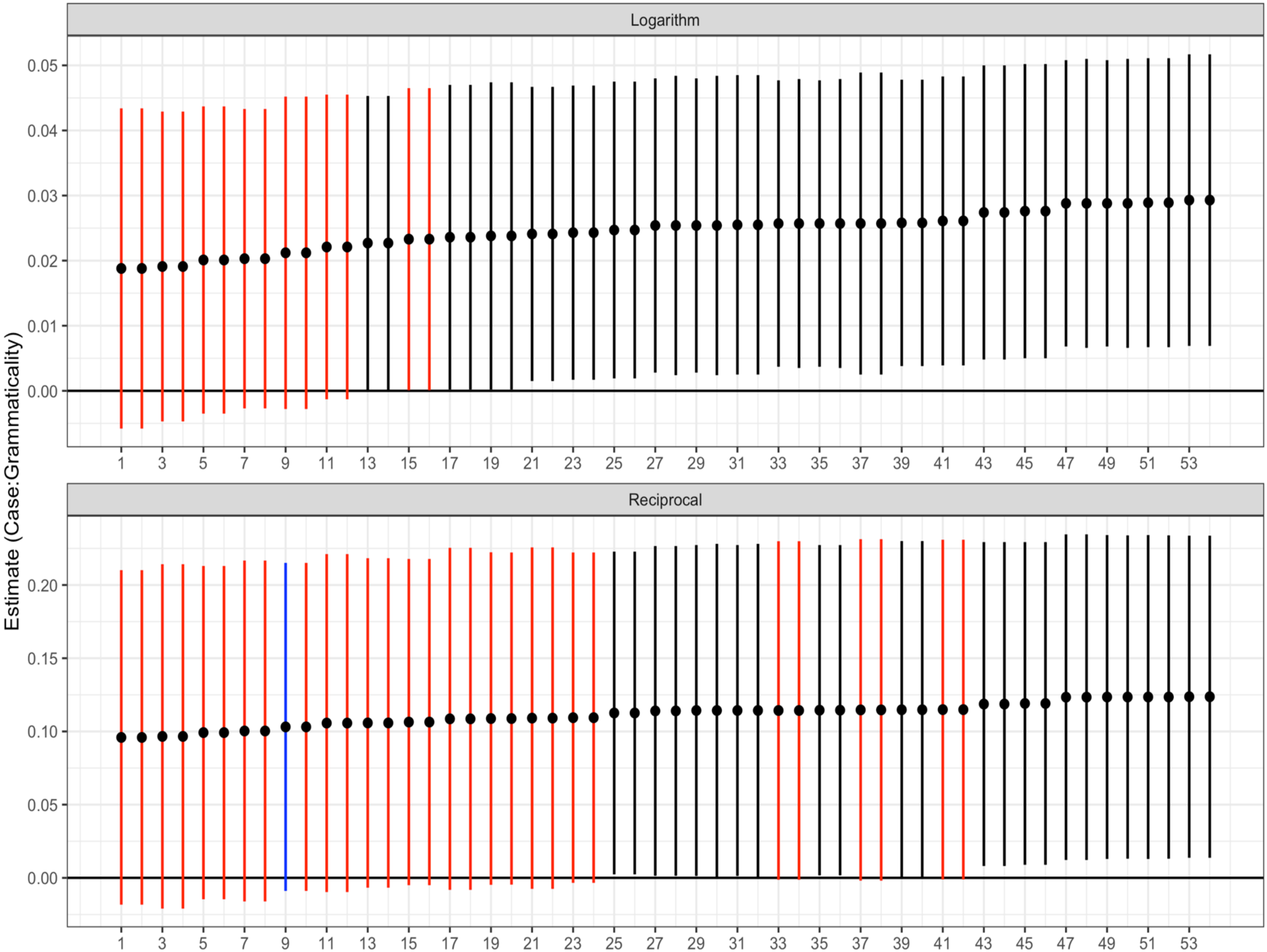
Creating data multiverse

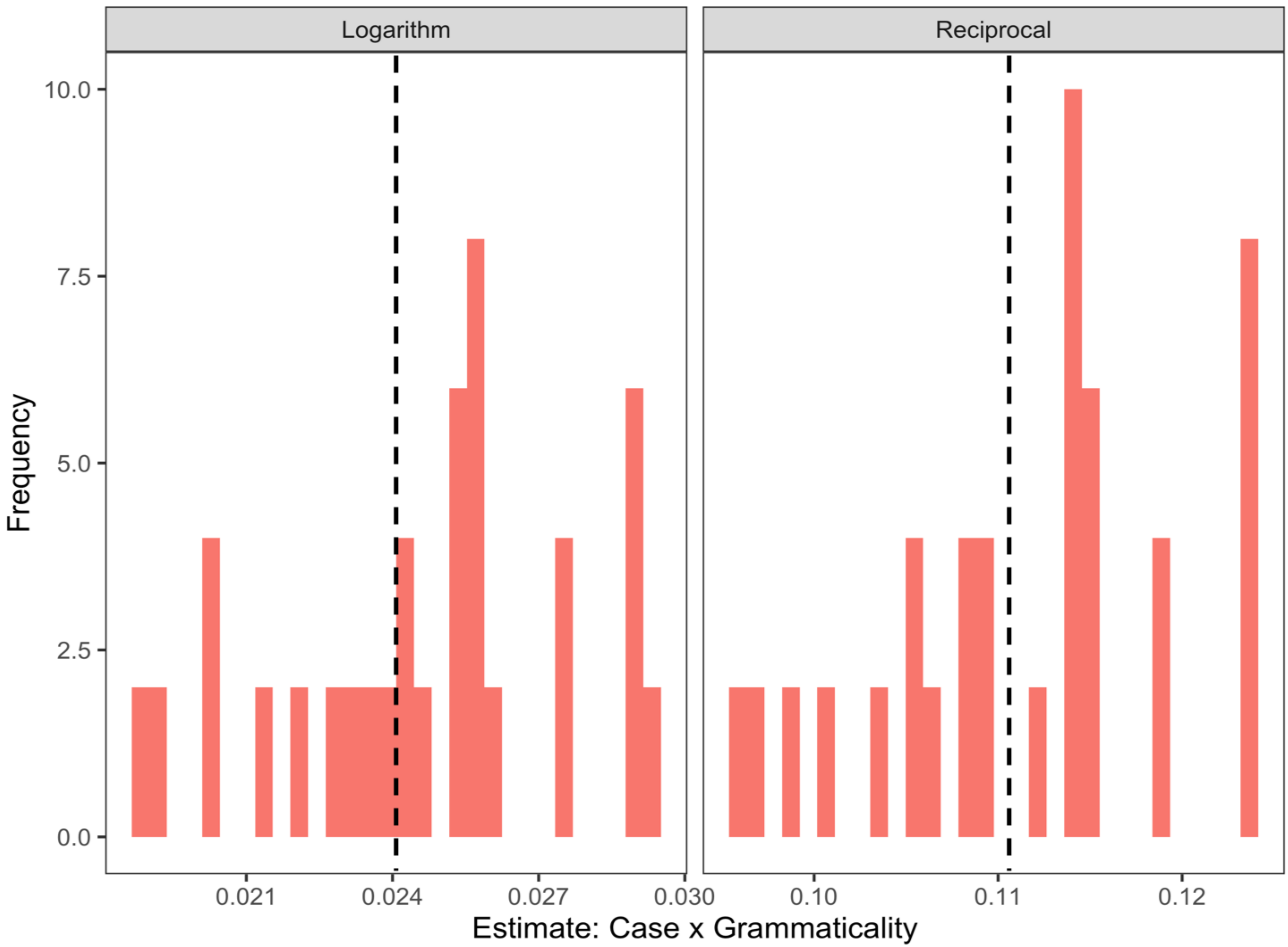
1. Excluding participants not reaching a critical level of accuracy in comprehension
2. Excluding outlier data points larger or smaller than absolute cutoff points
3. Excluding data points outside of the normal range of a participant
4. Excluding participants with mean word monitoring RT larger or smaller than the group
5. Transforming data for statistical analysis

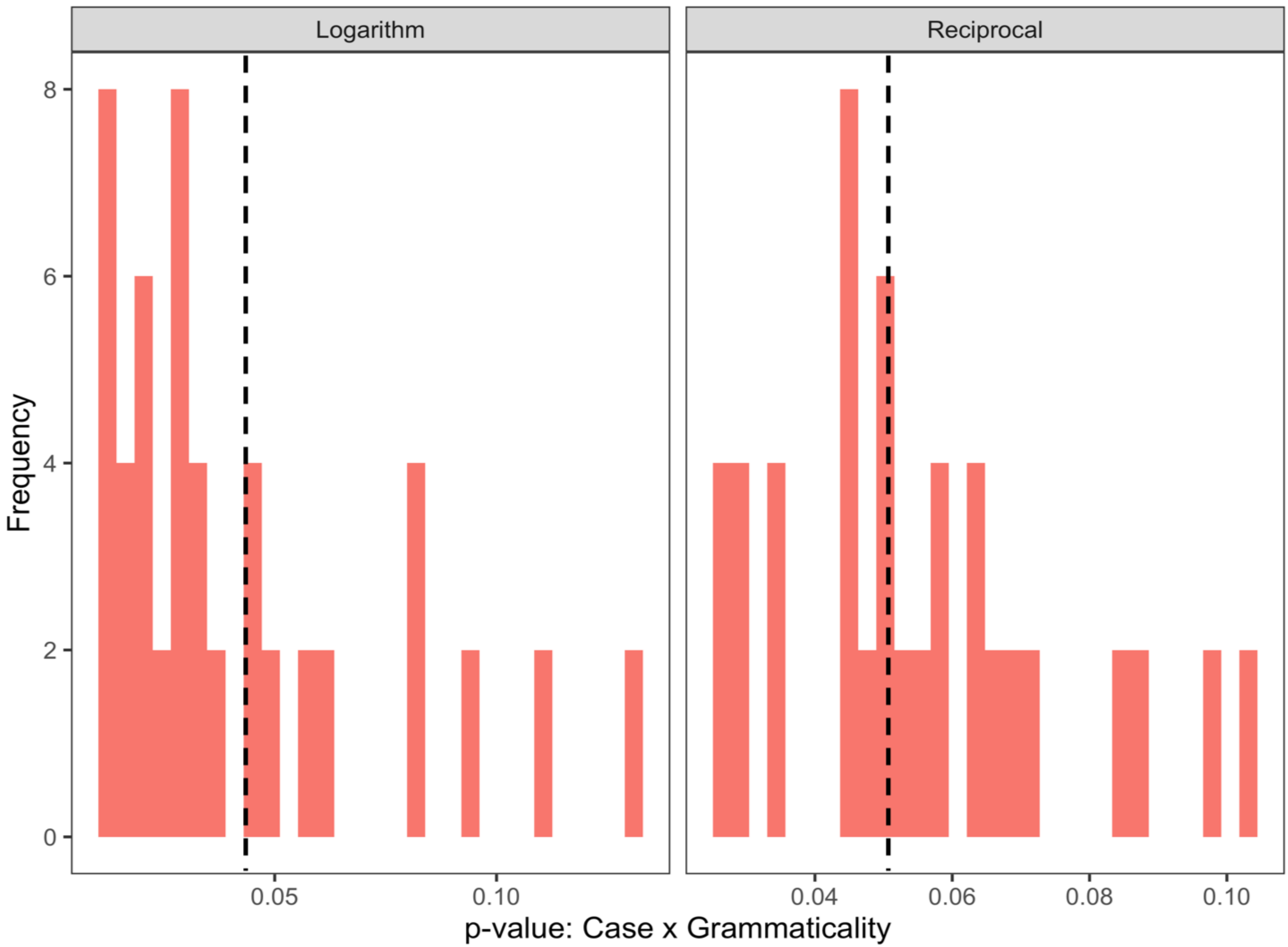
- **3 x 2 x 3 x 3 x 2 = 108 datasets**



Data
2600
2500
2400
2300
0
3
6
9
12
15
18
21
24
27
30
33
36
39
42
45
48
51
54
0.63
0.7
0.75


Logarithm
Reciprocal
Estimate (Case:Grammaticality)
0.05
0.04
0.03
0.02
0.01
0.00
0.20
0.15
0.10
0.05
0.00
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53


Logarithm
Reciprocal
10.0
7.5
5.0
2.5
0.0
Frequency
0.021
0.024
0.027
0.030
0.10
0.11
0.12
Estimate: Case x Grammaticality


Logarithm
Reciprocal
Frequency
8
6
4
2
0
0.05
0.10
0.04
0.06
0.08
0.10
p-value: Case x Grammaticality

AA [100, 1500] BA [120, 1500]
AB [100, 2000] BB [120, 2000]
AC [100, 2500] BC [120, 2500]

Step 4

Step 2

Step 3

Step 1

Step 1	Step 3	Log AA	Log AB	Log AC	Log BA	Log BB	Log BC	Rec AA	Rec AB	Rec AC	Rec BA	Rec BB	Rec BC
63%	2SD	0.010	0.010	0.034	0.010	0.010	0.034	0.028	0.027	0.044	0.028	0.027	0.044
63%	2.5SD	0.011	0.017	0.057	0.011	0.017	0.057	0.029	0.034	0.059	0.029	0.034	0.059
63%	3SD	0.011	0.017	0.080	0.011	0.017	0.080	0.028	0.033	<u>0.069</u>	0.028	0.033	0.069
70%	2SD	0.021	0.022	0.047	0.021	0.022	0.047	0.044	0.046	0.046	0.046	0.046	0.063
70%	2.5SD	0.023	0.036	0.081	0.023	0.036	0.081	0.059	0.047	0.056	0.047	0.058	0.084
70%	3SD	0.022	0.035	0.112	0.022	0.035	0.112	0.069	0.045	0.063	0.045	0.056	0.097
75%	2SD	0.028	0.029	0.062	0.028	0.029	0.062	0.049	0.051	0.070	0.049	0.051	0.070
75%	2.5SD	0.030	0.047	0.093	0.030	0.047	0.093	0.052	0.066	0.089	0.052	0.066	0.089
75%	3SD	0.029	0.046	0.129	0.029	0.046	0.129	0.051	0.064	0.103	0.051	0.064	0.103

Discussion: What to do?

- Variation in statistical results due to arbitrary choices during data processing and **you can reach different conclusions depending on which datasets you analyzed**!
- This cannnot be solved by pre-registering a study!
- Multiverse analysis is similar in concepts to sensitivity and outlier analysis, but it differs by **explicitly recognizing and incorporating uncertainty in statistical results induced by arbitrary choices**.

Discussion: What to do?

- One way to go around this problem is to take the average over the multiverse of statistical results
  - Estimate:
    - Mean = .024 (.018-.029) for Log and .114 (.095-.123) for Reciprocal
  - *p*-value:
    - Mean = .042 (.010-.129) for Log and .055 (.026-.106) for Reciprocal

Discussion: What to do?

And then make a statistical inference

- with an effect-based term:
  1. average over all the (multiverse of) regression models
     - conduct the same analysis for all other regression coefficients
  2. enter the value of indepdent variables (e.g., Gram vs Ungram) to linear predictors
  3. then, interpret the predicted value from the averaged model
     - What if the predicted difference between Gram and Ungram is 50ms?
- with uncertainty embraced and admitted
  - "**Statistics is the science of uncertainty and variation**" (Gelman, 2018, p. 41), but often, it is treated as a form of alchemy that coverts randomness into certainty - **uncertainty laundering**

Discussion: Moving forward

- Multiverse analysis is certainly recommended!
  - It does not have to be on the main document
- In Bayesian analysis, you can not only incorporate uncertainty of parameter estimation but also uncertainty that is caused by the multiverse of statistical results

Thank you for listening!

Acknowledgement

- Robert DeKeyser (University of Maryland)
- Michael Long (University of Maryland)
- Steve Ross (University of Maryland)
- Bronson Hui (Michigan State University)
- Masaki Eguchi (University of Oregon)

References - 1

Gelman, A. (2018). Ethics in statistical practice and communication: Five recommendations. *Royal Statistical Society*, *15*, 40-43.

Godfroid, A. (2016). The effects of implicit instruction on implicit and explicit knowledge development. *Studies in Second Language Acquisition*, *38*, 177-215.

Granena, G. (2012). *Age differences and cognitive aptitudes for implicit and explicit learning in ultimate second language attainment* (Unpublished doctoral dissertation). University of Maryland, College Park, College Park, MD.

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, *25*, 603-634.

References - 2

Jiang, N. (2012). *Conducting reaction time research in second language studies*. London, UK: Routledge.

Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). Morphological congruency and the acquisition of L2 morphemes. *Language Learning*, *61*, 940-967.

Maie, R., & DeKeyser, R. M. (2020). Conflicting evidence of explicit and implicit knowledge from objective and subjective measures. *Studies in Second Language Acquisition*, *42*, 359-382.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366

References - 3

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702-712.

Suzuki, Y. (2015). *Using new measures of implicit L2 knowledge to study the interface of explicit and implicit knowledge* (Unpublished doctoral dissertation). University of Maryland, College Park, College Park, MD.