

---

# Conceptual aspects in experimental and corpus data visualization

Guilherme D. Garcia

Université Laval | [gdgarcia.ca](https://gdgarcia.ca)

UNIVERSIDADE DE LISBOA

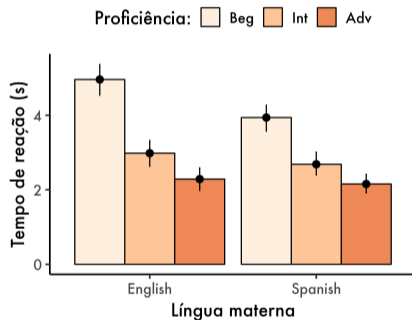
September, 2024



# Why visualize data?

Maximize our understanding of empirical patterns

	L1	Proficiency	$\bar{x}$	95% CI
1	English	Beg	4.96	[5.38, 4.53]
2	English	Int	2.98	[3.36, 2.60]
3	English	Adv	2.29	[2.60, 1.98]
4	Spanish	Beg	3.94	[4.30, 3.57]
5	Spanish	Int	2.68	[3.02, 2.35]
6	Spanish	Adv	2.16	[2.42, 1.89]

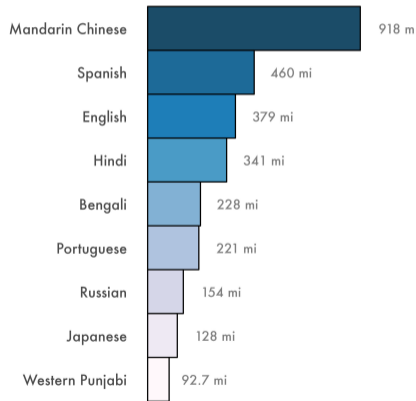
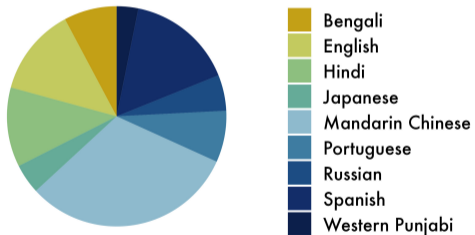


- Data from Garcia (in press): [doi.org/10.31219/osf.io/8r4ec](https://doi.org/10.31219/osf.io/8r4ec)



# Why visualize data?

Maximize our understanding of empirical patterns



# Why visualize data?

## Pre-submission

- Explore patterns
- Adjust methods

## Publication

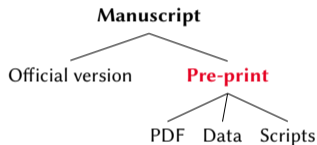
- Methodological clarity
- Efficiency in communication


☞ Acquisition studies: **a lot of data + variation** → another important reason

# Before we start

## A recommended workflow

- The notion of [Open science](#): open access to studies, data and publications



- Repositories such as [OSF](#) are highly recommended 
- Open access, SEO, control over formatting and updates, etc.
- ☞ Some journals tend to **limit** visualization options (e.g., colours)

# The issue

- Studies in language acquisition underuse quantitative methods
- The same applies to **visual exploration**

(Plonsky 2011)

☞ This can impact readers' understanding of key points  
*Inappropriate analyses also affect the reliability of a study*

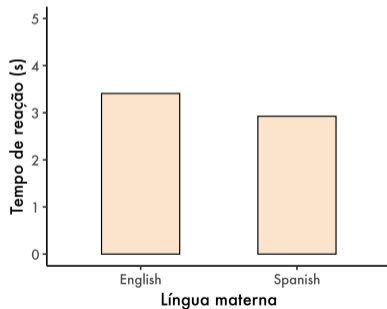
**Good and bad:** complex models are easily executed these days

## CONTINUOUS DATA

# Example

## General patterns

- Two groups of French learners: L1 English and Spanish
- Data: force-choice task → accuracy (0/1), reaction time (s) and certainty level (1–4)



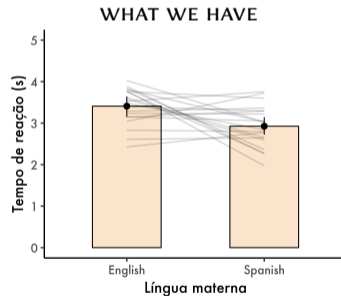
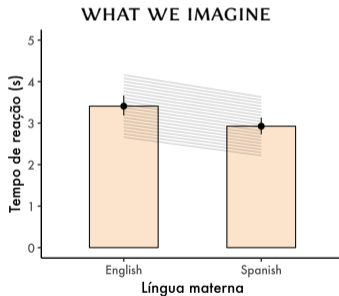
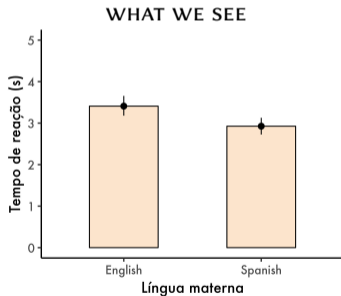
- Little information ( $\bar{x}$ ): only two variables
- Absence of error bars
- ☞ Variation is not shown



# Visual exploration

## Visualizing variation: items

- Learners' behaviour: **rarely** constant across **items** in a study

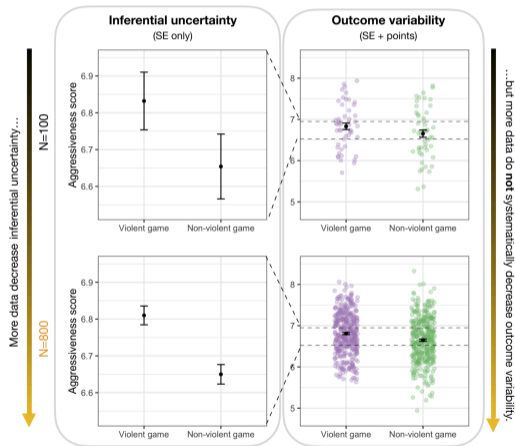


# Visual exploration

## The illusion of great magnitude

- Inferential uncertainty vs variability
- **Very** distinct perceptions

👉  $n$  may not fix the issue

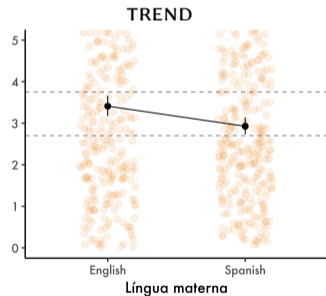
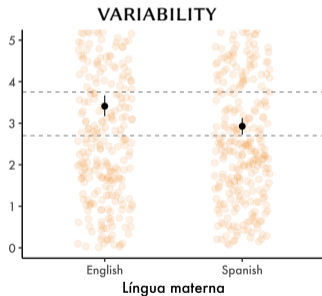
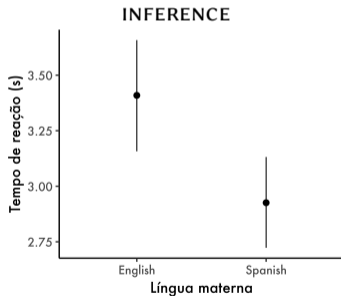


(Zhang et al. 2023, p. 2)

# From general to specific

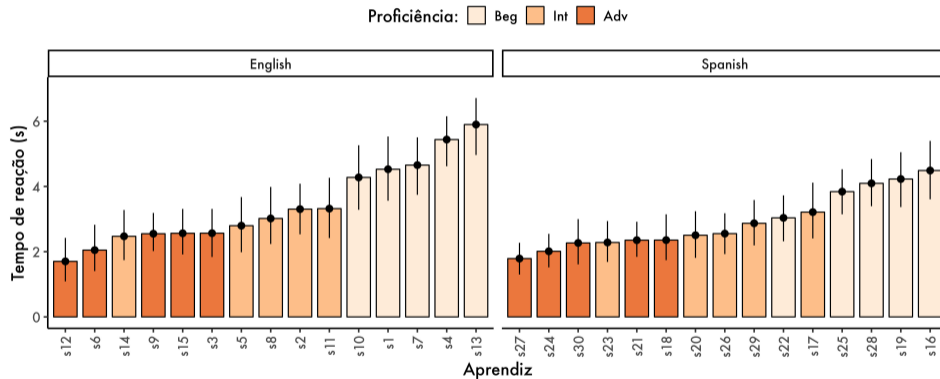
## The illusion of great magnitude

- Figures affect our conclusions on effect sizes for L1



# Visual exploration

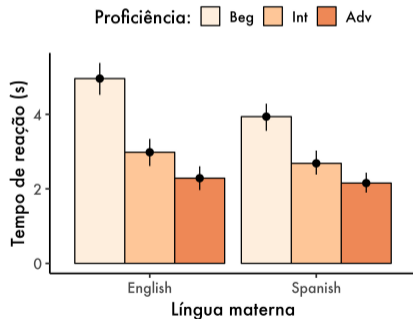
## Visualizing variation: learners



- Variation in reaction times: learners, L1, and proficiency levels

# Visual exploration

- L1 and proficiency are relevant here
- ☞ Main variable: **proficiency** (L1 *probably* doesn't matter)

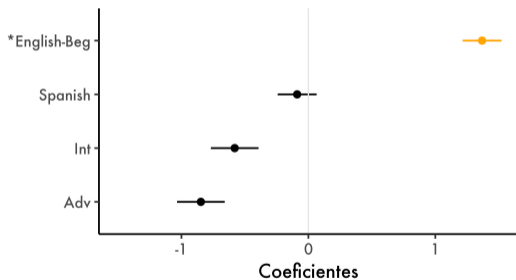


# Analysis

## Visualization of statistical models

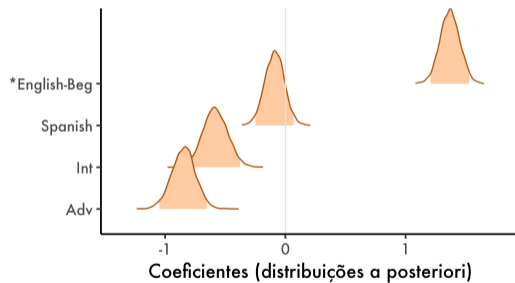
### Frequentista

Intervalos de confiança 95%



### Bayes

Intervalos de credibilidade 95%

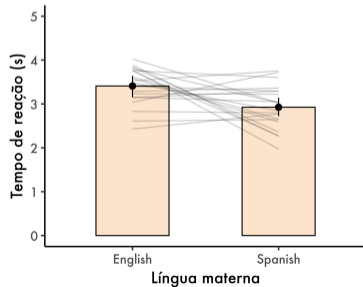
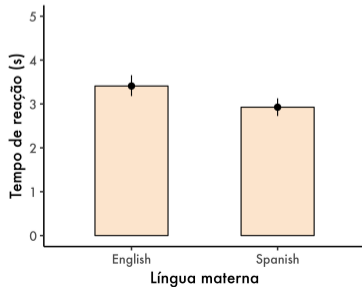


# Análise

## Visualization of statistical models

- Not all variation needs to be included in a model: here, **singular fit**
- Too much complexity when we consider main variables + random effects
- ☞ Here, **L1 + Proficiency**  $\succ$  variation across items and learners

**Question:** which figure is more appropriate?



## SCALAR DATA



# Visual exploration

## Scales

- Scales are often used in acquisition studies

*How certain are you about your response?*

1	2	3	4
---	---	---	---

- ☞ Like binary variables, ordinal variables often require **transformation**

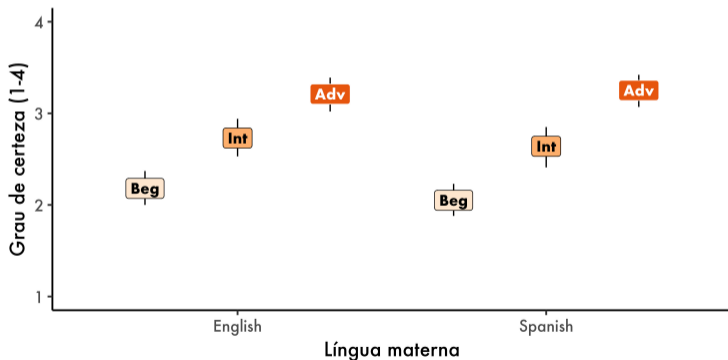
### **How to visualize scales...?**

- ☞ Discussion based on Garcia (2021, ch. 5 e 8) and Garcia (in press)

# Visual exploration

## Scales

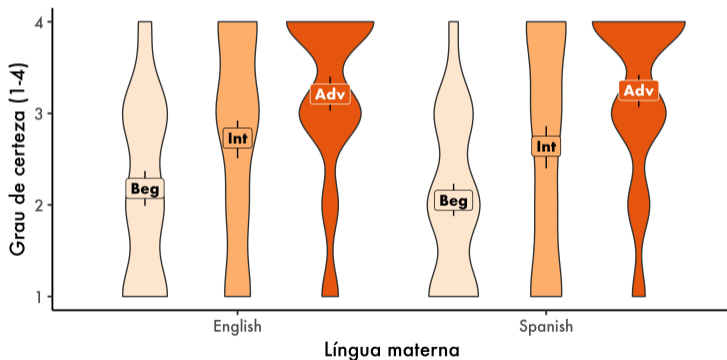
- Ordinal data are not continuous (**ordered factor**)
- Distribution is rarely normal → means are not very representative



# Visual exploration

## Scales

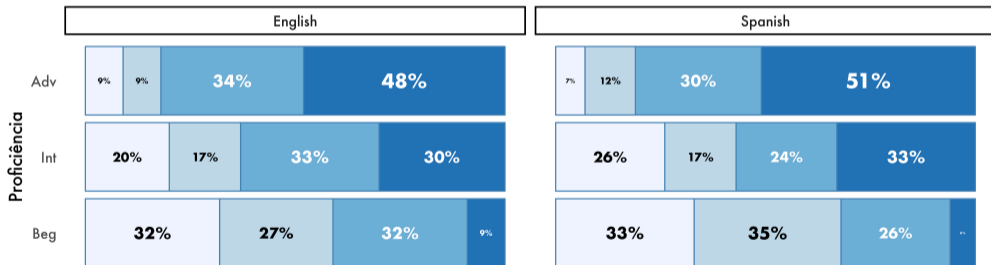
- Ordinal data are not continuous (**ordered factor**)
- Distribution is rarely normal → means are not very representative



# Visual exploration

## Scales

- Visualization **without** transformation → better aligned with ordinal models
- Bars and colours that mirror the original scale (adapted from Garcia 2021, p. 100)



Escala de certeza: 1-4

# Visual exploration

## Scales

- Grey scale for physical publications (here again Proficiency  $\succ$  L1)
- Easy adaptation with different palettes using `ggplot2`



Escala de certeza: 1-4

# Visual exploration

## Scales

### Data preparation

1. Group relevant variables
2. Count  $n$  for each point along scale
3. Calculate percentages

```
code
1 prop = viz ▷
2   summarize(n = n(),
3             .by = c(L1, Proficiency, Certainty)) ▷
4   mutate(Prop = n / sum(n),
5          .by = c(L1, Proficiency),
6          Dark = if_else(Certainty %in% c("3", "4"),
7                         "yes", "no"))
```

- ☞ Adding variable to help with customization in figure (lines 6–7)

# DATA PREDICTABILITY

# What's the goal of our analysis?

1. Examine the role of variables in a study with specific theoretical underpinnings
2. Generate the best possible model to predict new data (e.g., *machine learning*)
3. ...

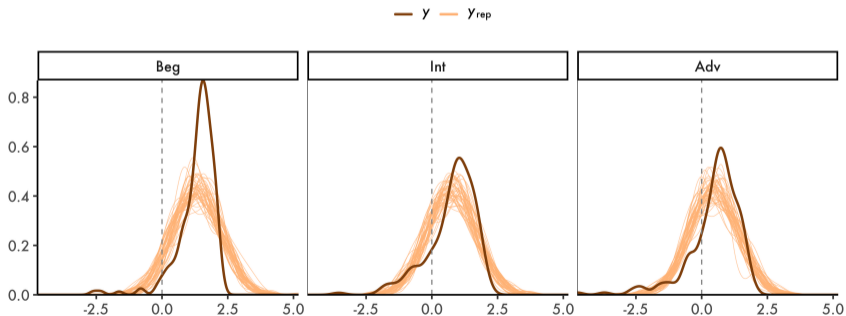
- ☞ Normally, predictability is related to goal (2)
  - But visualizing predictions can be quite informative



# What's the goal of our analysis?

## Posterior predictive check

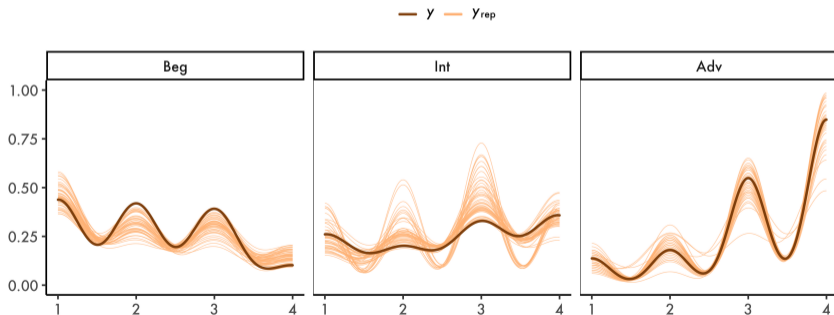
- Comparison between real data and data simulated from a model
- Model examining **reaction time** (log) as a function of L1 and Proficiency:



# What's the goal of our analysis?

## Posterior predictive check

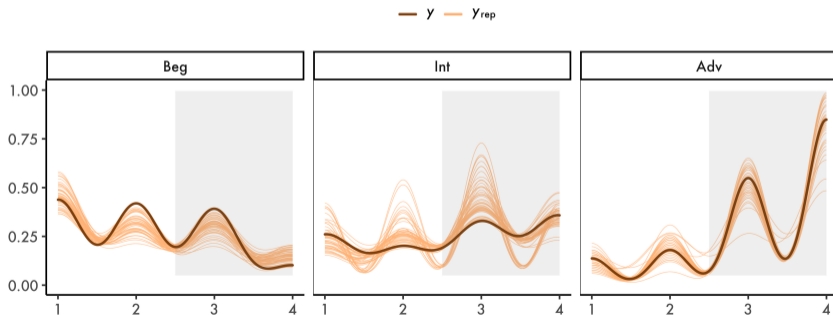
- Comparison between real data and data simulated from a model
- Model examining **certainty** as a function of L1 and Proficiency:



# What's the goal of our analysis?

## Posterior predictive check

- Comparison between real data and data simulated from a model
- Model examining **certainty** as a function of L1 and Proficiency:



# Final thoughts

- Visualize patterns before, during and after analysis
- Carefully consider aesthetic aspects: colours, sizes, amount of info
- Align visualization with analysis and goals: maximize efficiency
- Host materials and analysis on line (pre-prints + extras)

# Visual resources

## R packages

- [tidyverse](#) ([ggplot2](#))
- [plotly](#) (gráficos interativos)
- [MoMAColors](#)
- [RColorBrewer](#)

## Links e livros

- [r4ds.hadley.nz](#)
- [gdgarcia.ca](#) + [blog](#)
- Garcia (2021, 2023)
- Winter (2019)

- These slides are already available at [gdgarcia.ca/downloads](#)

👉 Now, on to RStudio<sup>1</sup> to see [ggplot2](#) in action


---

<sup>1</sup>One alternative to downloading RStudio: use RStudio cloud: [posit.cloud](#)

Muito

BRIGADO!

Dúvidas?



# References I

- Garcia, G. D. (2021). *Data visualization and analysis in second language research*. New York NY: Routledge.
- Garcia, G. D. (2023). Quantitative data visualization. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd. To appear. Pre-print, data and code available at <https://doi.org/10.31219/osf.io/8r4ec>.
- Plonsky, L. (2011). *Study quality in SLA: A cumulative and developmental assessment of designs, analyses, reporting practices, and outcomes in quantitative L2 research*. Ph. D. thesis, Michigan State University.
- Winter, B. (2019). *Statistics for linguists: an introduction using R*. New York: Routledge.
- Zhang, S., P. R. Heck, M. N. Meyer, C. F. Chabris, D. G. Goldstein, and J. M. Hofman (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences* 120(33), e2302491120.