Università degli Studi di Padova

PSICOSTAT – April 4, 2025

Reanalyzing published science

The power and perils of commentary articles as an early career researcher

Simone Gastaldon

Dipartimento di Psicologia dello Sviluppo e della Socializzazione (DPSS), Università di Padova

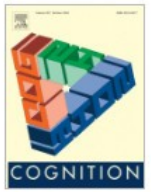
simone.gastaldon@unipd.it

Joint work with **Giulia Calignano** (DPSS, Università di Padova)

A tale of two papers

ELSEVIER

Cognition
Volume 227, October 2022, 105213



Conceptual alignment in a joint picture-naming task performed with a social robot

Giusy Cirillo [a b], Elin Runnqvist [a b], Kristof Strijkers [a b], Noël Nguyen [a b], Cristina Baus [c]

Received 5 May 2021, Revised 31 May 2022, Accepted 27 June 2022, Available online 5 July 2022, Version of Record 5 July 2022

ELSEVIER

Cognition
Volume 259, June 2025, 106099

Linguistic alignment with an artificial agent: A commentary and re-analysis

Simone Gastaldon [a b], Giulia Calignano [a]

Received 3 July 2024, Revised 24 January 2025, Accepted 25 February 2025, Available online 28 February 2025, Version of Record 28 February 2025

What I am NOT going to do in this talk

- Deep dive into all the specific methodological issues of the target paper and the details of our commentary (only main points)
- Focus on the details of the statistical analyses
- Lecture you on mistakes and how to revolutionize scientific publishing

What I AM going to do in this talk

- Focus on ***why*** we worked on a reanalysis & commentary paper
- Focus on ***how*** we proceeded and how we communicated with the authors
- Propose that ***Early Career Researchers*** should make commentaries and re-analyses
- Ask you about ***your opinion*** on and ***your experience*** with commentaries

A timeline

Jul 22
Target paper published (VOR)

Jul 24
Commentary submitted

Feb 25
Commentary published (VOR)

On & off work on commentary & reanalysis

3 rounds of review

Feb 23

- Target paper discussed at our weekly lab meeting as usual with other papers
- Beginning of dataset exploration

Jun 24

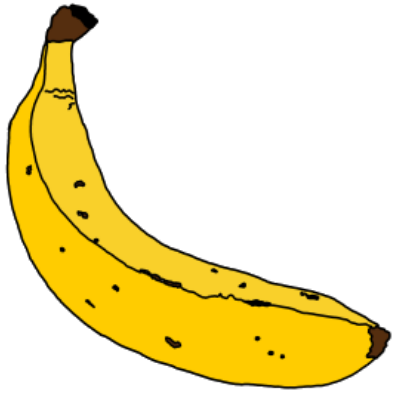
- First draft ready
- Draft sent to original authors
- Email exchange with authors

Ok, but can you tell us what the paper is about?

- Joint picture naming task (in French)
- Human – social robot partner (Furhat Robotics)
- Do participants (N = 24) align with the naming behavior of the robot?

BASIC CONDITION

For the category "fruit", the robot uses basic labels.
For 10/15 categories (mixed); 3600 trials in total.



Ok, but can you tell us what the paper is about?

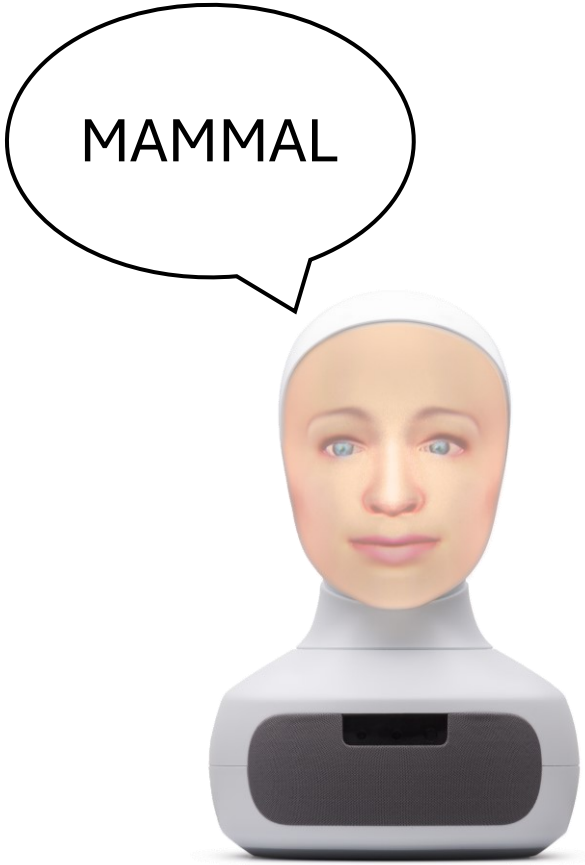
- Joint picture naming task (in French)
- Human – social robot partner (Furhat Robotics)
- Do participants (N = 24) align with the naming behavior of the robot?

Ok, but can you tell us what the paper is about?

- Joint picture naming task (in French)
- Human – social robot partner (Furhat Robotics)
- Do participants (N = 24) align with the naming behavior of the robot?

CATEGORY CONDITION

For the category "mammals", the robot uses category labels (superordinate). For 5/15 categories (mixed); 1800 trials in total.



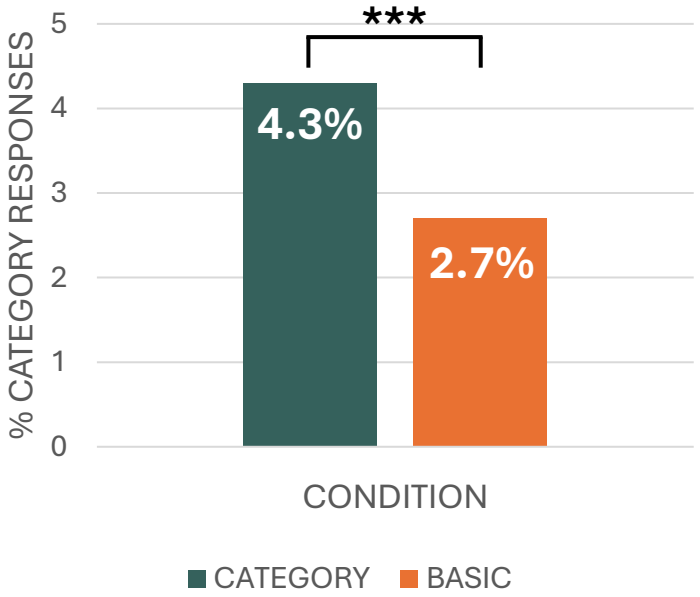
Ok, but can you tell us what the paper is about?

- Joint picture naming task (in French)
- Human – social robot partner (Furhat Robotics)
- Do participants (N = 24) align with the naming behavior of the robot?

Ok, but can you tell us what the paper is about?

% category responses in CATEGORY condition > BASIC condition

(GLM binomial family, random intercepts for participants and items)



ALIGNMENT WITH THE SOCIAL ROBOT

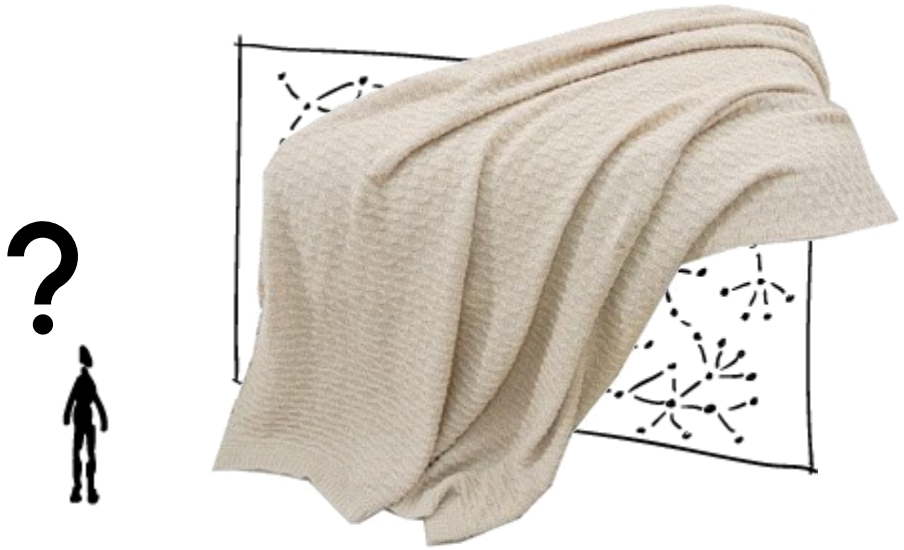
(DISCLAIMER: the authors did not use this kind of graph; only for illustration purposes)

Complexity remains hidden under a blanket

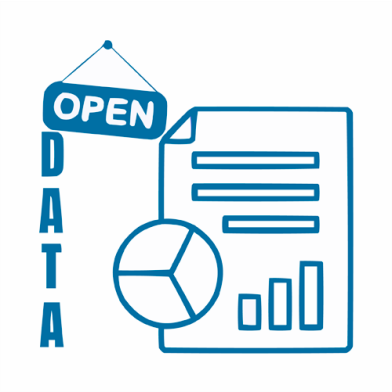
- 15 categories, 4 exp. blocks, lexical frequency, MultiPic (validated database)/new pictures: **statistics is blind to all this complexity** (if not/cannot be modeled)

- No visualization of **how responses are distributed**, despite category responses overall being very few to begin with (hence a rare naming behavior in the exp)

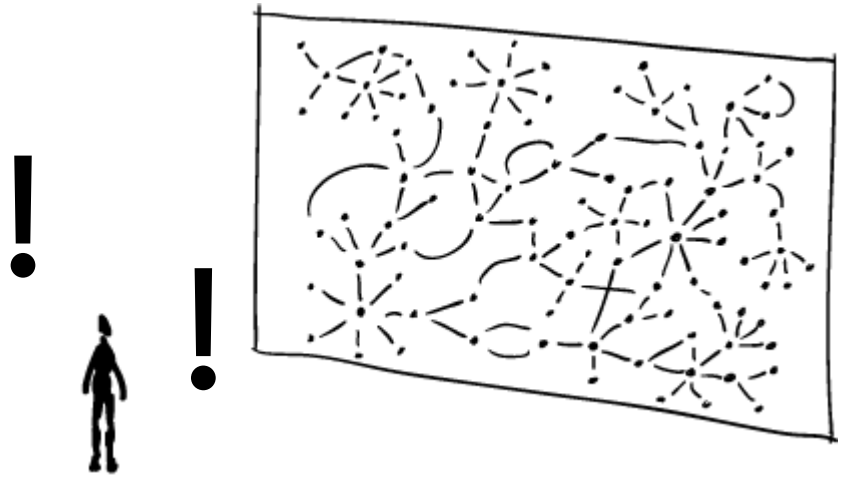
- To assess how **reliable and consistent** the effect is, data need to be thoroughly **explored**



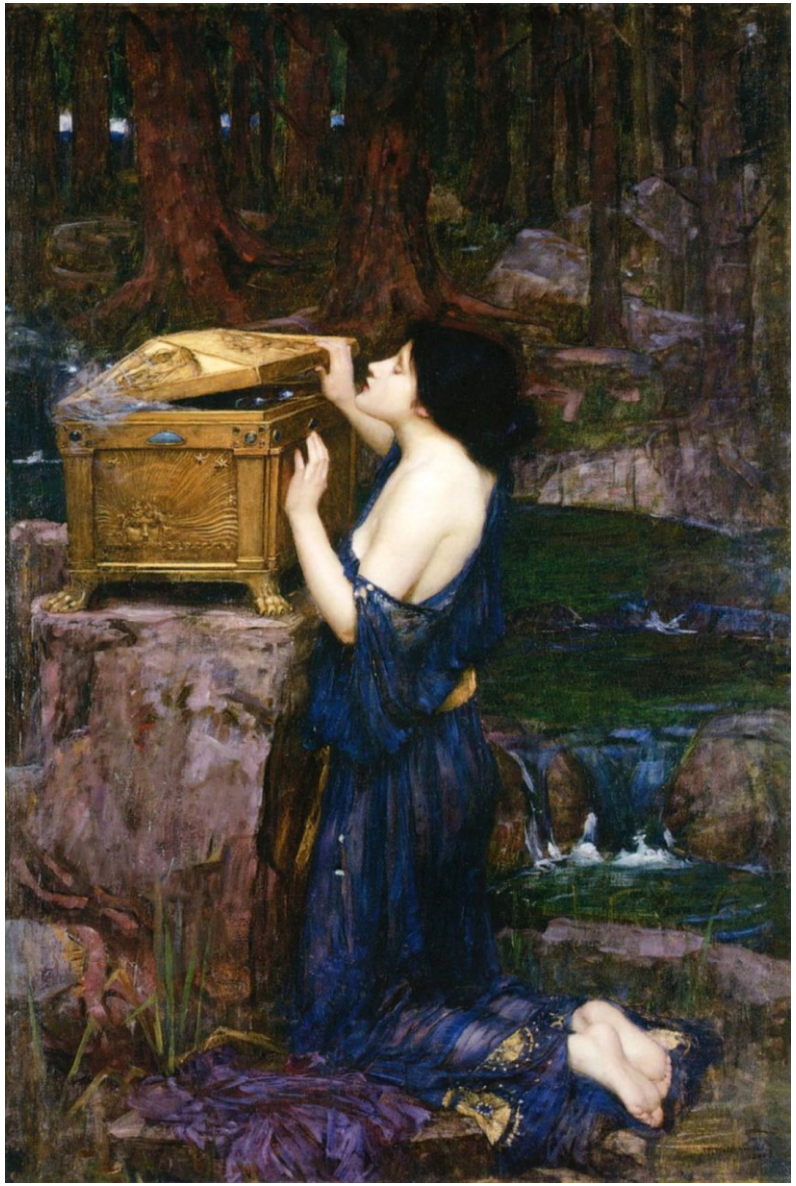
Praised be open data!



Possibility for the scientific community to explore them and **uncover (and appreciate) complexity**



Opening Pandora's box & uncovering complexity



- Analysis of response times to assess the claim of automaticity (not the focus here; see the papers)
- **Extensive data visualization** across many variables
- **Check for lexical frequency biases** in categories
- **Multiverse approach and robustness check**: how category-dependent is the effect?

Category responses cluster in some categories and new pictures

- 377 picture stimuli from the MultiPic database
- 73 picture stimuli were newly designed (novel pictures, not shared on OSF)
- 5400 trials in the exp.

 # RESPONSES

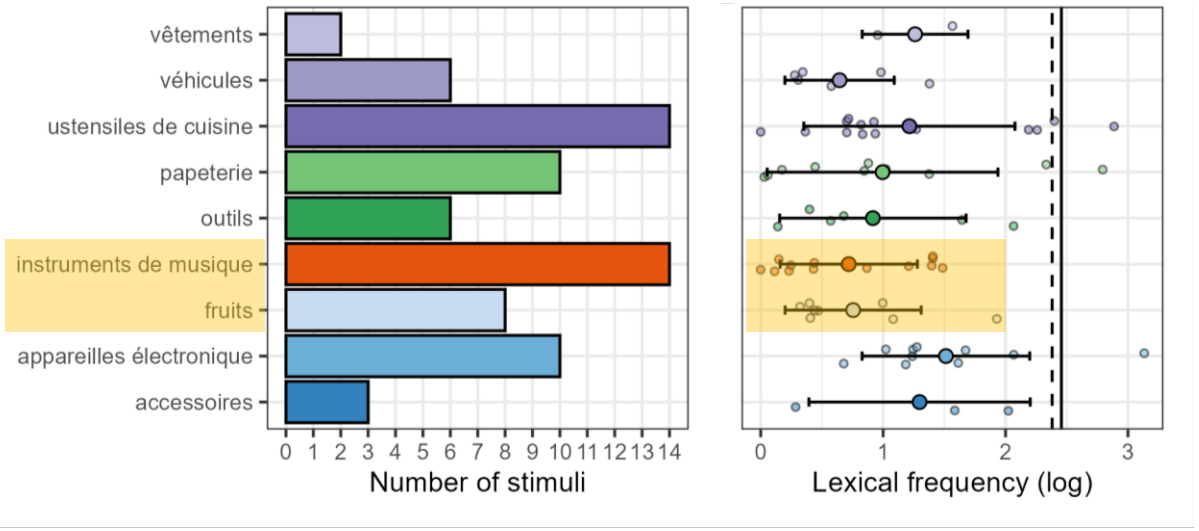
	BASIC	CATEGORY
ALIGNED	2862	**78**
NON-ALIGNED	**98**	1467

% RESPONSES

	BASIC	CATEGORY
ALIGNED	79.5%	**4.3%**
NON-ALIGNED	**2.7%**	81.5%

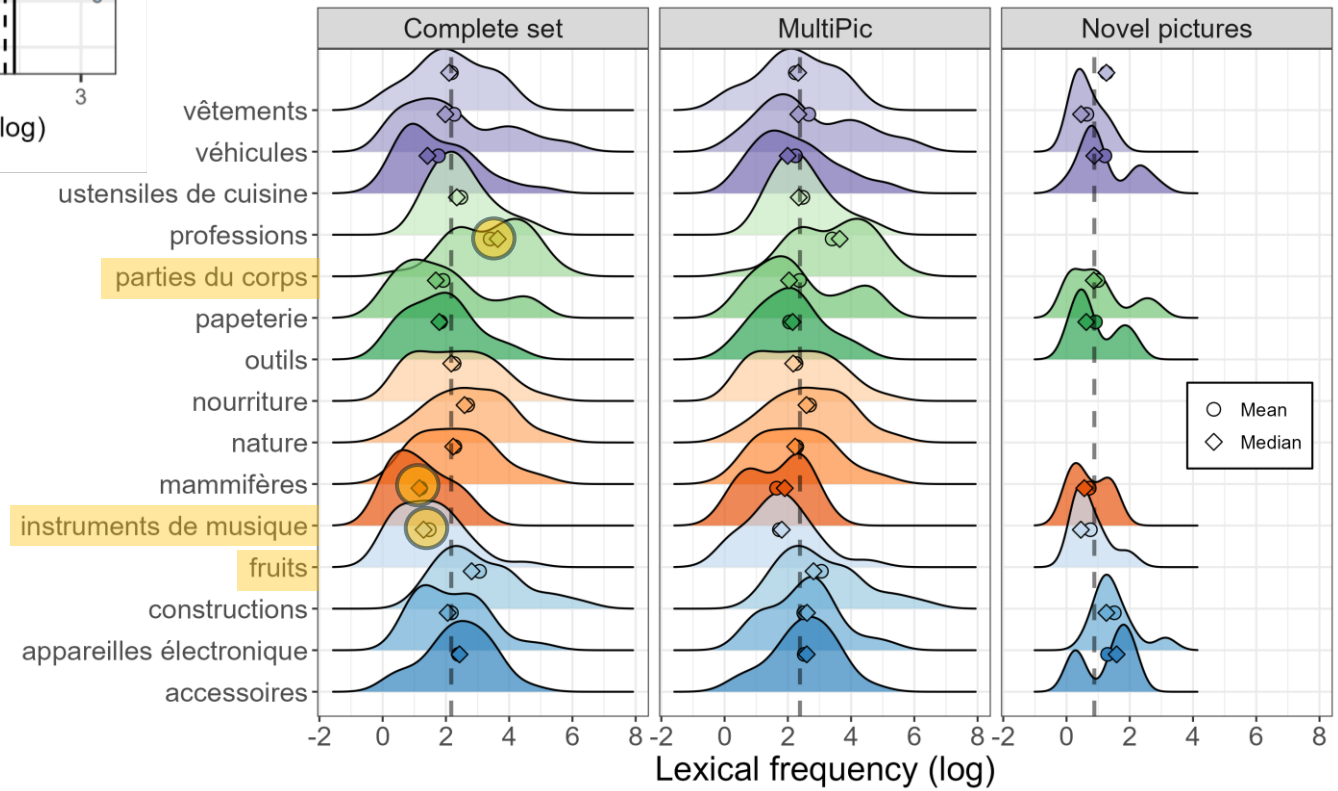
"Pure" alignment or non-alignment, excluding a variety of naming errors (N = 750)

Newly developed figures introduce biases in lexical frequency

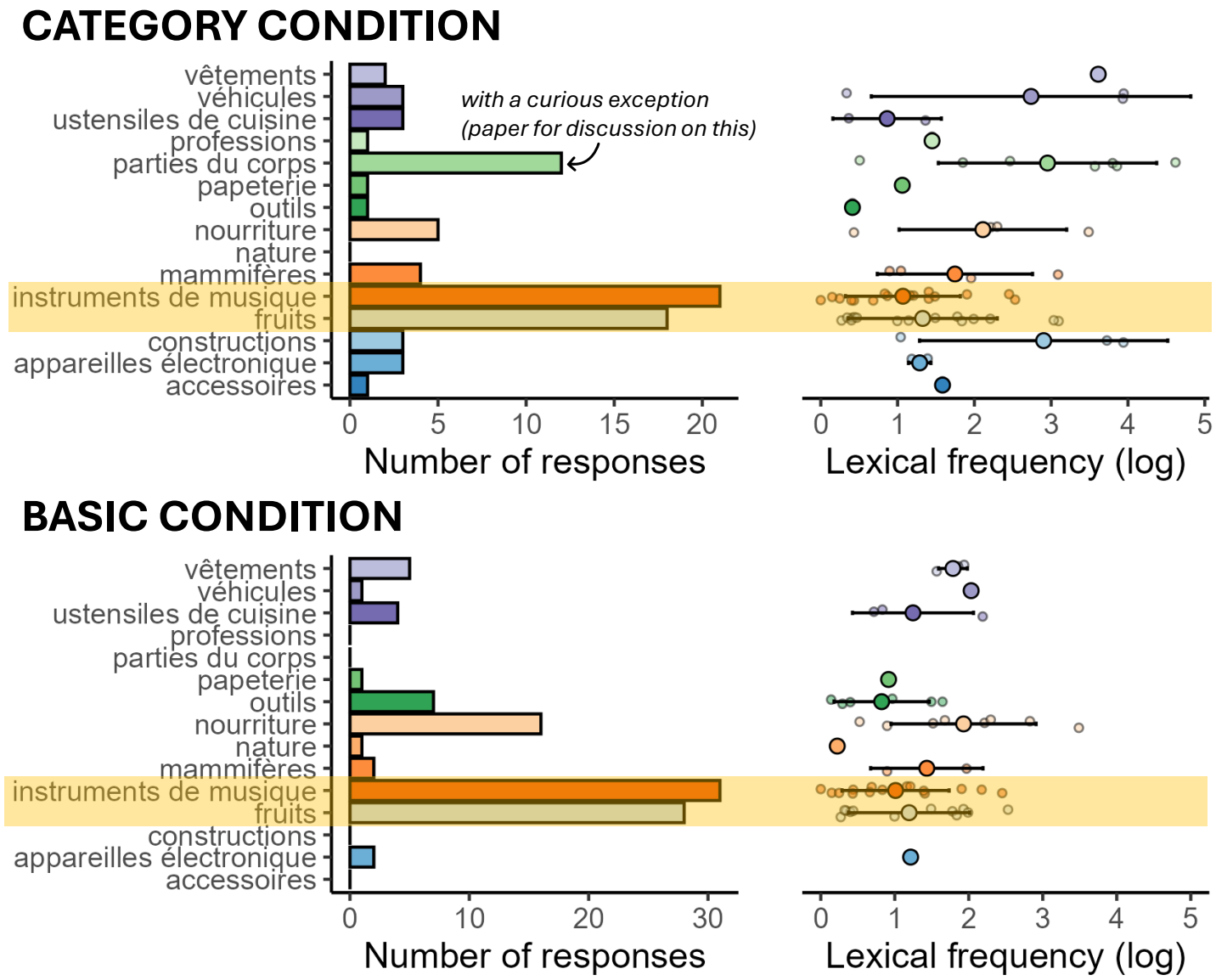


Bias in lexical frequency distributions of some categories that stand out relative to some of the others

New picture stimuli (N = 73) with words at low(er) lexical frequency (dashed and solid lines are mean and median for MultiPic *only* stimuli)

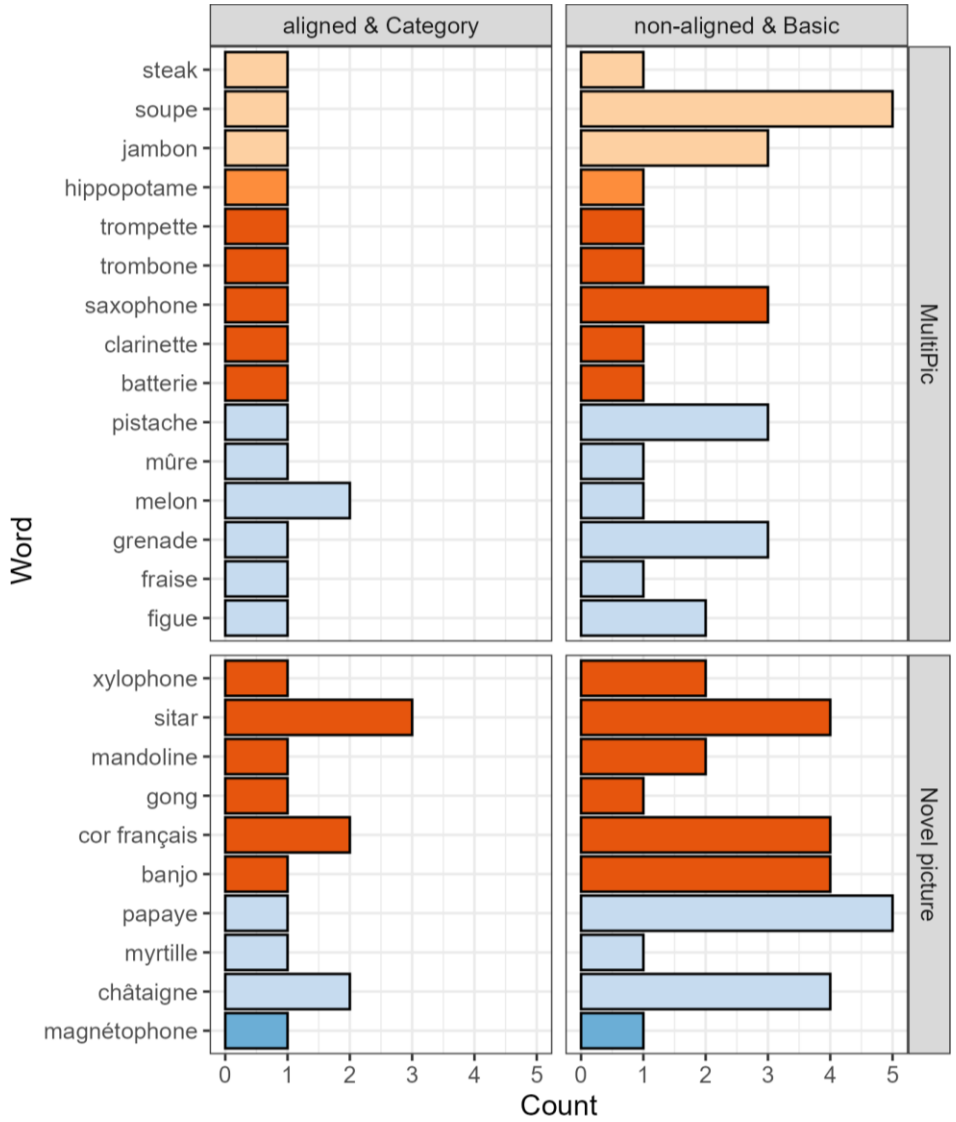


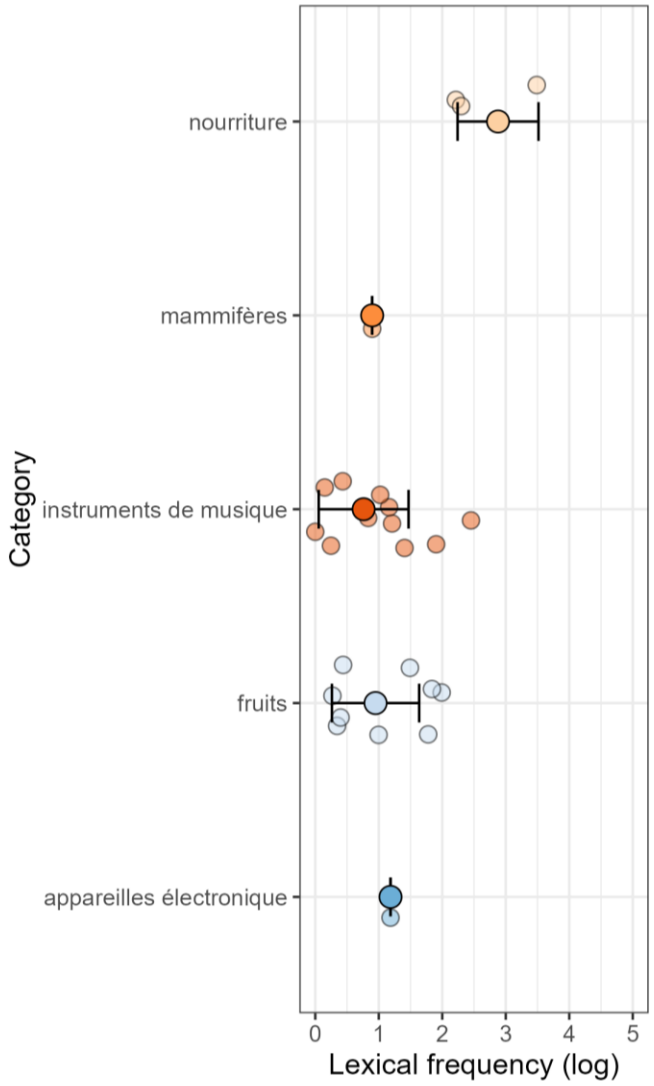
Category responses cluster primarily in low lexical frequency



In a naming task in which I learn that I am allowed to say the category, maybe I say the category more easily when I can't retrieve the basic label (and not because I align with the robot).

Category naming bias for a subset of items across conditions





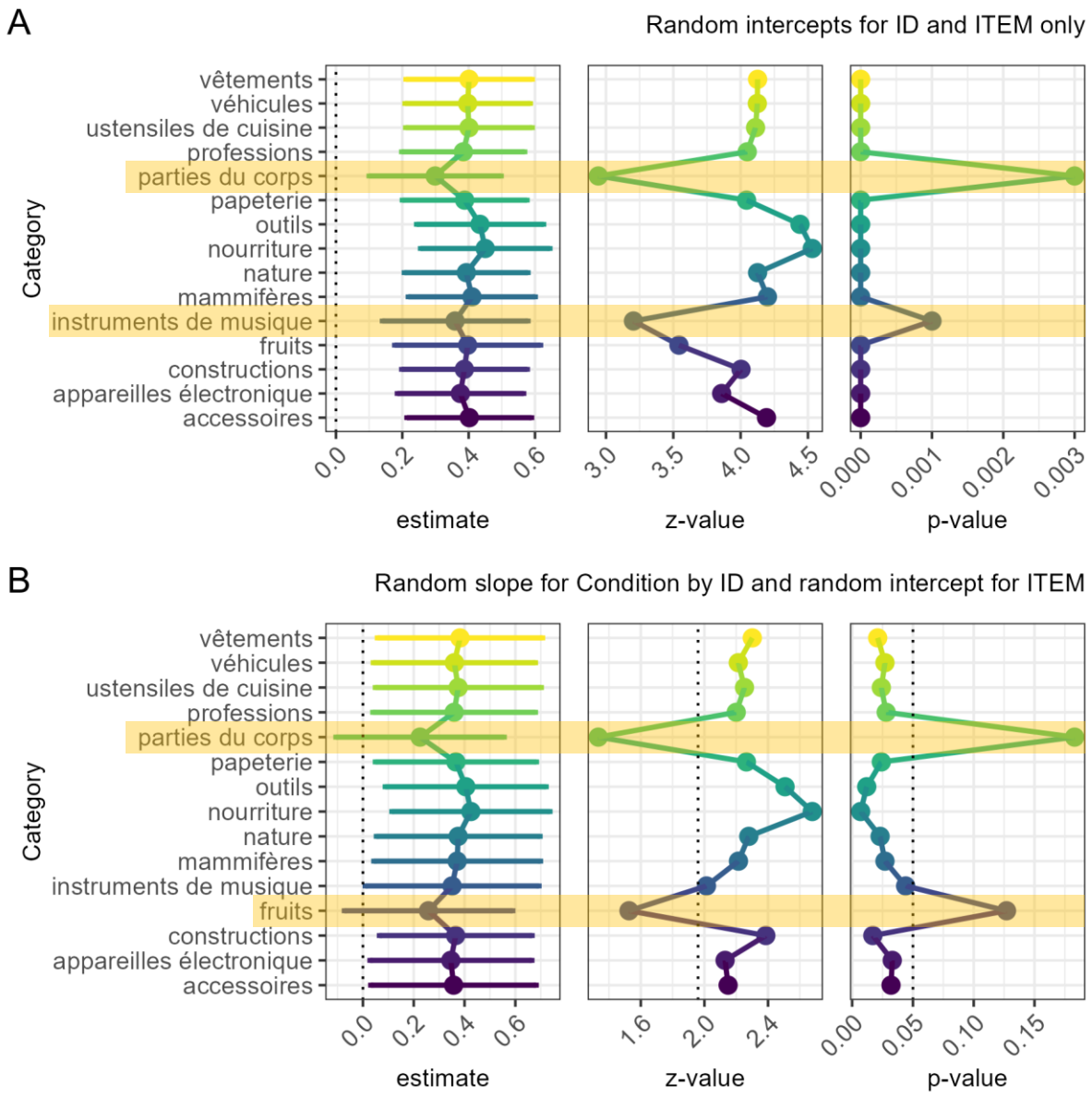
Category: appareilles électronique, fruits, instruments de musique, mammifères, nourriture

There is a bias for some items (N = 15) belonging to certain categories of being named with the category:

- low lexical frequency
- possibly unclear picture?
- low familiarity? (trompette, trombone, clarinette, sitar, cor français…) – no databases/ratings

In this context, a slight increased number of category responses in the category condition can greatly impact the statistics. But **the effect cannot be said to be robust, reliable, and generalizable!**

The effect is not robust and depends on category and participant



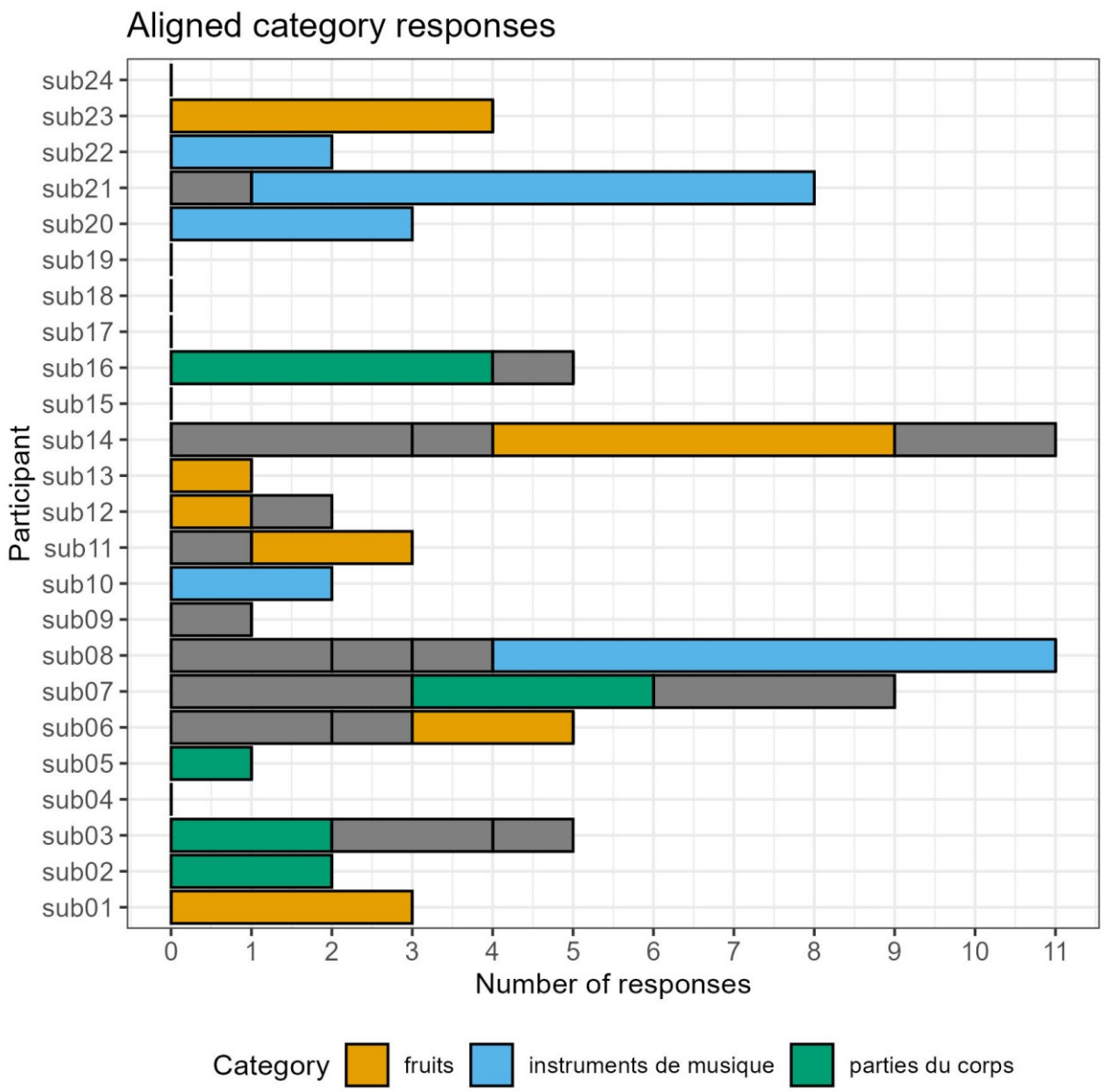
Leave-one-out approach for robustness check

The same influential categories:

- ❖ body parts
- ❖ fruit
- ❖ musical instruments

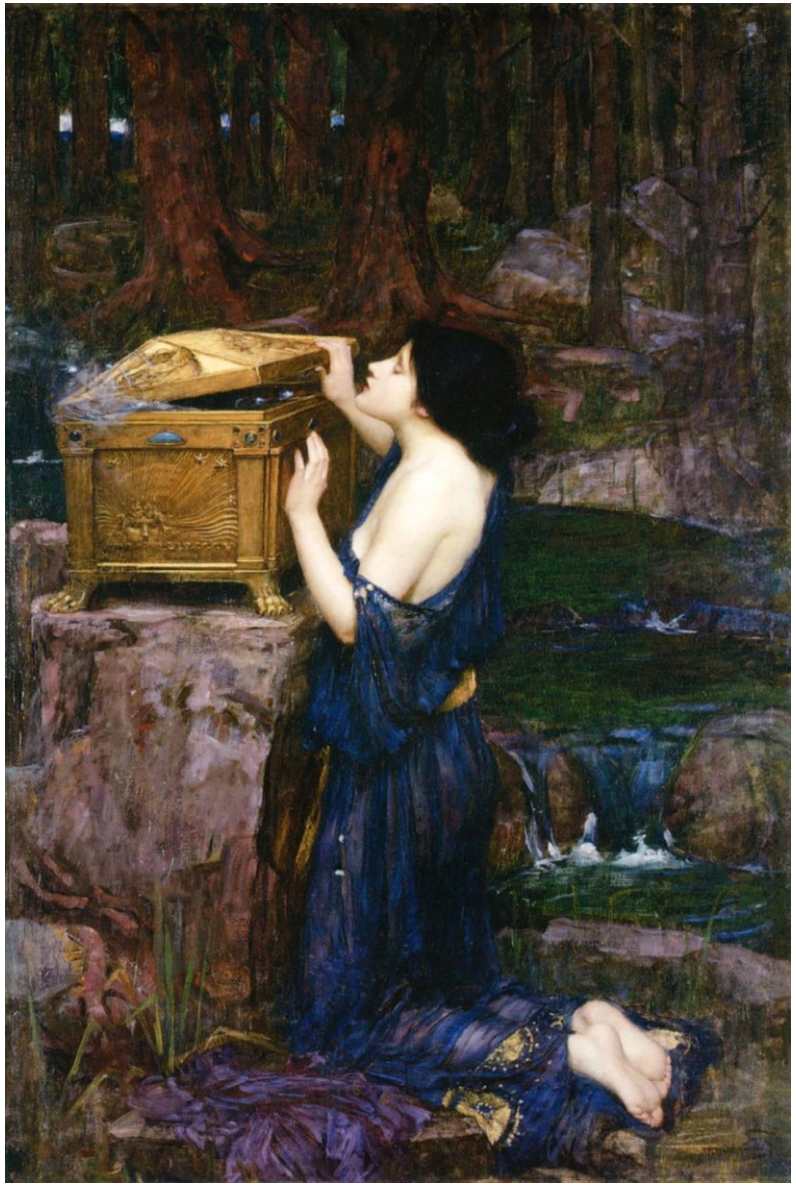
The effect also varies by participant (remember that participants had different categories in the category/basic condition!)

Few responses, for few categories, for few participants



- **50%** of aligned category responses **(N = 39)** are given by **4 participants**
- **65%** of aligned category responses **(N = 51)** are given for **3 categories** (body parts, fruit, musical instrument)

The curse of Pandora's box: limited generalizability



SPECIFIC INDIVIDUAL DATASET

EVERYONE & ALWAYS

The effect is due to **idiosyncrasies** in the experiment. **Stimuli are crucial! Data needs to be thoroughly understood!**

Influential variables influencing generalizability:

- Category
- Picture origin
- Lexical frequency
- Individual pictures' familiarity / recognizability (probably)
- (Participant)

The drive behind our commentary: open science principles

- **Genuine scientific interest in the phenomenon + open data** (kudos to the authors) that should serve a purpose: be explored by other researchers

- **Data is complex** and statistics may be blind to some complexities (if not modeled or when it can't be modeled; too many variables and few datapoints)

- **Clarify** the nature of an effect that was being cited as generalizable and given for granted

- **Discuss limitations and possible errors** with **no ethical stigma**, especially for ECRs

- But also: a bit worried about possible **negative effects** on us, but also on the first author of the original paper (an ECR like us!)

Mistakes
M[redacted]e happen

- We are afraid of **making mistakes**
- We are terrified of other people's finding out possible mistakes in our work = **stain in our career**
- We should be free of making genuine mistakes and acknowledge **limitations**

- Published science is not carved in stone but should be submitted to a **continuous process of collective correction**
- Making mistakes is **not** committing a fraud

The peer review process

- 8 months, 3 rounds: is it common for commentaries?
- "Too harsh", "too focused on picking on flaws", "should highlight more the positive aspects of the original work"...
- "It appears the original work offers no insights, this should be evidenced more strongly"
- Discussing construct validity is not relevant
- Too many supplementary figures

- Too much focus on Open Science for the journal scope, focus on *theoretical* contribution instead
  - Decision to eliminate "collaborative open science" from the title
  - Gradual limitation of OS first in Introduction and Discussion, then only in Discussion in a dedicated small paragraph

Did we approach it in the best way?

- We emailed the authors when we had a full manuscript and reproducible scripts; criticism: we should have contacted them earlier

- Informally, one of the senior authors knew we were working on a commentary (not informed by us and without us knowing beforehand); the other authors did not know until they read our email

- Availability on our part to modify before submission any sections in which we may have incorrectly reported their work, request to check fair reporting; denied, perceived by the authors as a request for thorough review of the work in little time

- New stimuli & naming data not shared: seen as an accusation, would have shared if contacted earlier

- Unfortunate timing: the first author was on maternity leave and rightfully had other priorities; however, none of the co-authors intervened to provide any feedback either

- Criticism that this is not collaborative open science

- Ultimately, the commentary and our emails were not well received, and our work was dismissed as not worthy of their time because of other priorities (especially by a senior author emailing us without including their coauthors: "my co-authors and I are all tied up with more pressing matters")

The power and perils of commentaries as ECRs

- Contribute to clarify & deepen understanding of **existing findings**
- Offer new ideas for **better constructs and measures**
- Promote a new approach in **facing uncertainty in a collective way**
- Boost the **usefulness of open data** (= real open science)

- Marked in the black book (or Deathnote) of someone

  You can even be blocked on X/Twitter! Don't tell anyone about this 🤫
- Risk of **negative reactions** due to a possible sense of **personal attack** (we tend to identify with our papers)

I want to hear from you now

- ❖What would you have done differently from us?
- ❖What is your experience with commentaries?
- ❖Have you ever been the target of a commentary?
- ❖Can we make commentaries less adversarial and more collegial (our attempt aimed at this direction)?

Readings & Resources

- Cirillo et al. (preprint; v1 Apr 21; v2 Dec 23): https://doi.org/10.31234/osf.io/cjy24

- Cirillo et al. (published, *Cognition*, Jul 22): https://doi.org/10.1016/j.cognition.2022.105213

- Original paper OSF repository: https://osf.io/f6gu3/

- Gastaldon & Calignano (preprint; v1 Jul 24; v2 Nov 24): https://doi.org/10.31234/osf.io/v9ufk

- Gastaldon & Calignano (published, *Cognition*, Feb 25): https://doi.org/10.1016/j.cognition.2025.106099

- Commentary OSF repository: https://osf.io/yeqgp/

- MultiPic: paper (https://doi.org/10.1080/17470218.2017.1310261) & database (https://www.bcbl.eu/databases/multipic)

PsyArXiv

MultiPic: A standardized set of 750 drawings with norms for six European languages

OSF

Thank you

(DISCLAIMER: Pictures with an * beside them were not actually used as stimuli in the experiment and are not taken from the MultiPic database. Since newly created pictures were not openly shared but some of the corresponding referents elicited aligned category responses, here we included pictures found on the web to exemplify such referents)