Experimental pragmatics

3. Organizing and presenting our data

Xavier Villalba

Dept. de Filologia Catalana & Centre de Lingüística Teòrica

Escola d'Estiu de Lingüística Catalana, 18-24 de juliol de 2021

1. Describing our experiment

2. Processing our data
   - Validity
   - Significance

3. Conclusions

Crucial point

If our experiment cannot be replicated, it is useless!

Replication crisis

Open Science Collaboration (2015):

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration[*,†]

+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

Article | Figures & Data | **Info & Metrics** | eLetters | PDF

Article Information

vol. 349 no. 6251

DOI: https://doi.org/10.1126/science.aac4716

Crucial point

Full disclosure refers to the process of describing in full the study design and data collected that underlie the results reported, rather than a curated version of the design, and/or a subset of the data collected. The need for disclosure is clear: in order to adequately evaluate results we need to know how they were obtained. For example, the informational value of a dependent variable exhibiting an effect of interest is different if only one variable was collected or if fifteen were. The probability of a single variable achieving $P < 0.05$ just by chance is 5%, but the probability of one of fifteen variables achieving $P < 0.05$ is 54%. It is obvious that cherry-picking one from fifteen variables invalidates the results unless it is clear that this has happened. If readers know, then they can adjust their interpretation accordingly. From this simple fact it follows that if authors do not tell us whether they collected one or fifteen variables readers cannot evaluate their research. Munafò et al. (2017)

Crucial point

- How did we obtain the data?
- How did we process the data?
- How did we analyze the data?

Predictions

Brunetti et al. (2020)

Hence, we predicted that ratings of sentences with right dislocation would be optimal with epithets and that they would decrease as the anaphoric link becomes weaker, up to the optional part bridging type, which would render the worst ratings. In contrast, for a sentence with left dislocation, the hypothesis is reversed: We expected the sentence to be preferred when the bridging link was less strong. In other words, the ratings of left and right were expected to peak at opposite ends of the scale of bridging possibilities.

Predictions

Mayol & Villalba (2018)

3.2 Predictions

Taking as a departing point the view represented Vallduví (1992), Villalba (2000: Chapter 3), and Bott (2007) that RD was better suited for closer anaphoric relationships, such as identity, we predicted that RD would be preferred in the most direct bridging types, namely those that required little or no deduction at all. Hence, EPITHET was clearly predicted to be better with RD than with LD, whereas at the other side of the scale, INDUCIBLE PARTS and OPTIONAL ROLE were predicted to be bad with RD, but fine with LD. As for less extreme cases, which involved a quite straightforward deduction step (HYPONYM, SET MEMBERSHIP, NECESSARY PARTS, and NECESSARY ROLE), we predicted LD to be preferred over RD by default.

Materials: stimuli

Mayol & Villalba (2018)

Two counterbalanced randomized lists were prepared with 42 target items (=3 sentences with LD + 3 sentences with RD × 7 bridging types) and 40 fillers. Since the experiment was posted in the web (Survey Monkey), it was preceded by a language proficiency questionnaire and a brief instruction section. 168 native Catalan speakers completed the experiment, in which they had to rate the acceptability of every item in a 10-point Likert scale, as shown in Figure 1. Participants evaluated one written target item at a time and the lists were constructed so that each person would only see each item either with a left-dislocation or with a right-dislocation.

Materials: stimuli

Figure 1

Materials: stimuli

Trotzke & Villalba (2020)

For each combination, there were four examples. To test whether participants understood the task of judging the minidialogues, we constructed four fillers we expected to get good judgments ('good' fillers), four fillers we expected to get bad judgments ('bad' fillers), and four fillers we expected to receive mixed judgments ('medium' fillers); see Appendix B.

Taken together, there were 36 stimuli in total; stimuli were divided into 2 lists, each consisting of 24 items. All items were designed in a strictly parallel fashion for the two languages (Catalan and German), but we also ensured that the items sounded most natural in the respective language by our choice of language-specific names, interjections, etc.

Materials: stimuli

We offer some examples for illustrating the structure of our items:

polar context + wh-exclamative

En Carles té un nou cap a la feina i parla amb un dels seus companys.
Company: "T'ha fet bona impressió el nou cap?"
Carles: "Déu meu! Que mesquí que és aquest paio!"

non-polar context + wh-exclamative

En Carles té un nou cap a la feina i parla amb un dels seus companys.
Company: "Quina impressió t'ha fet, el nou cap?"
Carles: "Déu meu! Que mesquí que és aquest paio!"

Materials: stimuli

We offer some examples for illustrating the structure of our items:

polar context + that-exclamative

En Carles té un nou cap a la feina i parla amb un dels seus companys.
Company: "T'ha fet bona impressió el nou cap?"
Carles: "Déu meu! Que n'és de mesquí aquest paio!"

non-polar context + that-exclamative

En Carles té un nou cap a la feina i parla amb un dels seus companys.
Company: "Quina impressió t'ha fet, el nou cap?"
Carles: "Déu meu! Que n'és de mesquí aquest paio!"

Materials: stimuli

When building the items, we were very careful to place the antecedent in focus position (the rightmost position in the main clause), and we opted for object dislocates to avoid unwanted readings, for Catalan object dislocates must be resumed by a clitic (unlike subjects), and objects are not clitic-doubled (unlike datives). Mayol & Villalba (2018)

(1) a. La meva cosina s'ha canviat de pis i ahir va fer una festa.
b. Els ha pintat de color taronja, els radiadors.

(2) a. La meva cosina s'ha canviat de pis i ahir va fer una festa.
b. Els radiadors, els ha pintat de color taronja.

Materials: stimuli

Full list in appendix and/or open repository: `https://osf.io/73bvq/`.

Materials: participants

Brunetti et al. (2020)

Sixty-seven native Catalan speakers participated in Experiment 1; they were university students at the Universitat Autònoma de Barcelona and at the Universitat Pompeu Fabra in Barcelona. All participants answered a version of the Bilingual Linguistic Profile (Gertken et al., 2014), adapted to Catalonia's situation, and 20 candidates were excluded because they were classified as Spanish language dominant speakers. Forty-seven participants were eventually included (38 women and 9 men), aged 17 to 51 years (mean age, 22.1).

Materials: participants

Trotzke & Villalba (2020)

We collected judgments from 34 native Catalan and 61 native German speakers; Catalan speakers were tested within the context of a university class, and German speakers were recruited through Clickworker's crowdsourcing service (https://www.clickworker.de), following previous literature on experimentation in pragmatics (Degen et al., 2019). Participants had to rate the acceptability of Speaker B's reactions on a scale ranging from 1 (= very bad) to 6 (= very good). All Catalan participants passed a version of the Bilingual Linguistic Profile (Gertken et al., 2014), adapted to Catalonia's situation, where several degrees of Catalan-Spanish bilingualism coexist, and we discarded any candidate who was classified as a Spanish-dominant speaker (see Appendix C).

Materials: procedure

Brunetti et al. (2020)

Participants had to give an acceptability judgment using a 10-point Likert scale. First they had to read the instructions and were tested for language dominance. Then, participants also had to rate three practice items. Only then did they start the actual experimental task. Data were collected using the web interface Ibex Farm (http://spellout.net/ibexfarm, Drummond, 2013).

Two keys

- Validity
- Significance

Validity

Our choices in preparing and conducting the experiment have consequences on the validity of our results:

- Was the number of participants sufficient?
- Was the sample representative?
- Was the sample homogeneous?

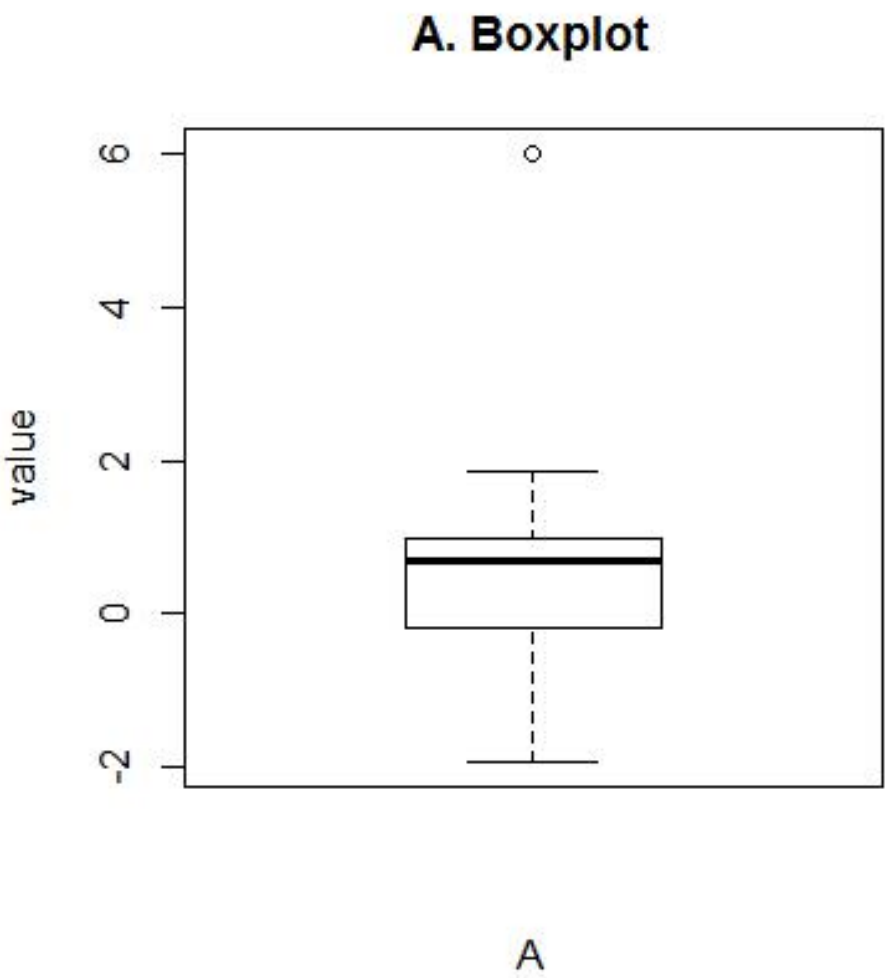
→ These considerations apply to corpus linguistics as well!

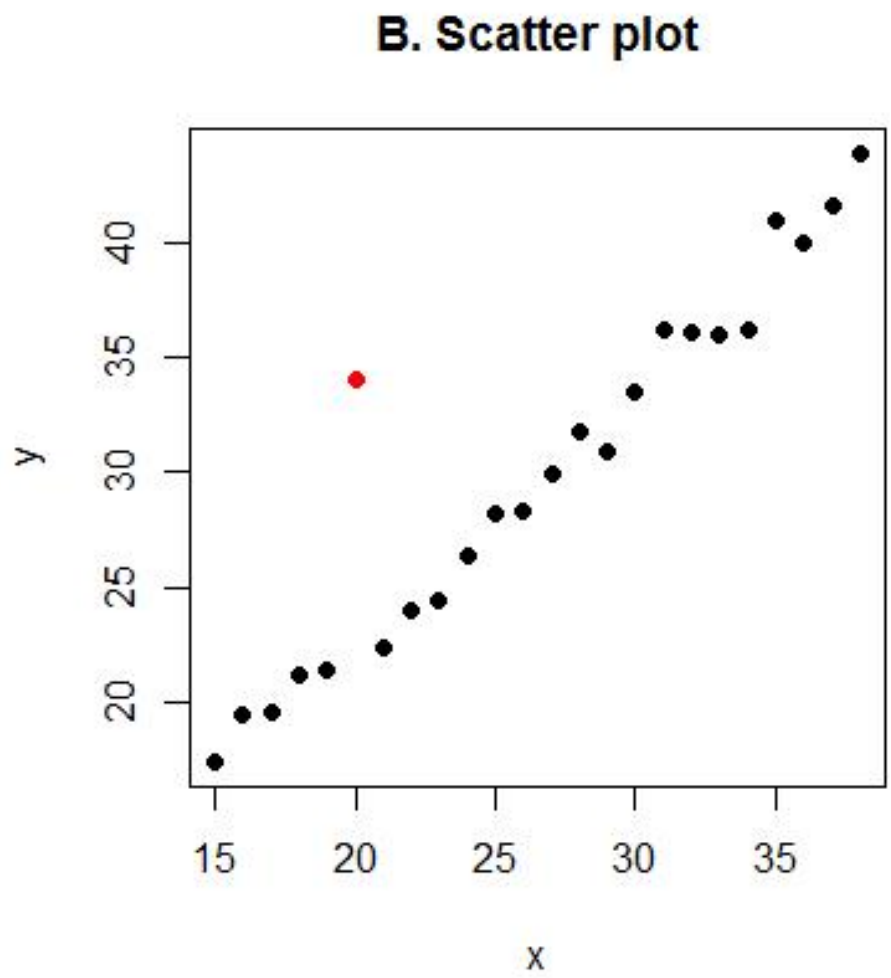
Validity

Yet, we can also sharpen our data to make them more general:

- discard unsuitable participants
- discard incomplete answers
- discard problematic items
- leave aside extreme values (outliers)
- normalize values

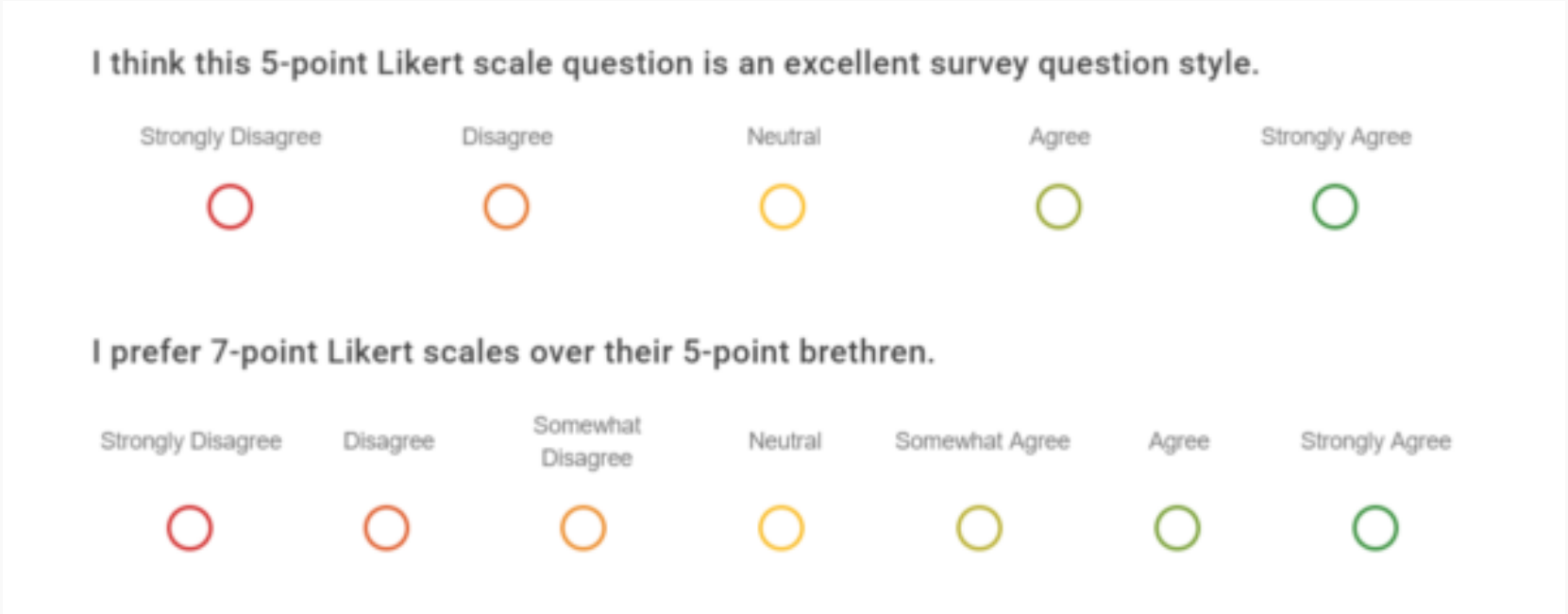
Outliers


A. Boxplot
value
6
4
2
0
-2
A


B. Scatter plot
y
40
35
30
25
20
15
20
25
30
35
x

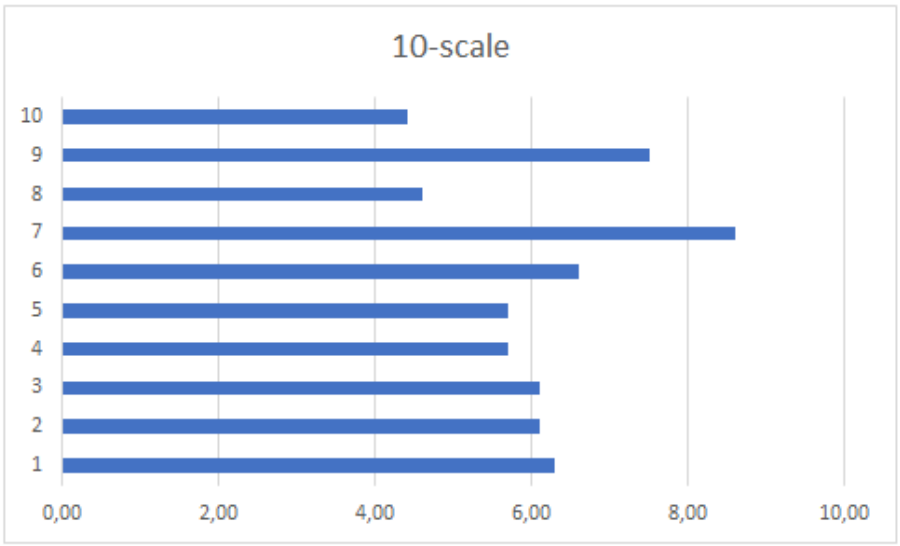
Validity: normalizing values

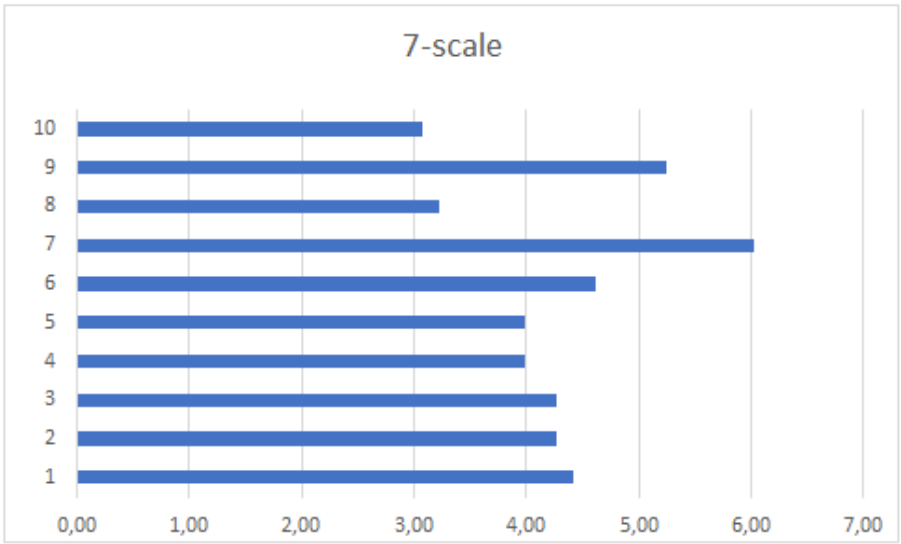
The results of our experiment are interpreted according to the numerical scale we use. So we can find two similar experiments differing in this respect:



How can we compare the data?

Validity: normalizing values



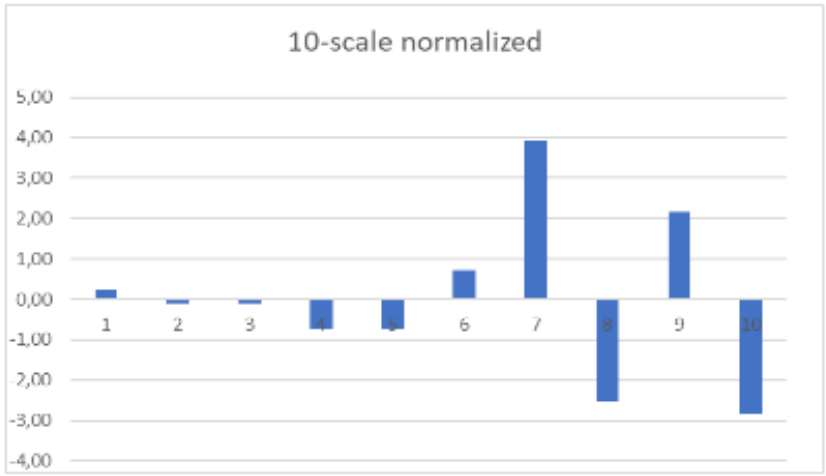


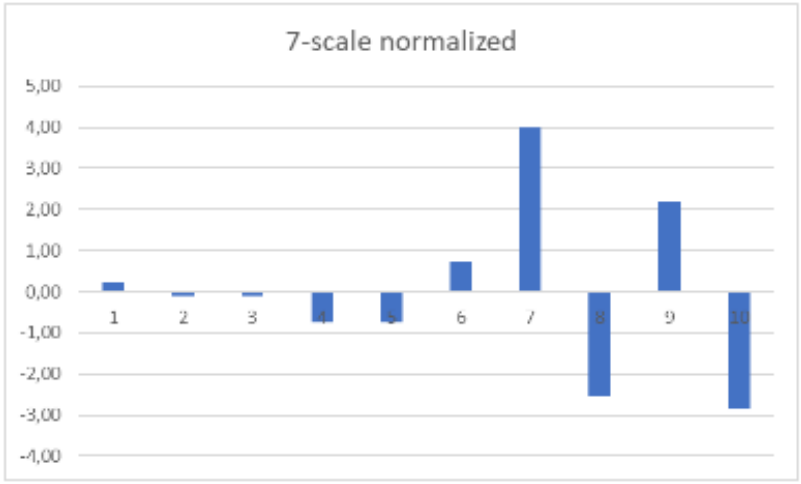
Validity: normalizing values

Normalized values (z-scores) are the result of transforming the obtained value into a measure of the distance of the value to the mean. This way we can show how prototypical our results are.

10-scale	s	z-score	7-scale	s	z-score
6,30	0,10	0,23	4,41	0,07	0,23
6,10	0,04	-0,10	4,27	0,03	-0,09
6,10	0,04	-0,10	4,27	0,03	-0,09
5,70	0,33	-0,74	3,99	0,23	-0,74
5,70	0,33	-0,74	3,99	0,23	-0,74
6,60	0,31	0,71	4,62	0,22	0,72
8,60	1,73	3,94	6,02	1,21	3,98
4,60	1,10	-2,52	3,22	0,77	-2,53
7,50	0,95	2,16	5,25	0,66	2,19
4,40	1,24	-2,84	3,08	0,87	-2,86
6,16	**0,62**		**4,31**	**0,43**	

Validity: normalizing values





Descriptive statistics

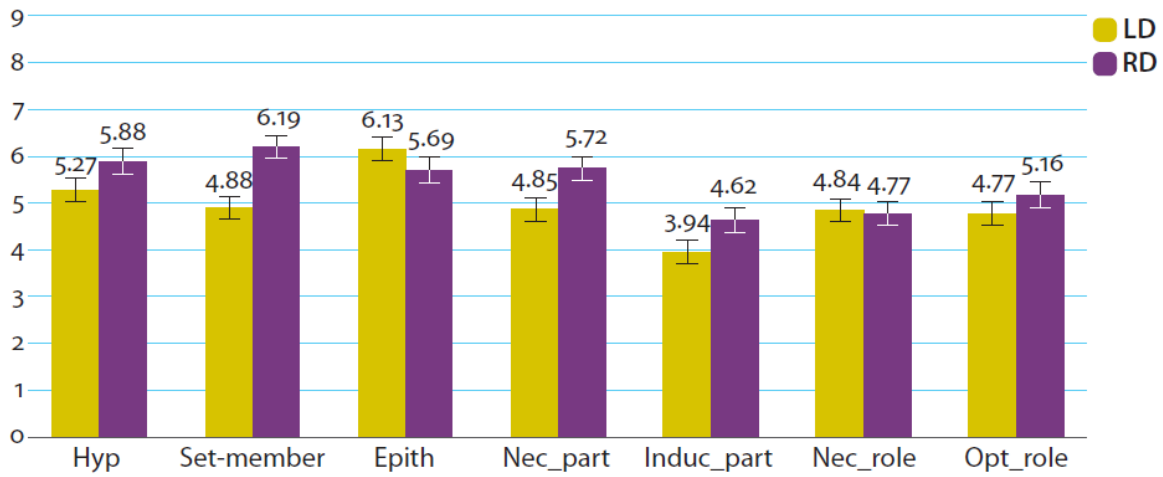


Figure 2. Means for LD and RD regarding bridging type

3.2 Predictions

Taking as a departing point the view represented Vallduví (1992), Villalba (2000: Chapter 3), and Bott (2007) that RD was better suited for closer anaphoric relationships, such as identity, we predicted that RD would be preferred in the most direct bridging types, namely those that required little or no deduction at all. Hence, EPITHET was clearly predicted to be better with RD than with LD, whereas at the other side of the scale, INDUCIBLE PARTS and OPTIONAL ROLE were predicted to be bad with RD, but fine with LD. As for less extreme cases, which involved a quite straightforward deduction step (HYPONYM, SET MEMBERSHIP, NECESSARY PARTS, and NECESSARY ROLE), we predicted LD to be preferred over RD by default.

Descriptive statistics

Frequency distribution (Catalan):

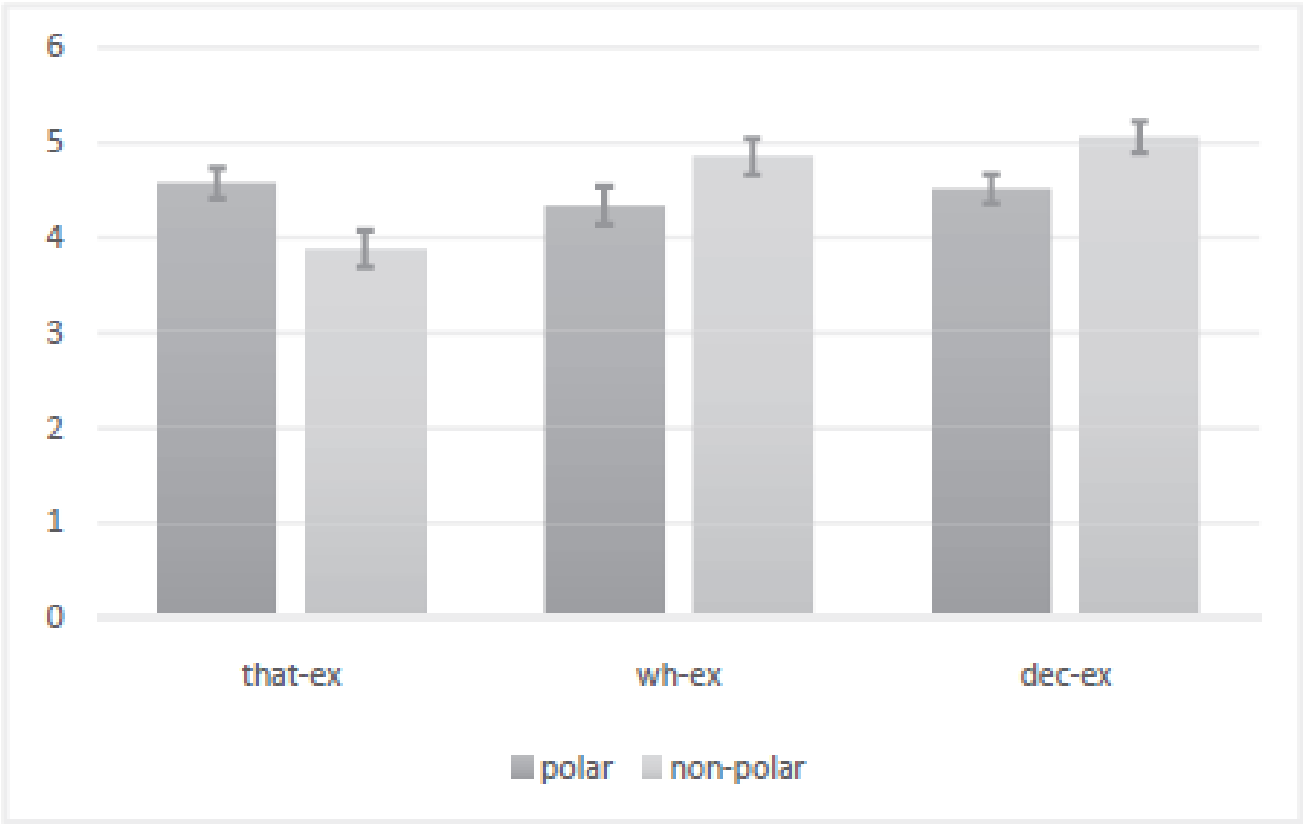


Fig. 3. Judgment of critical items (Catalan); whiskers represent SE.

Descriptive statistics

Frequency distribution (German):

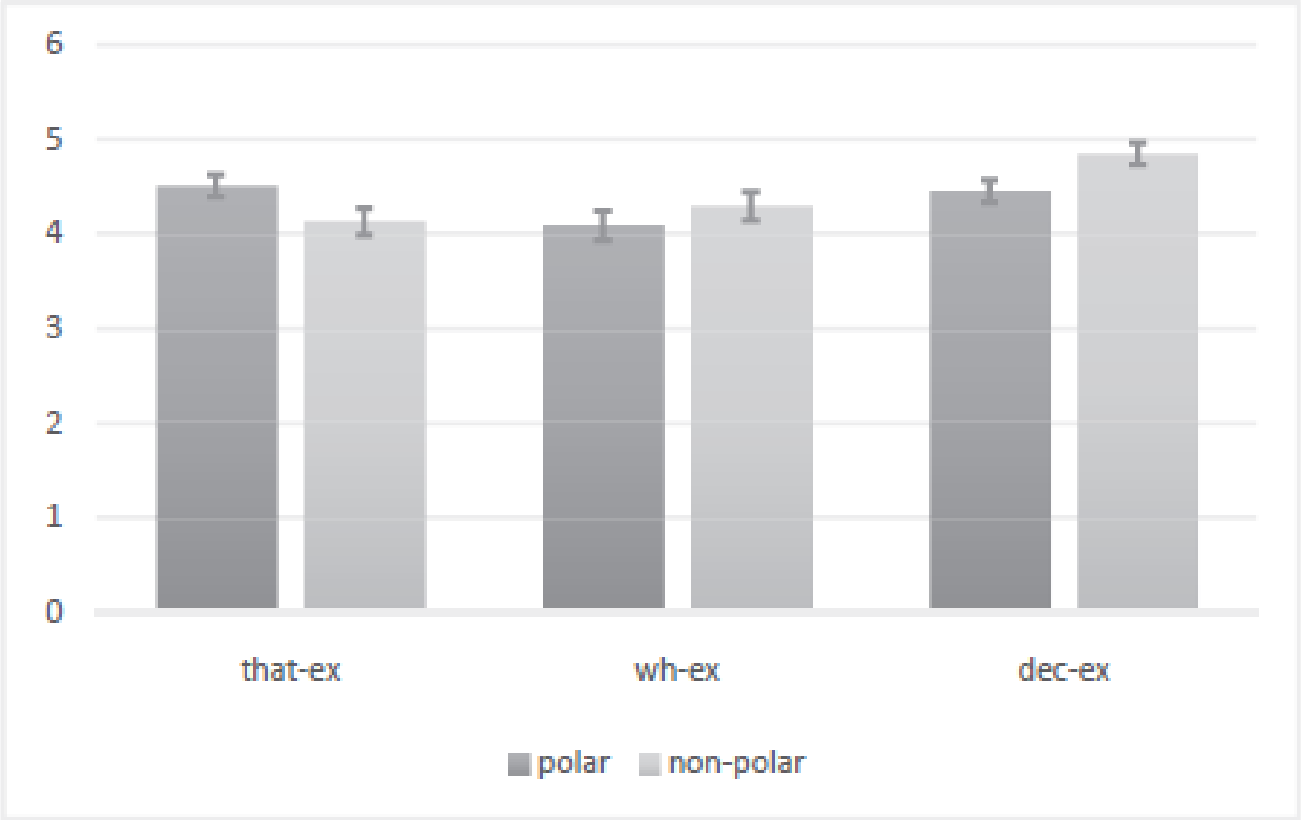


Fig. 4. Judgment of critical items (German); whiskers represent SE.

Significance tests

While mean distribution and standard error can suggest a clear interaction between our variables, further statistical analysis is needed to show the SIGNIFICANCE of the variation.

Null hypothesis

Even if our experiment shows differences in our independent variables, we must measure the degree of such differences. We know that some variation is inherent to any test, so by default, the null hypothesis is that our differences are due to chance.

- H_0 (null) = The independent value does not influence the dependent value.
- H_1 (alternative) = The independent value influences the dependent value.

Significance tests

The t-test and ANOVA (**An**alysis **of** **Va**riance) are the most common significance measures: they mesure the differences in means between different factors, so that the more different they are, the more probable is that the independent variable has a significant effect in the dependent variable [=not explained by the null hypothesis]. The result is a probability value (p-value) of obtaining our results if H_0 were true: if the p-value is small enough ($p < 0.05$), then we can reject the null hypothesis, and the difference is *statistically* significant.

Significance tests

Trotzke & Villalba (2020):
For Catalan, a two-way ANOVA (3x2) revealed a significant main effect of EXCLAMATION FORM ($F(2,39)=9.32$, $p<.001$) and a significant interaction of EXCLAMATION FORM and CONTEXT ($F(2,39)=9.43$, $p<.001$), but we found no significant effect of CONTEXT ($F(1,27)=.92$, $p<.05$).

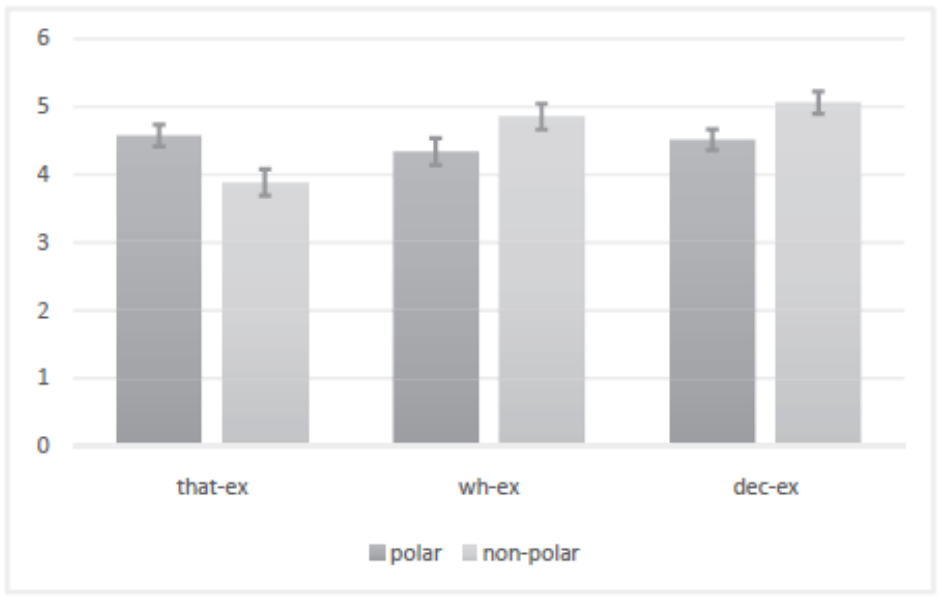


Fig. 3. Judgment of critical items (Catalan); whiskers represent SE.

Significance tests

Significance tests are a useful tool for interpreting the impact of our independent variables, but they are not the solution to all our variation.

- The p-value only shows the compatibility of the data observed with the null hypothesis.
- Even if a significant effect is found, The p-value does not directly reflect the size of the effect.

Significance tests: a true life story

Nuzzo (2014)

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white. The results were "plain as day", recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was sexy," he says, "and the data provided clear support." The *P* value, a common index for the strength of evidence, was 0.01 –usually interpreted as 'very significant'. Publication in a high-impact journal seemed within Motyl's grasp. But then reality intervened. Sensitive to controversies over reproducibility, Motyl and his adviser, Brian Nosek, decided to replicate the study. With extra data, the *P* value came out as 0.59 –not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame.

Significance tests: a true life story

Nuzzo (2014)

Last year, for example, a study of more than 19,000 people showed that those who meet their spouses online are less likely to divorce ($p < 0.002$) and more likely to have high marital satisfaction ($p < 0.001$) than those who meet offline (see *Nature* http://doi.org/rcg; 2013). That might have sounded impressive, but the effects were actually tiny: meeting online nudged the divorce rate from 7.67% down to 5.96%, and barely budged happiness from 5.48 to 5.64 on a 7-point scale.

Beyond significance tests

Discussion about including size effects measures, and confidence levels, which show the magnitude and relevance of effect.

Ask your family statistician!

Main ideas

- To favor reproducibility, include the the rellevant information about your experiment and data analysis in the paper, and an open repository `https://osf.io/73bvq/`.
- Do it in a standard and perspicuous way.
- Ask experts for advice.

References

Brunetti, L., Mayol, L. & Villalba, X. (2020), 'Bridging Strength, Monotonicity, and Word Order Choices in Catalan', *Discourse Processes* **57**(8), 703–724.

Mayol, L. & Villalba, X. (2018), Bridging and dislocation in Catalan, *in* L. Repetti & F. Ordóñez, eds, 'Romance Languages and Linguistic Theory 14. Selected papers from the 46th Linguistics Symposium on Romance Languages (LSRL), Stony Brook, NY', John Benjamins, Amsterdam / Philadelphia, pp. 201–213.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J. & Ioannidis, J. P. A. (2017), 'A manifesto for reproducible science', *Nature Human Behaviour* **1**(1), 0021.

Nuzzo, R. (2014), 'Scientific method: Statistical errors', *Nature* **506**(7487), 150–152.

Open Science Collaboration (2015), 'Estimating the reproducibility of psychological science', *Science* **349**(6251).

Trotzke, A. & Villalba, X. (2020), 'Exclamatives as responses at the syntax-pragmatics interface', *Journal of Pragmatics* **168**, 139–171.