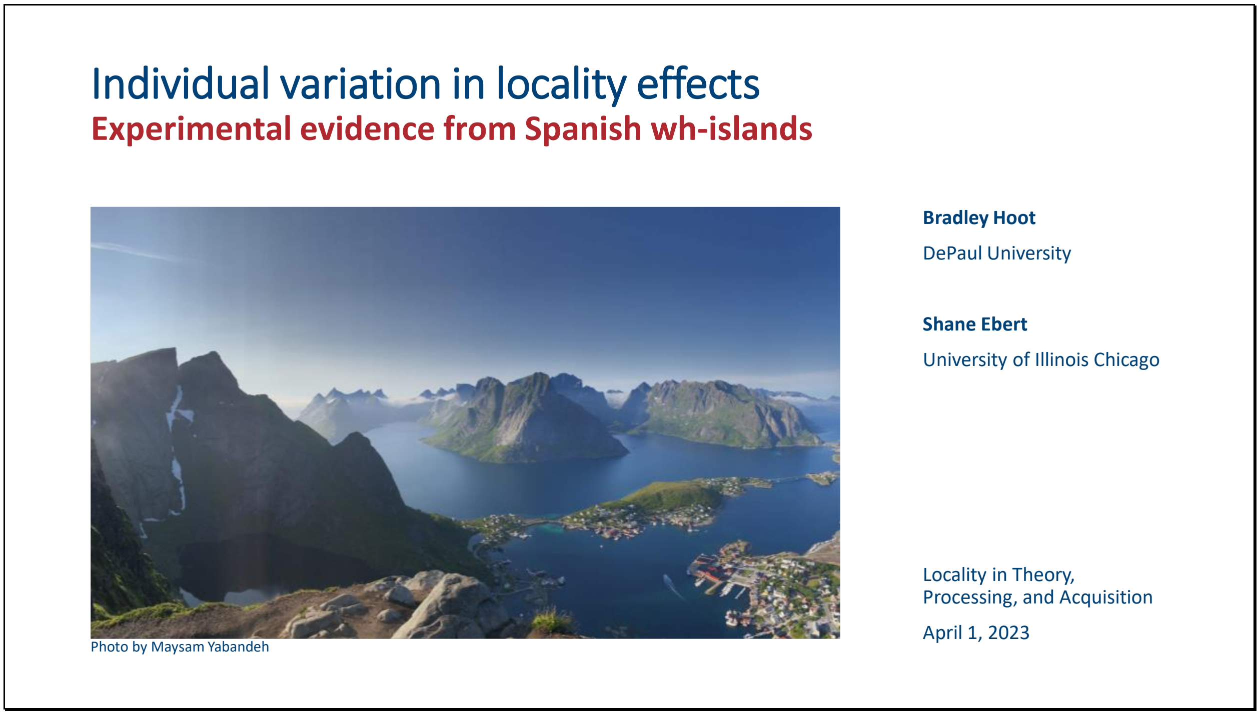


Individual variation in locality effects

Experimental evidence from Spanish wh-islands

Photo by Maysam Yabandeh

Bradley Hoot

DePaul University

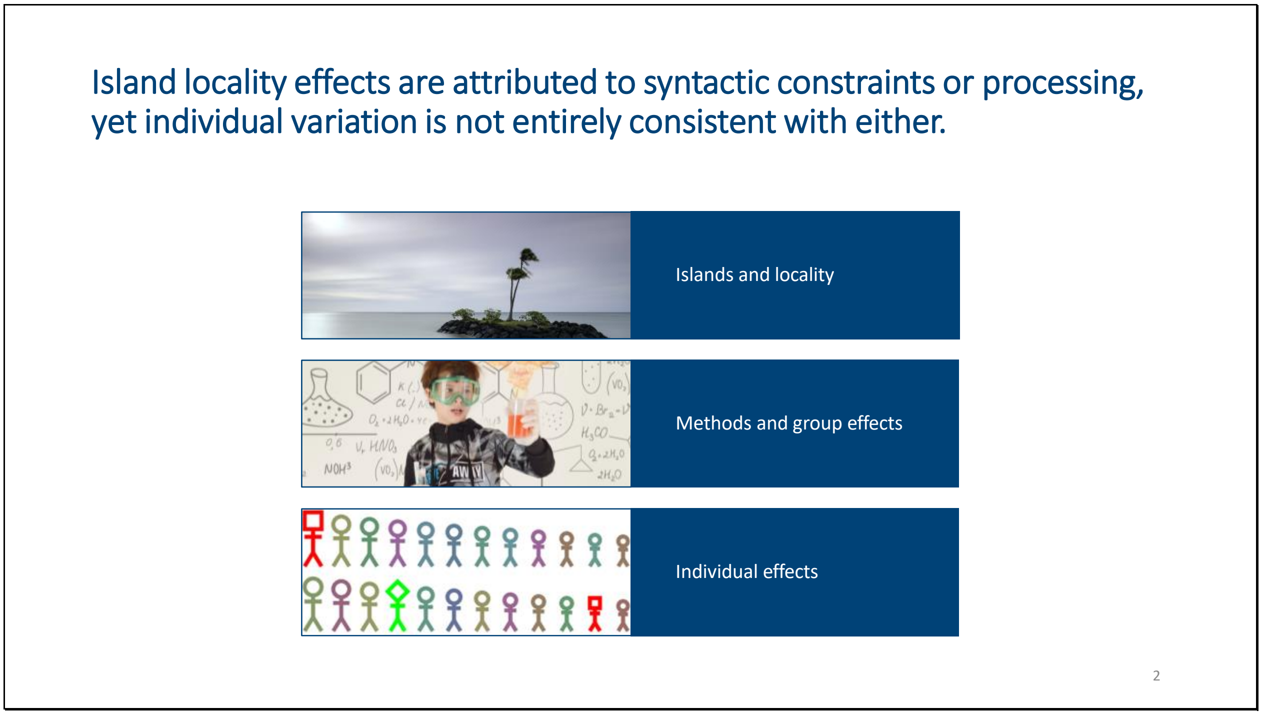
Shane Ebert

University of Illinois Chicago

Locality in Theory,
Processing, and Acquisition

April 1, 2023

Contact: bhoot@depaul.edu, sebert2@uic.edu.

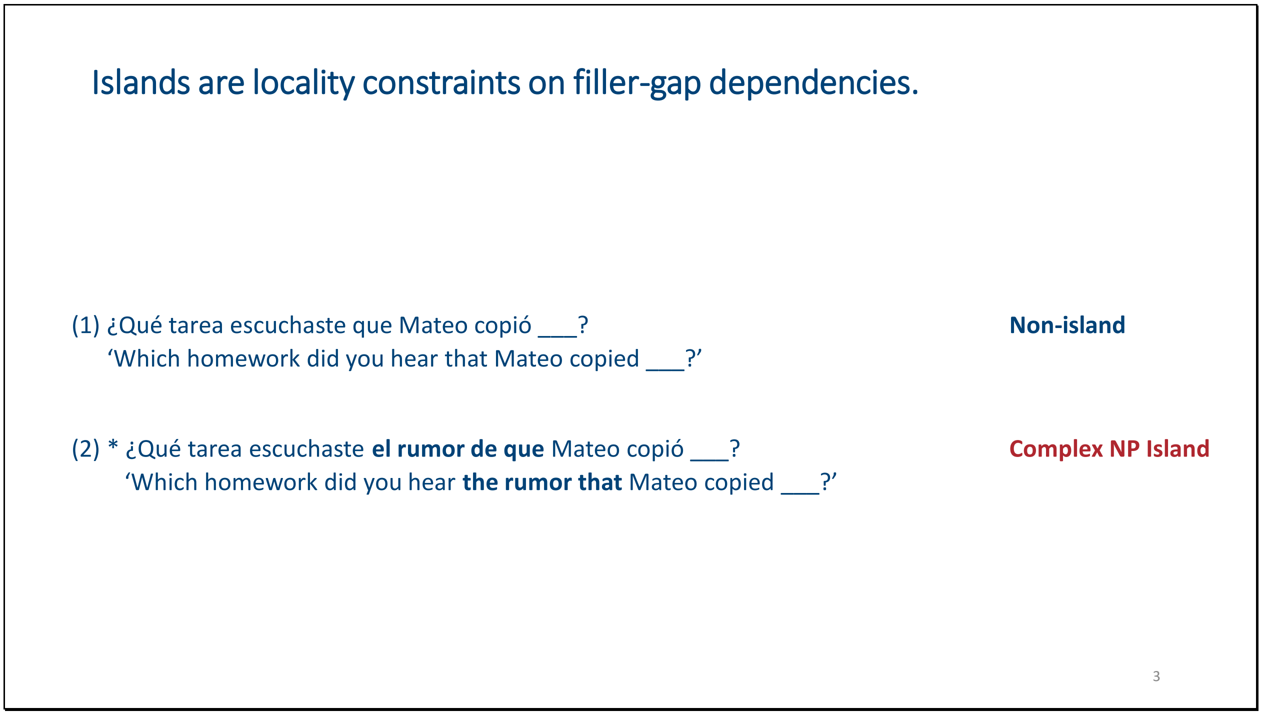


Island locality effects are attributed to syntactic constraints or processing, yet individual variation is not entirely consistent with either.

Islands and locality

Methods and group effects

Individual effects



Islands are locality constraints on filler-gap dependencies.

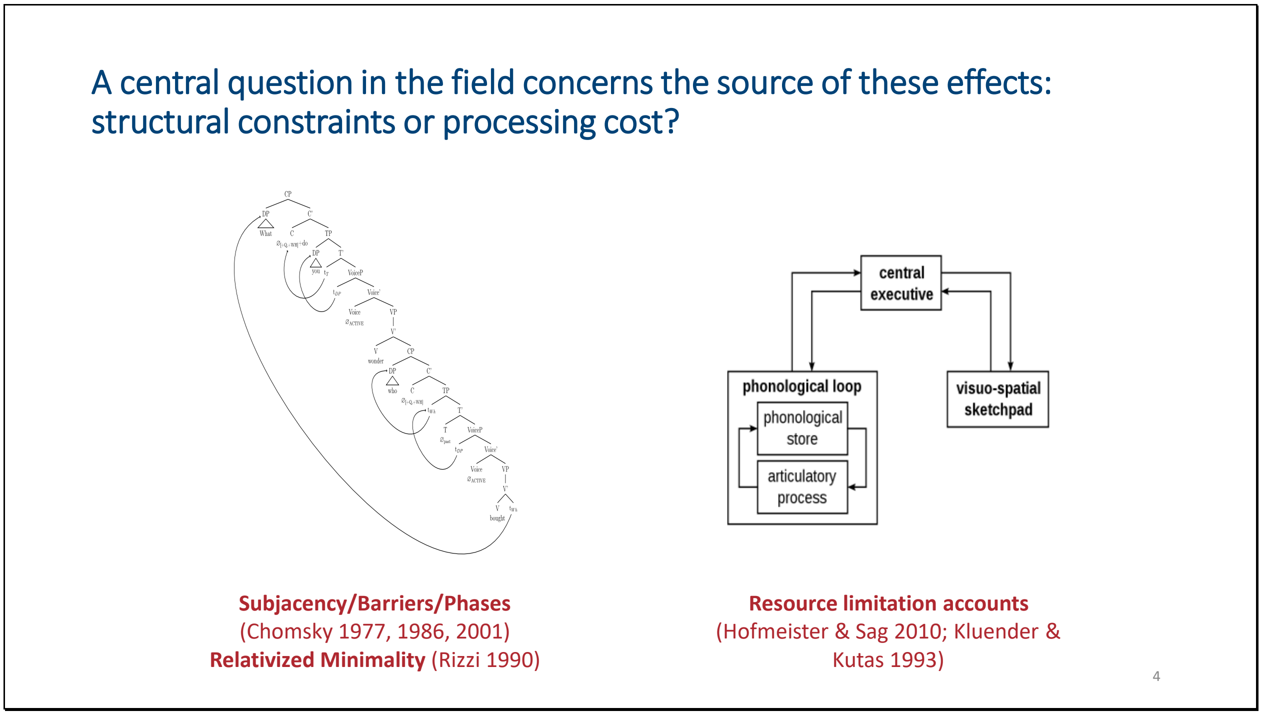
(1) ¿Qué tarea escuchaste que Mateo copió ___?
‘Which homework did you hear that Mateo copied ___?’ **Non-island**

(2) * ¿Qué tarea escuchaste **el rumor de que** Mateo copió ___?
‘Which homework did you hear **the rumor that** Mateo copied ___?’ **Complex NP Island**

Islands have received extensive attention in the literature since Ross (1967). For an overview, see Boeckx (2012), Citko (2016), and Szabolcsi & Lohndal (2017).

Wh-phrases can be extracted out of their clause quite productively, such as in (1)

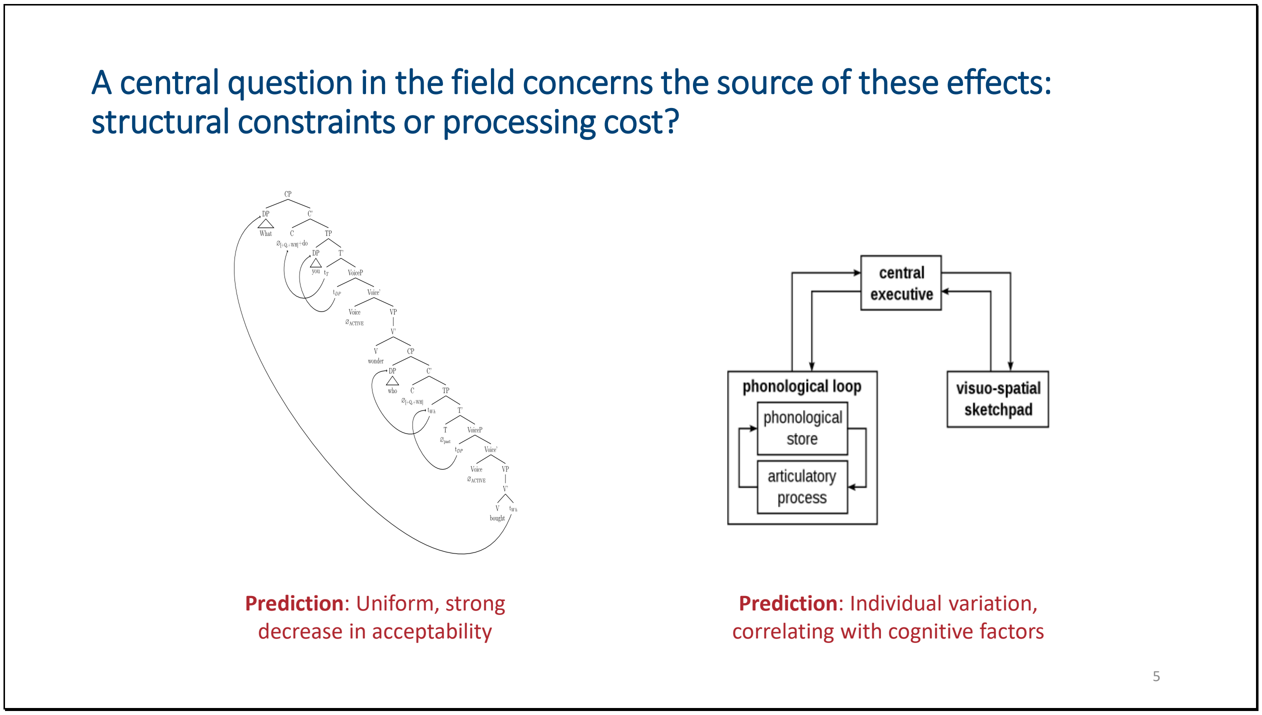
But in some contexts, there are constraints on how local that movement or filler-gap dependency needs to be. For example, you cannot move it out of a clause that’s part of a complex NP, as in (2).



A central question in the field concerns the source of these effects: structural constraints or processing cost?

Subjacency/Barriers/Phases
(Chomsky 1977, 1986, 2001)
Relativized Minimality (Rizzi 1990)

Resource limitation accounts
(Hofmeister & Sag 2010; Kluender & Kutas 1993)



A central question in the field concerns the source of these effects: structural constraints or processing cost?

Prediction: Uniform, strong decrease in acceptability

Prediction: Individual variation, correlating with cognitive factors

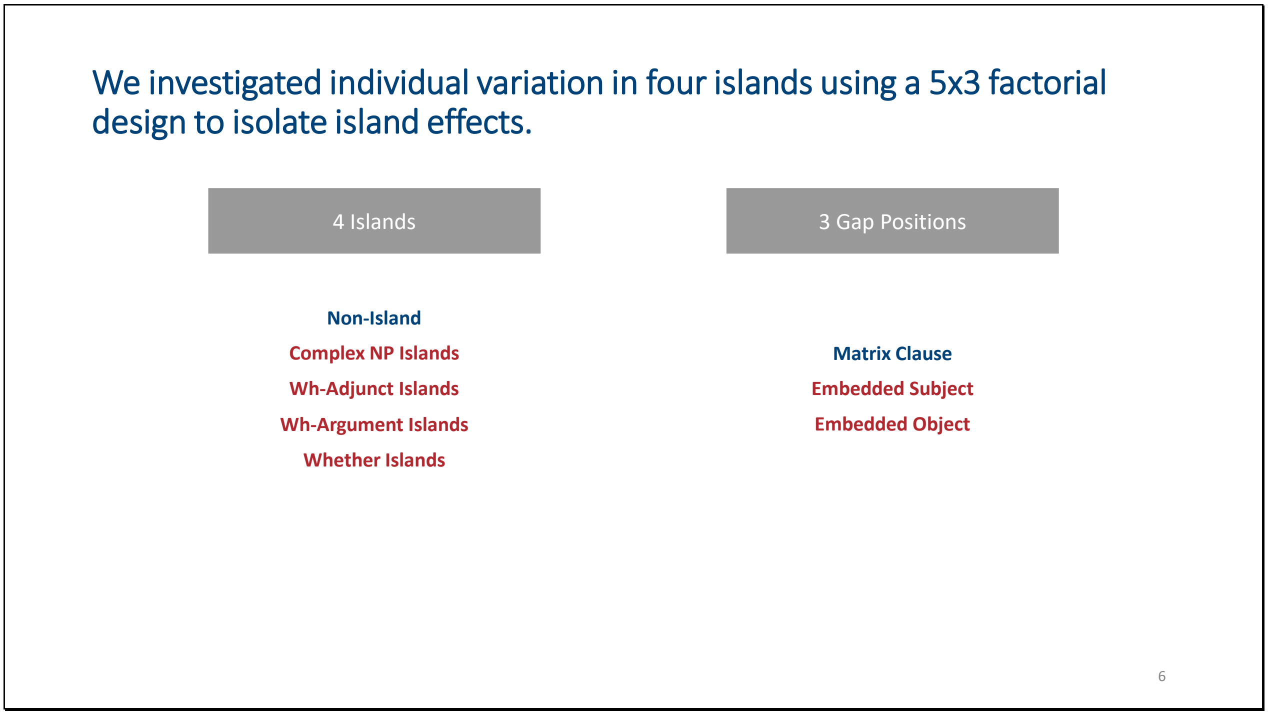
5

Syntactic accounts predict uniform, strong decreases in acceptability for island effects (Kush et al. 2019), although they may additionally be modulated by other factors.

Processing accounts account straightforwardly for variation by individual much as other language processing abilities vary by individual. If so, they also predict that the variation will correlate with variation on other measures of processing capacity (Sprouse et al. 2012).

Many studies have investigated correlations between island sensitivity and working memory, and largely have concluded that there is no relationship (e.g., Michel 2014; Pañeda et al. 2020; Pham et al. 2020; Sprouse et al. 2012). However, some claim these tests do not measure the right aspect of working memory because they include only recall and not processing (see Pham et al. 2020 for discussion), and others conclude that nearly all island effects are due to processing (Liu et al. 2022).

We report here on an experiment that can shed new light on this debate by thoroughly examining individual variation in island effects in Spanish, including a correlation with a cognitive measure that indexes both recall and processing.

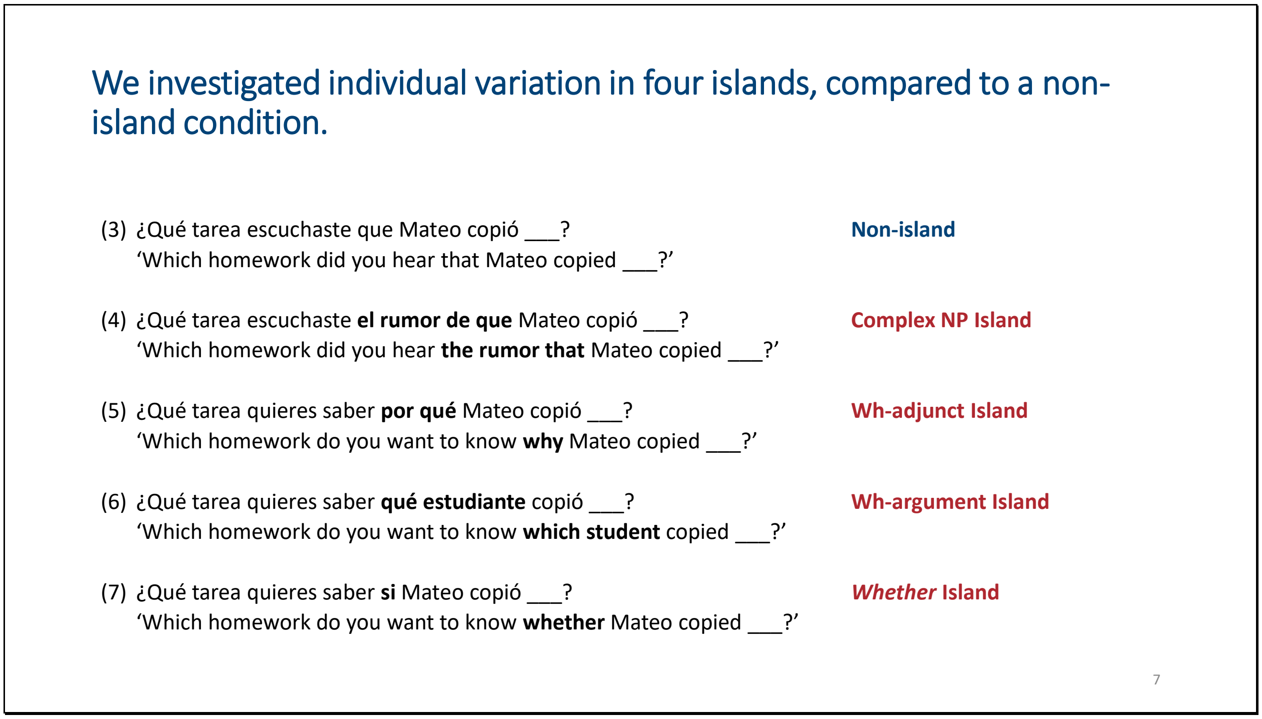


We investigated individual variation in four islands using a 5x3 factorial design to isolate island effects.

4 Islands	3 Gap Positions
Non-Island	
Complex NP Islands	**Matrix Clause**
Wh-Adjunct Islands	**Embedded Subject**
Wh-Argument Islands	**Embedded Object**
Whether Islands	

6

This design adapts the factorial design created by Sprouse and colleagues (Sprouse et al. 2012, 2016). We follow Stigliano & Xiang (2021) in comparing multiple islands against a single non-island condition in a larger design. We also expand the database by including extraction from both subject position and object position inside the embedded clauses.



We investigated individual variation in four islands, compared to a non-island condition.

(3) ¿Qué tarea escuchaste que Mateo copió ___? **Non-island**
‘Which homework did you hear that Mateo copied ___?’

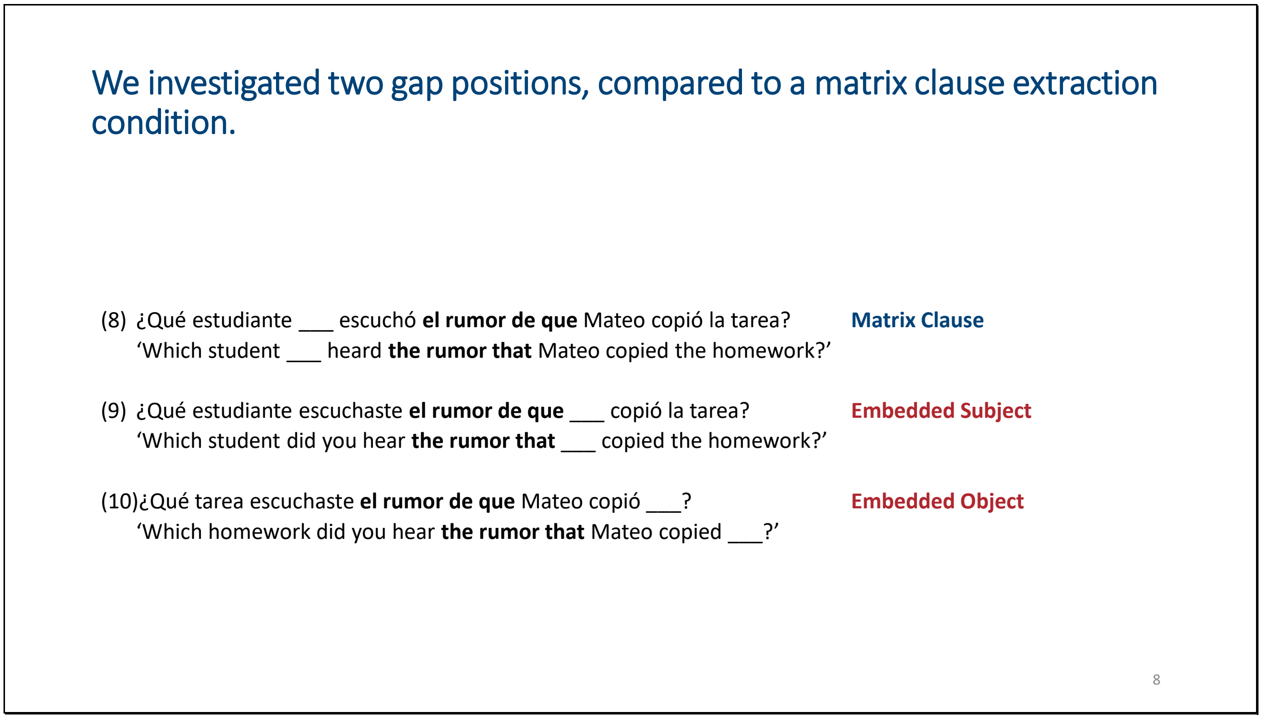
(4) ¿Qué tarea escuchaste **el rumor de que** Mateo copió ___? **Complex NP Island**
‘Which homework did you hear **the rumor that** Mateo copied ___?’

(5) ¿Qué tarea quieres saber **por qué** Mateo copió ___? **Wh-adjunct Island**
‘Which homework do you want to know **why** Mateo copied ___?’

(6) ¿Qué tarea quieres saber **qué estudiante** copió ___? **Wh-argument Island**
‘Which homework do you want to know **which student** copied ___?’

(7) ¿Qué tarea quieres saber **si** Mateo copió ___? ***Whether* Island**
‘Which homework do you want to know **whether** Mateo copied ___?’

Torrego (1984) claims (5) and (7) are grammatical in Spanish for extracting subject or objects, while (6) is grammatical in Spanish for extracting subjects, but ungrammatical for extracting objects. (4) is predicted to be ungrammatical for all extractions.



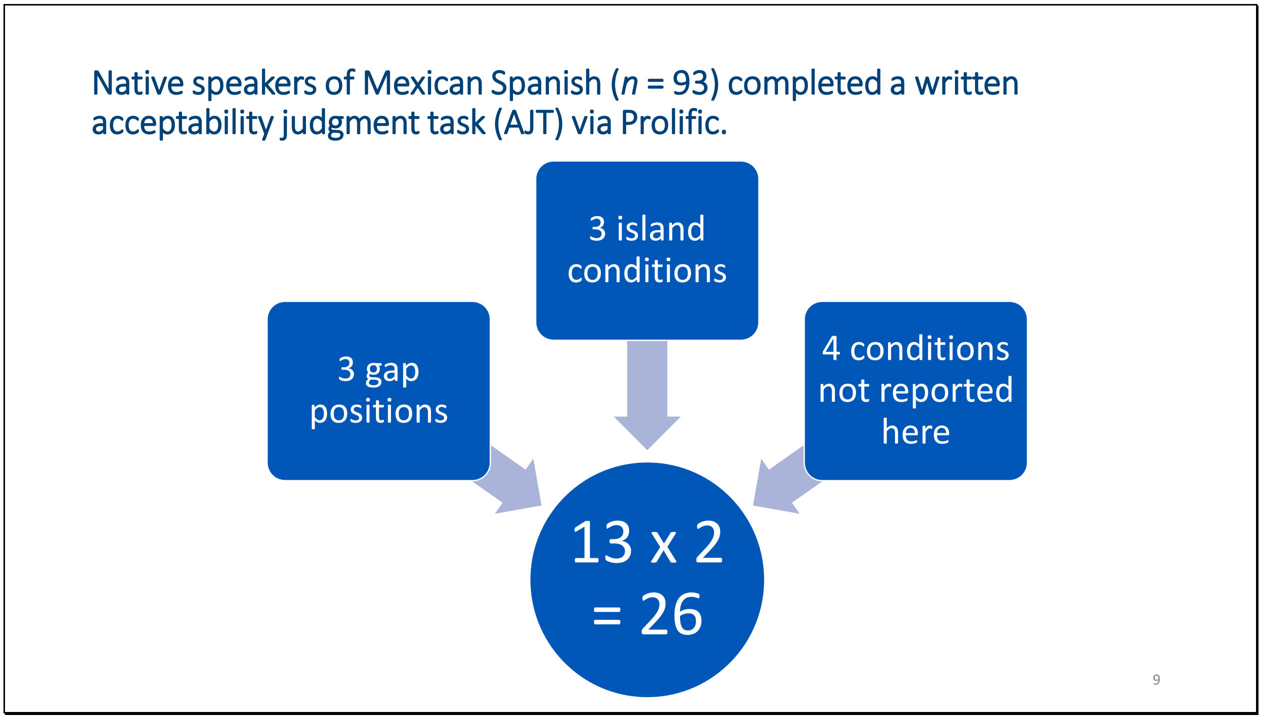
We investigated two gap positions, compared to a matrix clause extraction condition.

(8) ¿Qué estudiante ___ escuchó **el rumor de que** Mateo copió la tarea? **Matrix Clause**
‘Which student ___ heard **the rumor that** Mateo copied the homework?’

(9) ¿Qué estudiante escuchaste **el rumor de que** ___ copió la tarea? **Embedded Subject**
‘Which student did you hear **the rumor that** ___ copied the homework?’

(10)¿Qué tarea escuchaste **el rumor de que** Mateo copió ___? **Embedded Object**
‘Which homework did you hear **the rumor that** Mateo copied ___?’

We tested both subjects and objects because these are predicted to be different for at least some islands (Torrego 1984).



Native speakers of Mexican Spanish (*n* = 93) completed a written acceptability judgment task (AJT) via Prolific.

3 island conditions

3 gap positions

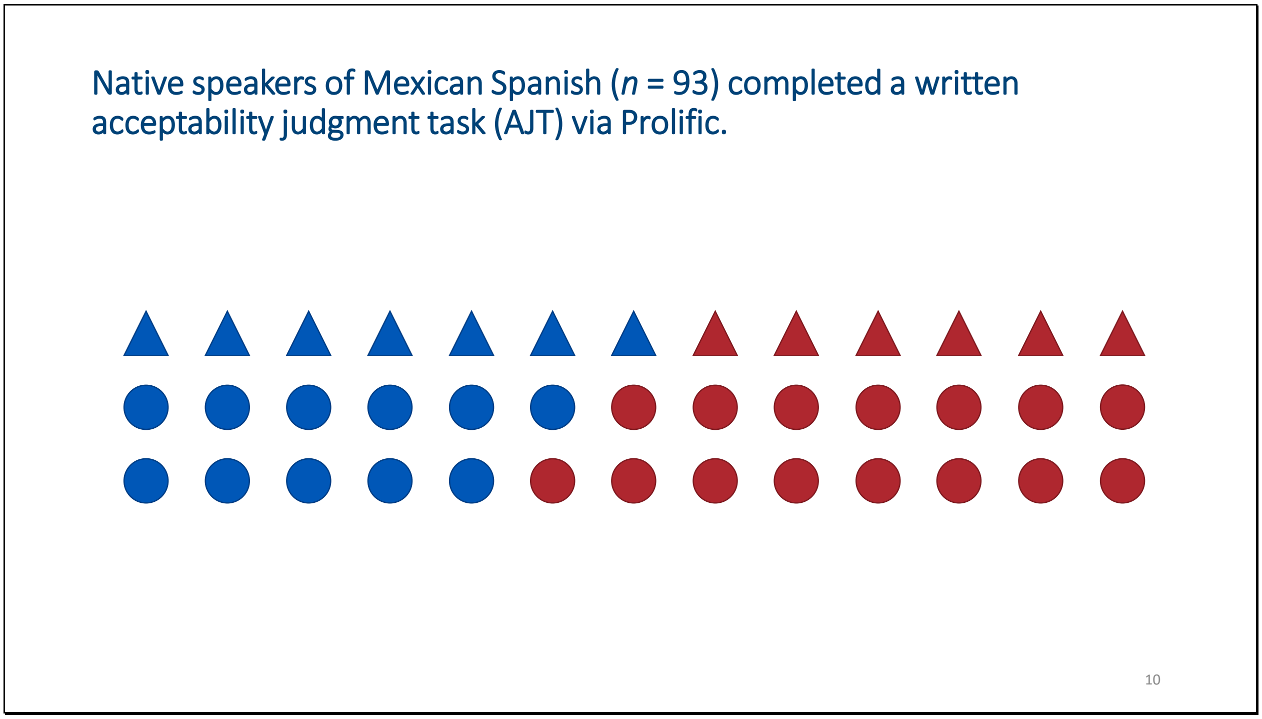
4 conditions not reported here

13 x 2 = 26

We divided the conditions into two tasks for the sake of length, so each person judged 2 wh-islands and the baseline non-island condition, as well as some other island conditions we don’t report on here.

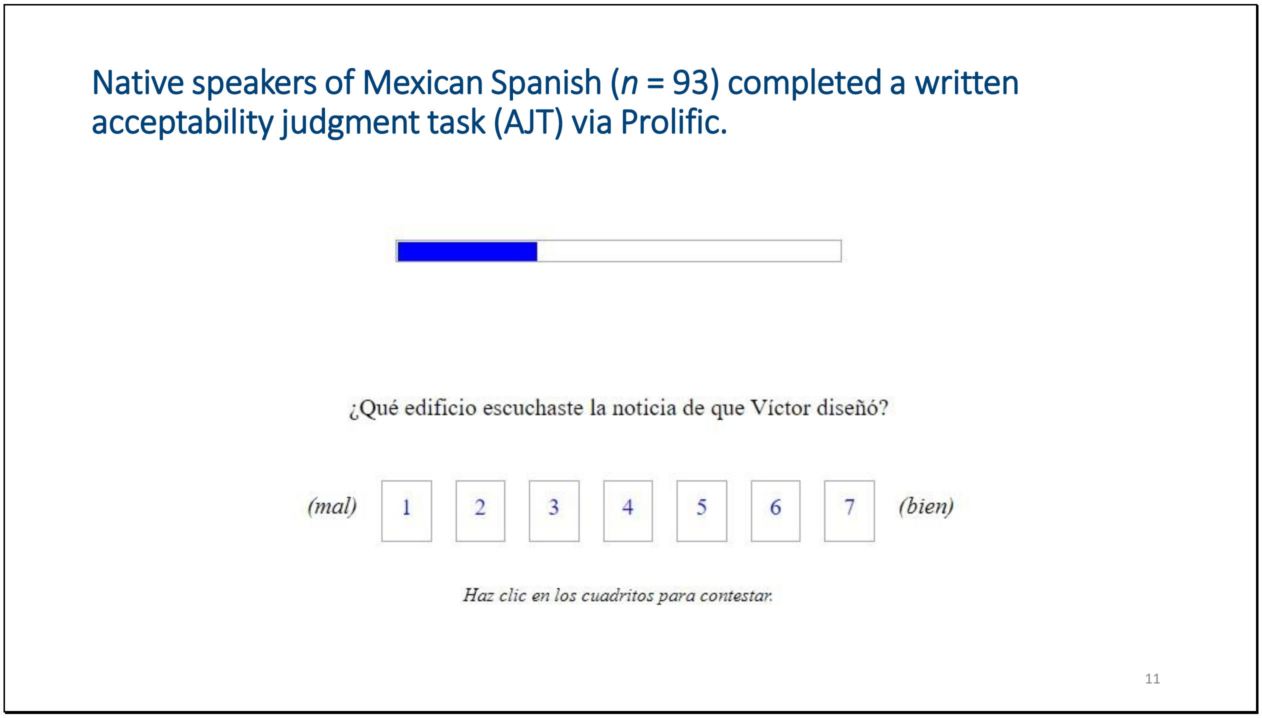
Each person therefore judged sentences in 13 conditions. They judged 2 sentences per condition, 26 target sentences total.

Sentences were distributed by Latin square, so no lexicalization was repeated for any individual.



Native speakers of Mexican Spanish (*n* = 93) completed a written acceptability judgment task (AJT) via Prolific.

Additionally, they judged 52 fillers for a 2:1 ratio of fillers to target, a 1:1 overall ratio of grammatical to ungrammatical, and items with the full range of acceptability.



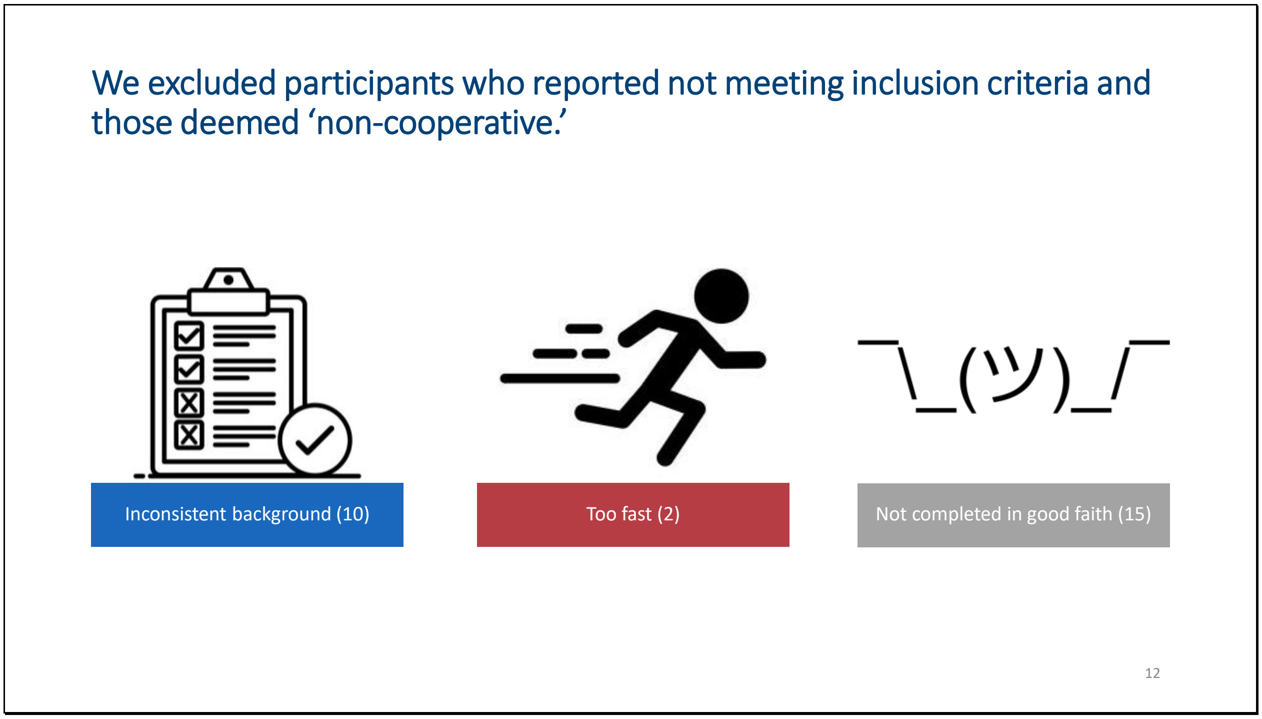
Native speakers of Mexican Spanish (*n* = 93) completed a written acceptability judgment task (AJT) via Prolific.

¿Qué edificio escuchaste la noticia de que Víctor diseñó?

(mal) □1 □2 □3 □4 □5 □6 □7 *(bien)*

Haz clic en los cuadritos para contestar.

Participants judged the sentences on a seven-point scale without accompanying contexts, as shown here.



We excluded participants who reported not meeting inclusion criteria and those deemed ‘non-cooperative.’

Inconsistent background (10)

Too fast (2)

¯_(ツ)_/¯

Not completed in good faith (15)

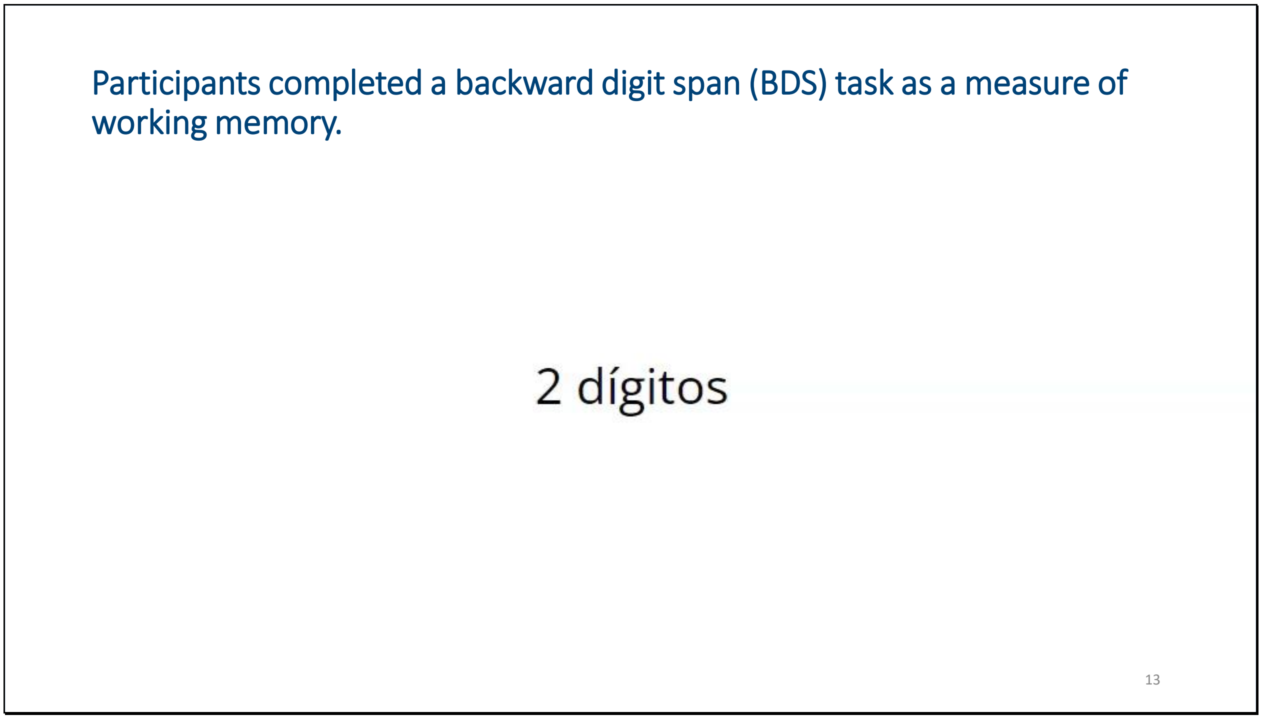
12

We excluded participants for three reasons:

- Responses inconsistent with their pre-screeners, such as growing up speaking more than one language, or some other irregularity = 10
- ‘Non-cooperative’ participants who responded to more than 20% of sentences below a reasonable threshold for the time required to read the sentence and make a judgment (calculated as 1200 ms, following Häussler & Juzek 2021, Juzek 2016) = 2
- Participants who did not complete the task correctly, as indicated by:
    - mean ratings for the ungrammatical filler sentences at the midpoint of the scale (4) or higher (following Pañeda & Kush 2022) = 8
    - ratings on the opposite side of the scale for two or more of three clear attention check items (which have clear ratings of 1 or 7, what Juzek 2016 calls ‘booby-trap items) = 7

Additionally, we included three ‘instructional manipulation checks’ (i.e., “please select 3”), but no participants were excluded for this reason.

After all exclusions, 93 participants remained in the sample, of whom 48 completed task 1 and 45 completed task 2. They all acquired Spanish before age 5 without any other home language, were born and educated in Mexico, and lived in Mexico at the time of testing. Although those who reported other languages spoken by caregivers in childhood were excluded, remaining participants necessarily know some English, because English is required to navigate Prolific’s website, so although they are native speakers of Mexican Spanish, they are not monolingual. Fifty-three were female, 40 male, and none non-binary or another gender identity. Their mean age was 25.6 years (range: 19-51).



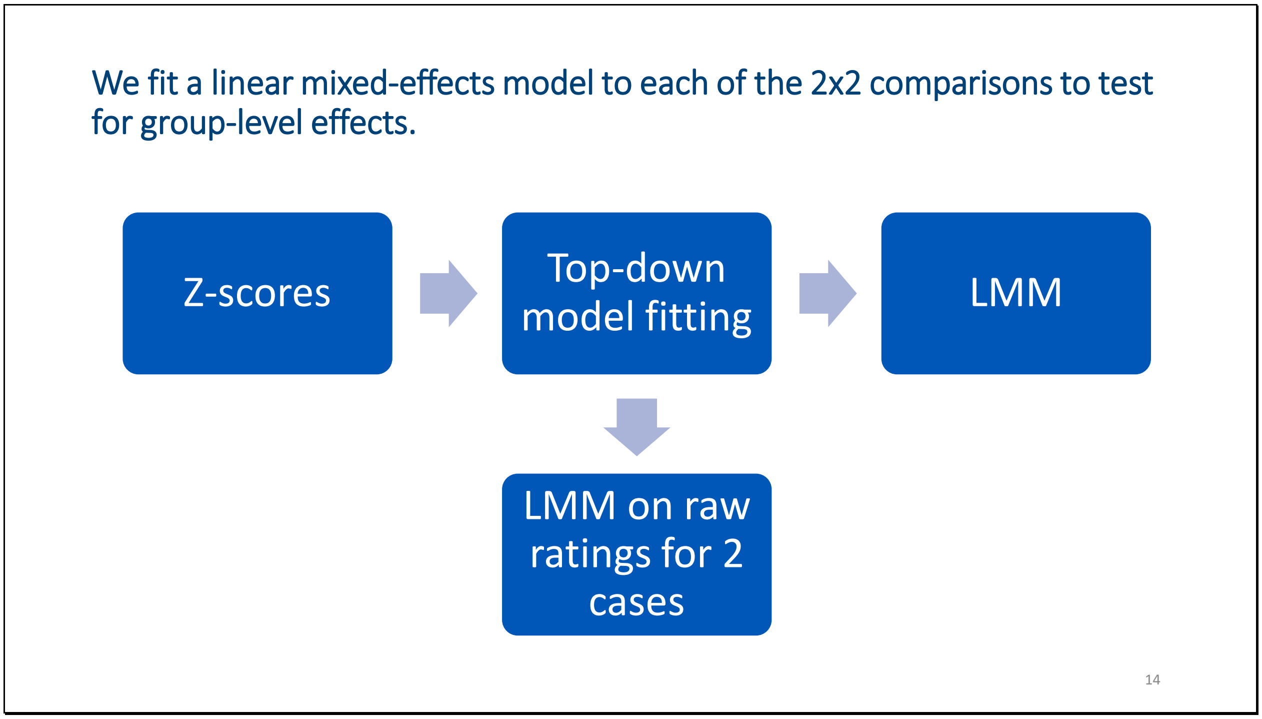
Participants completed a backward digit span (BDS) task as a measure of working memory.

2 dígitos

13

Some early tests of the resource-limitation view of islands (i.e., Sprouse et al. 2012) were criticized for using “simple span tasks which do not include both storage and processing components” (Pham et al. 2020:4). The backward digit span task includes both recall and a processing operation.

In this task, a sequence of single-digit numbers is presented serially in the center of the screen and participants type out the sequence in reverse order after it concludes. It begins by presenting two sequences of length 2, then two sequences of length 3, and so on up to length 8. Each correctly recalled inverse sequence scores one point. When a participant fails to correctly recall both sequences of a given length, the task ends. This task produces a score from 0 to 16.



We fit a linear mixed-effects model to each of the 2x2 comparisons to test for group-level effects.

Z-scores

Top-down model fitting

LMM

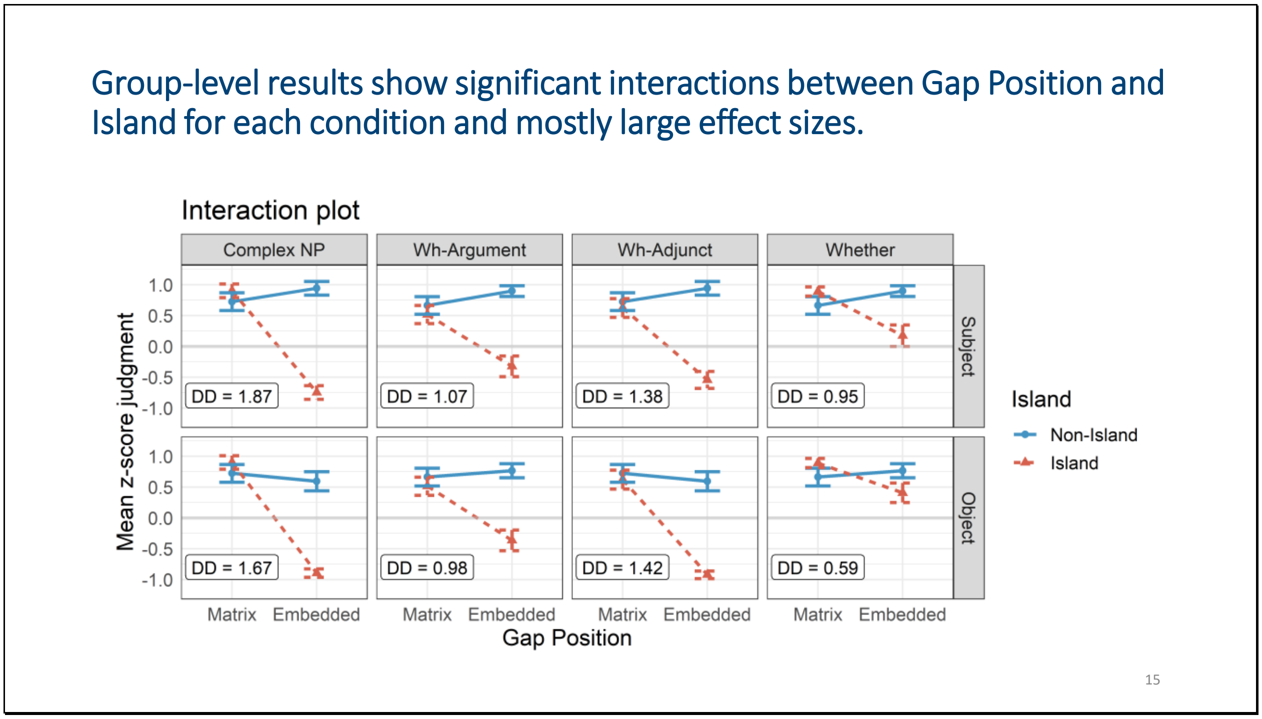
LMM on raw ratings for 2 cases

14

We z-score transformed the raw ratings by participant (Schütze & Sprouse 2013) to address scale compression and skew.

We fit a linear mixed-effects model to each of the 2x2 comparisons, using a top-down model-fitting procedure that results in the maximal random effects structure that fits the data. This minimally includes a random intercept by participant. However, in two cases (subject extraction for complex NP and wh-argument islands) even that random effects structure failed to converge; we instead carried out the statistical test on the raw ratings, instead of the z-scores, rather than conducting the test without accounting for repeated measures via random effects. The reason these two models did not converge with the z-scores is that the z-score transformation had already removed essentially all the by-participant variation, so a random effect by participant could not be modeled. Given that both the z-score transformation and LMMs have become essentially standard in experimental syntax, it is worth having a conversation about how we choose to address by-participant variance in our statistical models.

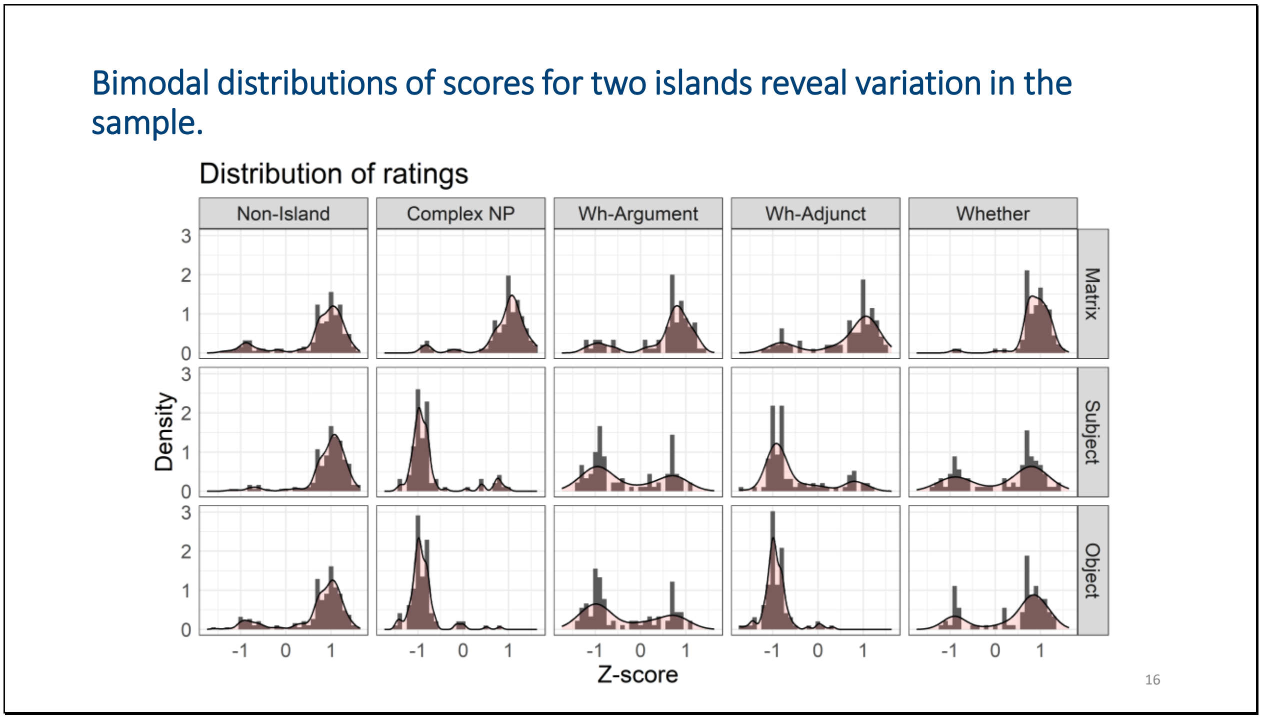
Nevertheless, the interactions are clear here, and we report the z-score data in the graph.



Group-level results show significant interactions between Gap Position and Island for each condition and mostly large effect sizes.

All these interactions are statistically significant at the group level.

We calculated a differences-in-differences (DD) score as a measure of effect size (Maxwell & Delaney 2004). Kush et al. (2018) characterize DD scores above 1 as 'large' effects.

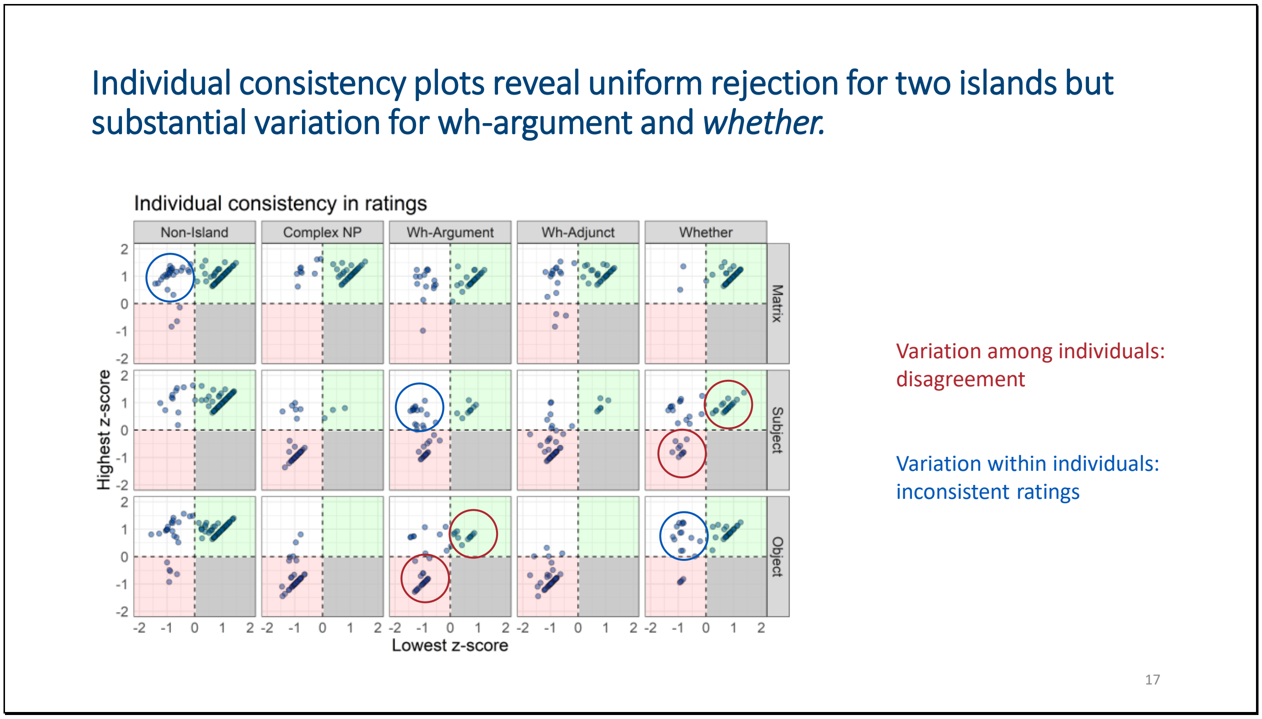


Bimodal distributions of scores for two islands reveal variation in the sample.

16

To better understand these effects, we can examine “second-order acceptability effects” (Kush et al. 2019), including the scores’ distribution by group and individual.

First, we examine the distribution of the scores at the group level by examining histograms (with overlaid density plots) of the z-scores for each condition.



Individual consistency plots reveal uniform rejection for two islands but substantial variation for wh-argument and *whether.*

Variation among individuals: disagreement

Variation within individuals: inconsistent ratings

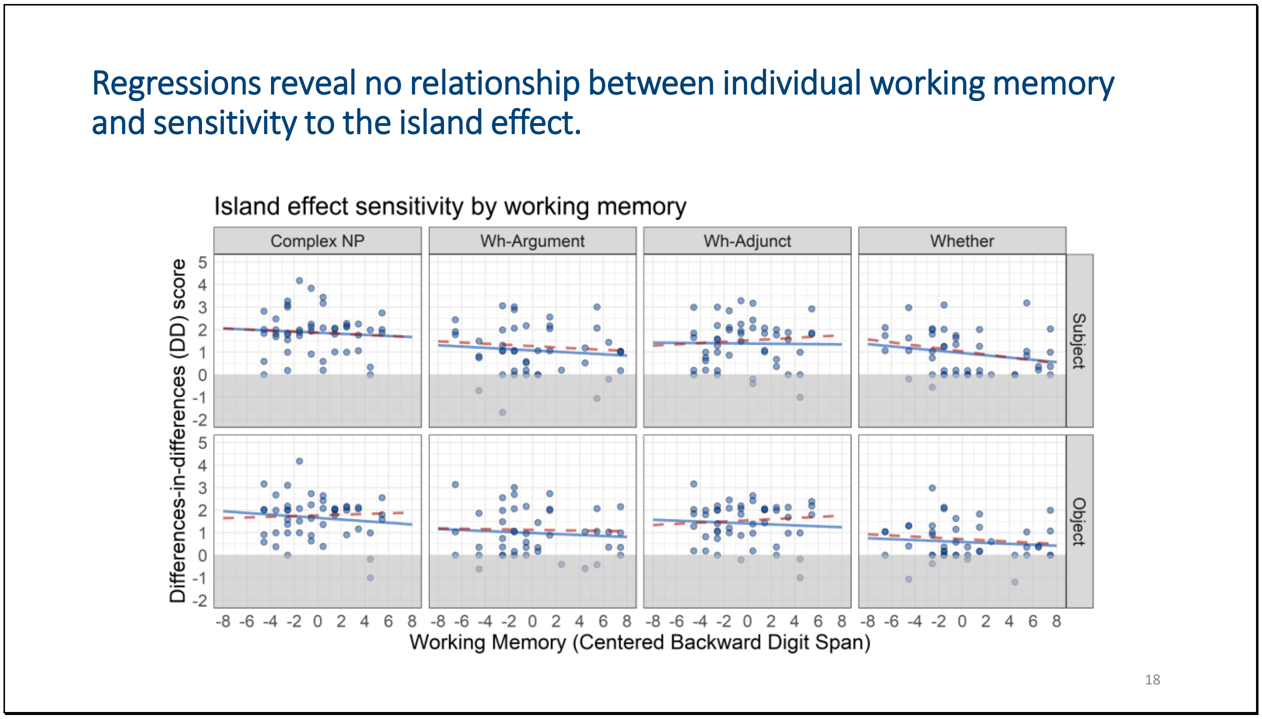
17

Second, we examine individual consistency by plotting each person's highest rating against their lowest rating (following Pañeda & Kush 2022).

Each person gave two ratings for each condition. Each person gave two ratings for each condition. Those with z-scores above 0 for both sentences in a condition—consistent acceptors—appear in the upper right (green) quadrant. Those with z-scores below 0 for both sentences in a condition—consistent rejectors—appear in the lower left (red) quadrant. Those who split their ratings—inconsistent raters—appear in the upper left (white) quadrant. Those who split their ratings—inconsistent raters—appear in the upper left (white) quadrant.

Note that we observe variation both between individuals and within individuals.

It is also noteworthy that some participants are inconsistent in their ratings of non-island structures as well.



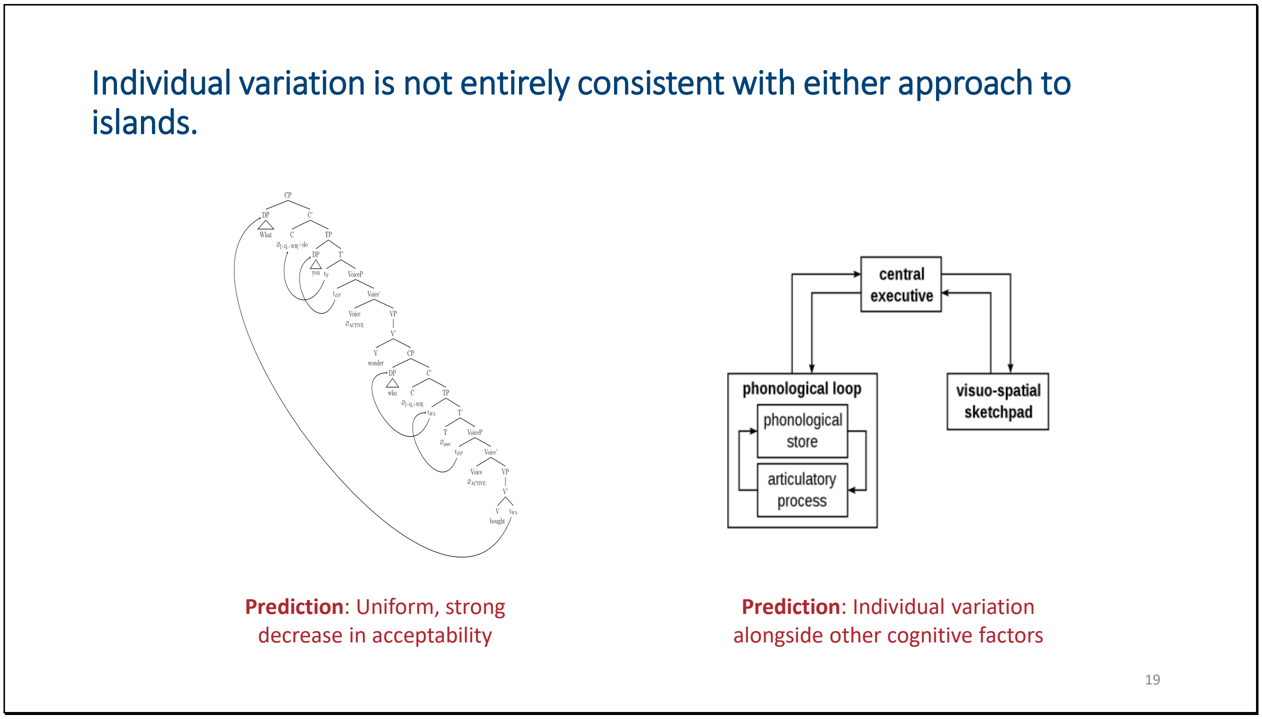
Regressions reveal no relationship between individual working memory and sensitivity to the island effect.

18

To investigate whether individual differences correlated to working memory scores, we carried out a linear regression. Nothing approaches significance.

We ran it twice, once with all the scores (blue solid line) and once with scores below 0 removed (red dashed line), following Pham et al. (2020).

The important thing about this measure is that the BDS indexes both recall and processing, avoiding some of the pitfalls of earlier attempts to measure working memory as it relates to islands.



Individual variation is not entirely consistent with either approach to islands.

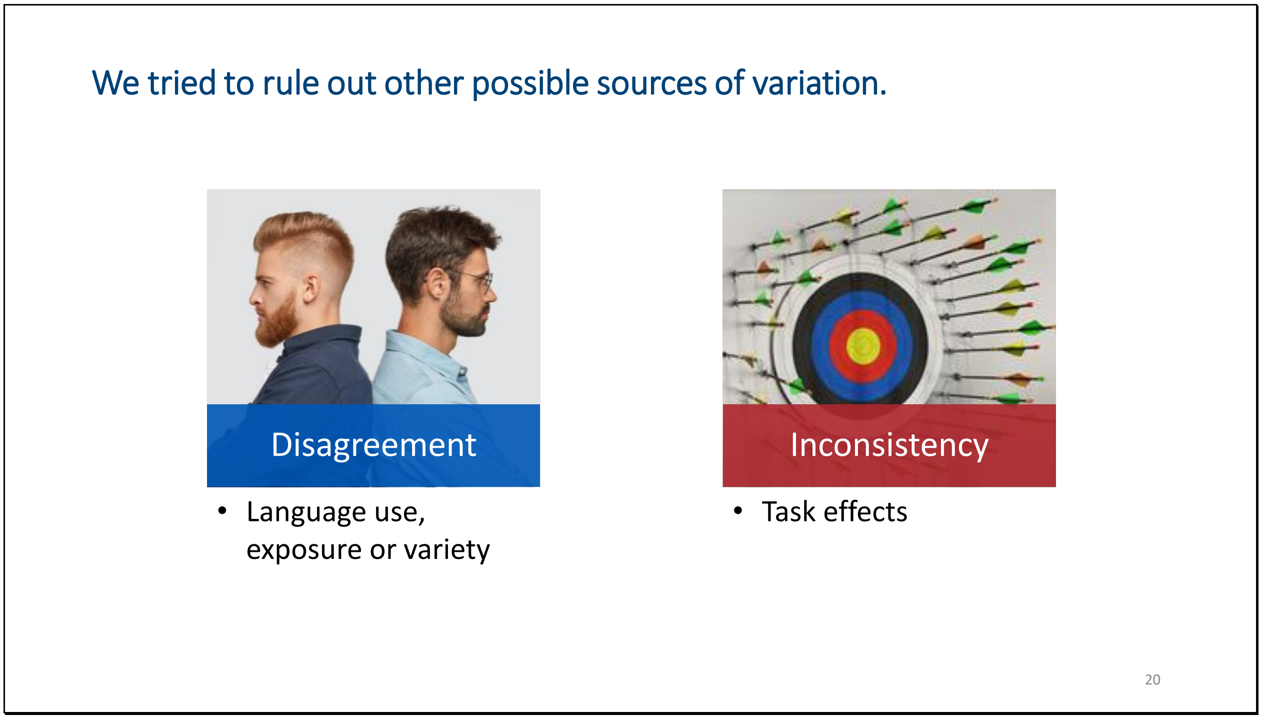
Prediction: Uniform, strong decrease in acceptability

Prediction: Individual variation alongside other cognitive factors

19

We observe substantial inter- and intra-individual variation, contrary to the predictions of grammatical accounts.

However, the individual variation we observe is not as expected. First, it does not corelate with individual variation in processing, as predicted by the resource-limitation account. Second, it is not consistent across islands, even between similar islands such as wh-arguments and wh-adjuncts.



We tried to rule out other possible sources of variation.

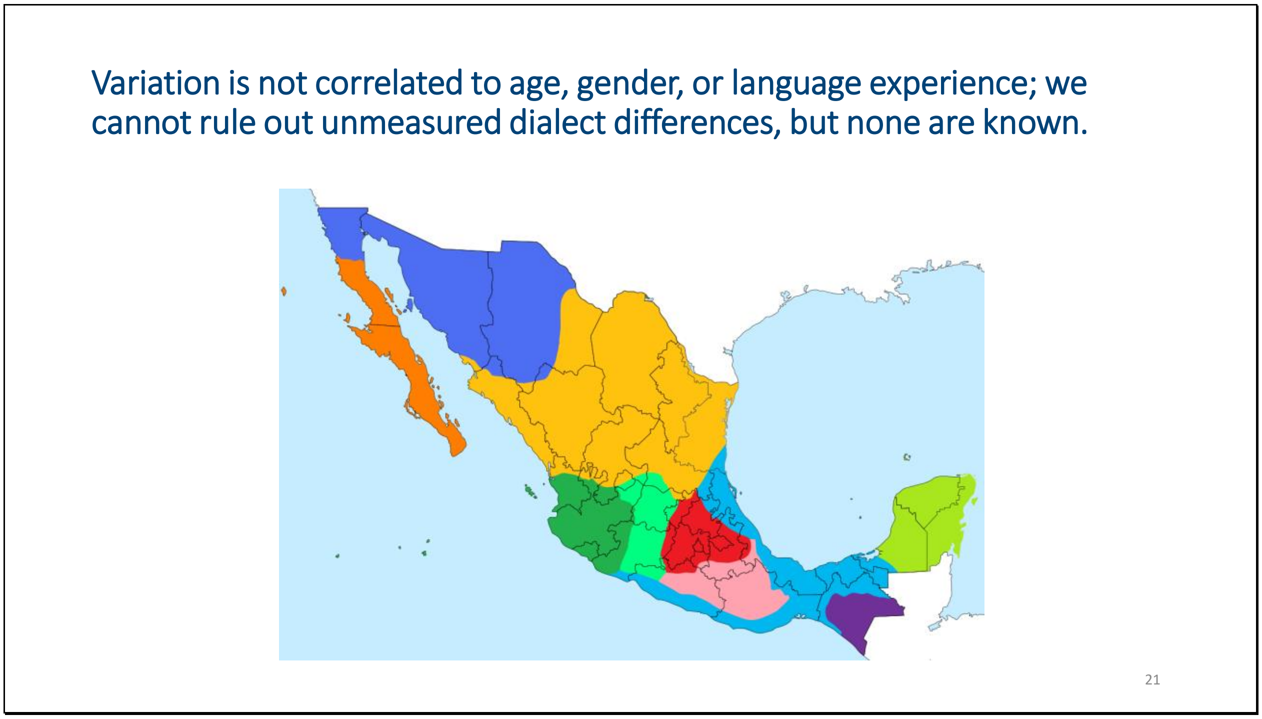
Disagreement

- Language use, exposure or variety

Inconsistency

- Task effects

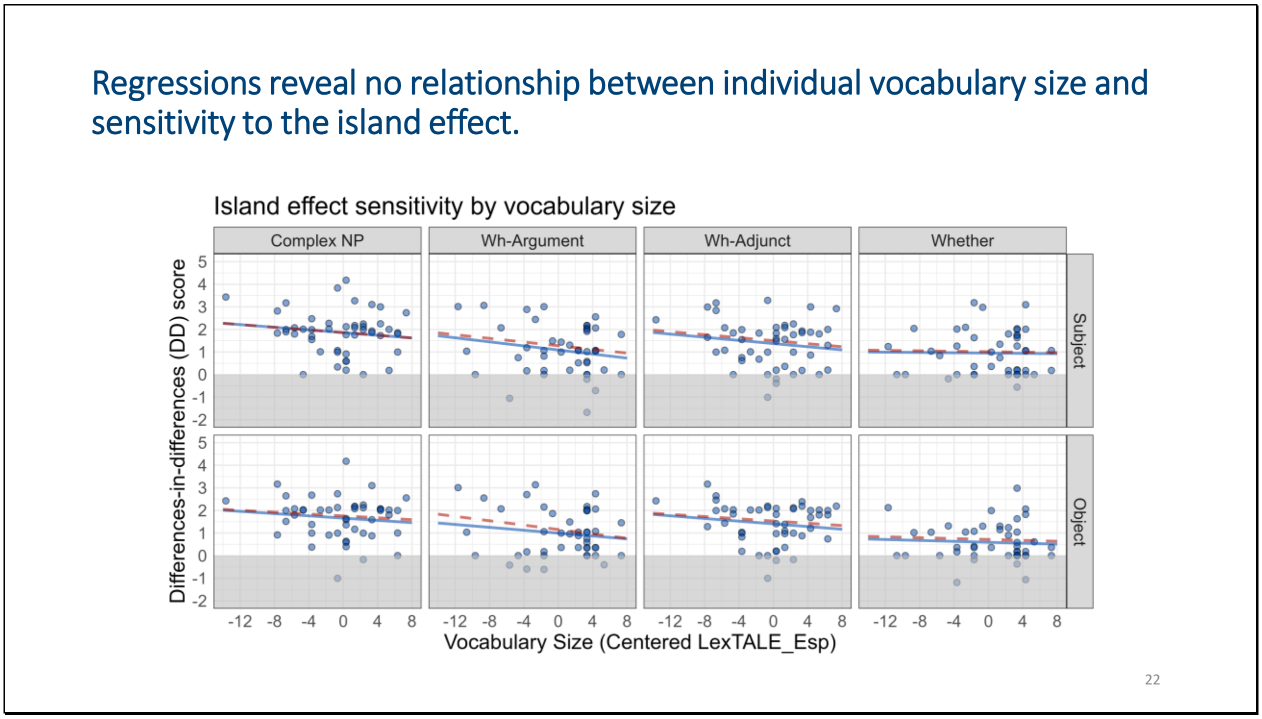
20



Variation is not correlated to age, gender, or language experience; we cannot rule out unmeasured dialect differences, but none are known.

21

Visual examination of both between-participant disagreement and within-participant inconsistency by these extralinguistic factors reveals no patterns that we could see. That said, we collected very limited demographic data, and it is always possible there exist undocumented but subtle dialect differences, although we have no reason to expect this is so, nor that it would apply only to some islands. For instance, Gutiérrez-Bravo (2020) does not mention any such difference between dialects in Mexico.

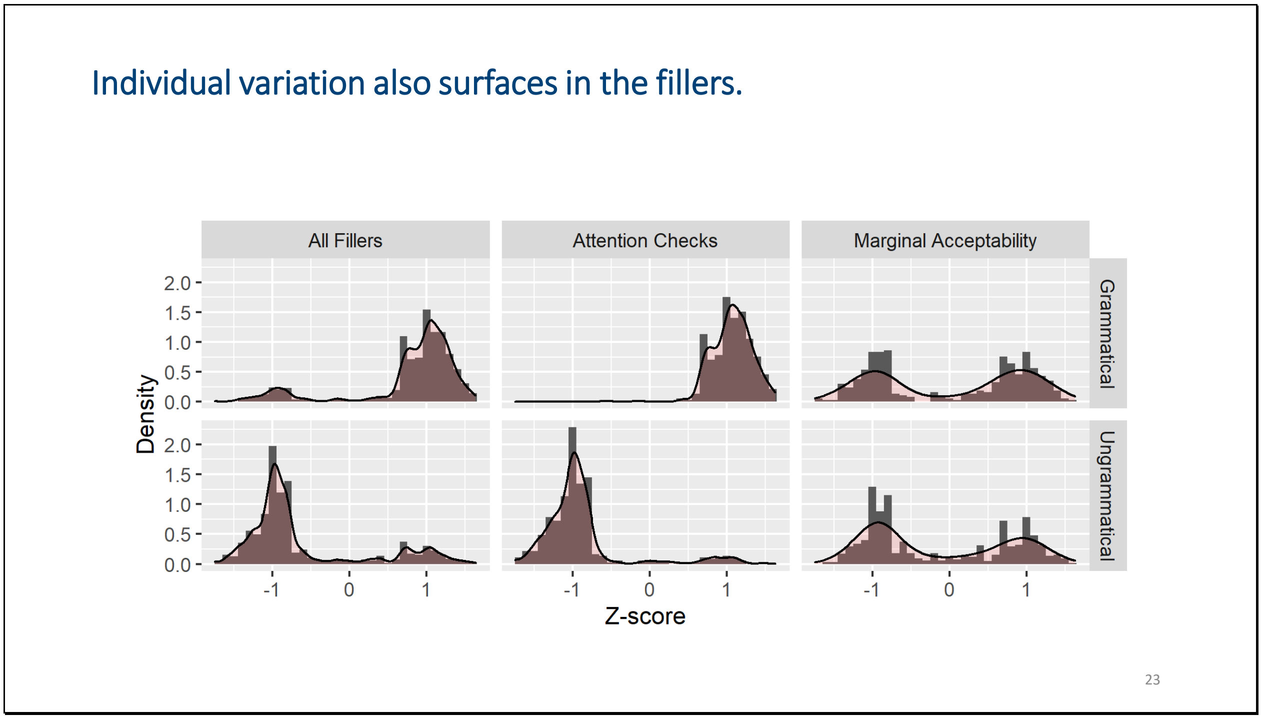


Regressions reveal no relationship between individual vocabulary size and sensitivity to the island effect.

22

We ran a similar regression by vocabulary size (measured by the LexTALE_Esp lexical decision task, Izura et al. 2014), and it showed no effects either. Language experience affects vocabulary size in various ways that might have produced an effect, but we don't observe that here.

So, of the individual cognitive differences we were able to measure, we don't see any evidence that these affect island sensitivity.



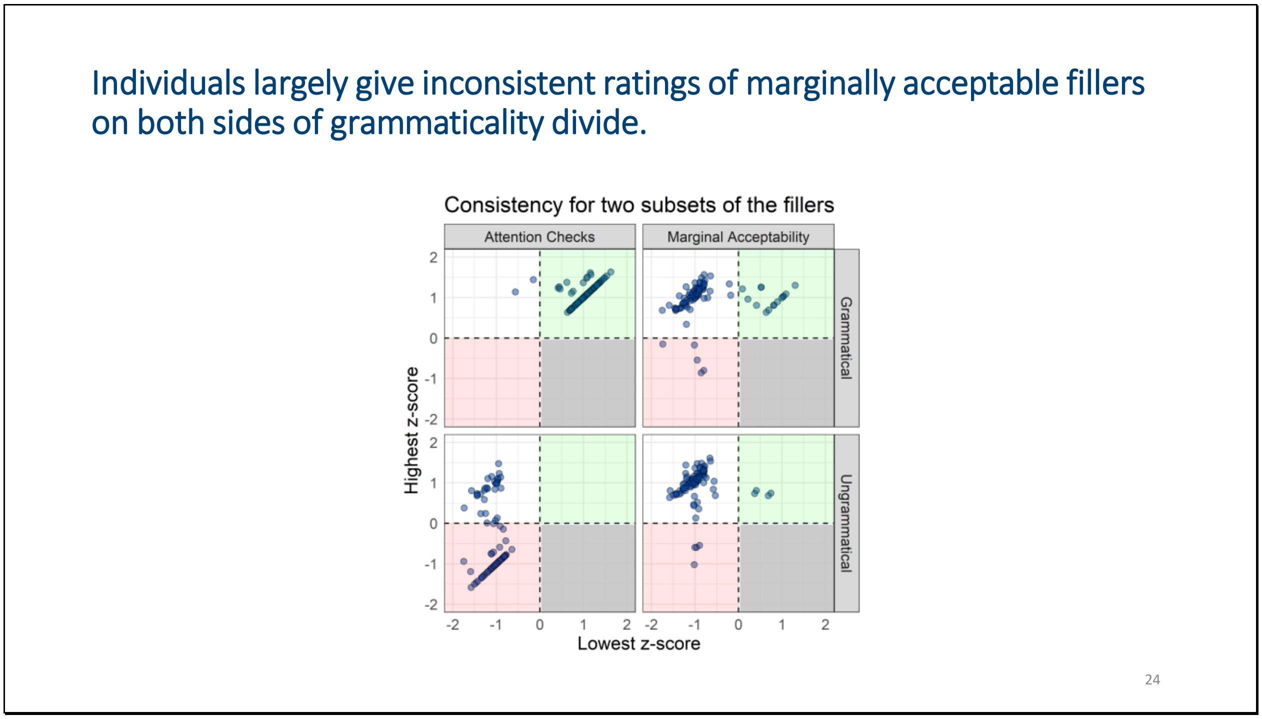
Individual variation also surfaces in the fillers.

23

Returning to the variation in ratings more generally, let's look at within-subject variation in more detail. One way to understand whether participants are largely consistent or not is to look at how they behave when rating the filler items.

Is the variation something about islands, or about their judgments more generally?

Here we plot histograms (with overlaid density plots) of the ratings of the fillers, which reveal consistent ratings overall but significant individual variation on fillers found in previous studies to be of marginal acceptability, whether grammatical or ungrammatical.



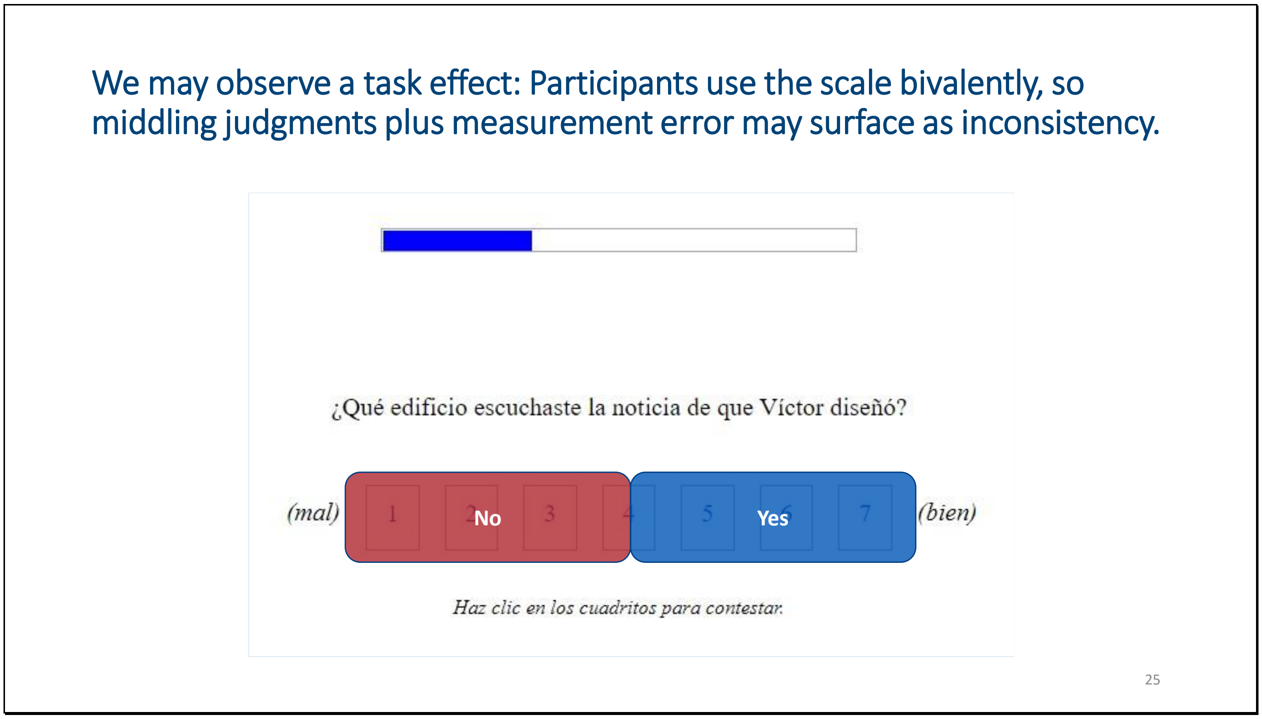
Individuals largely give inconsistent ratings of marginally acceptable fillers on both sides of grammaticality divide.

24

A few people give inconsistent ratings even of the very clear attention check items when they are ungrammatical (and this is after we controlled for people who didn't answer the attention checks consistently).

So, what we see is that sentences that are in the middle of the scale of acceptability, whether grammatical or not, yield the same pattern: inconsistency.

This also suggests that participants may be using the scale bivalently – they are sometimes treating it as a yes/no judgment, such that sentences near the middle of the scale tend to vacillate between yes and no rather than receiving middling ratings.



We may observe a task effect: Participants use the scale bivalently, so middling judgments plus measurement error may surface as inconsistency.

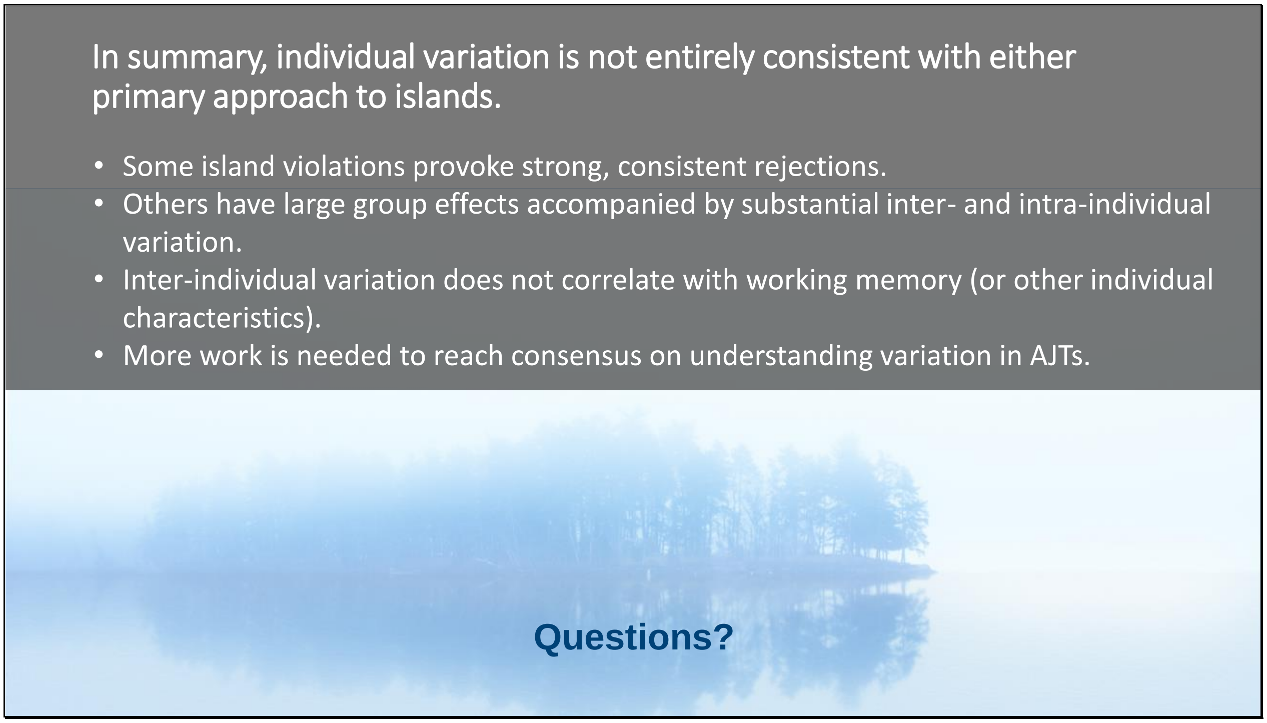
¿Qué edificio escuchaste la noticia de que Víctor diseñó?

Haz clic en los cuadritos para contestar.

25

The role of measurement error in AJTs is poorly understood. As Schütze (2020) points out “we should never underestimate subjects’ creativity in finding ways of looking at sentences that would not have occurred to us, or in being bothered by aspects of sentences that we find mundane” (p. 213), which implies that “the non-convergent data probably do not represent genuine judgment disagreements” (p. 194).

The field has not reached a consensus on how to understand intra-individual variable judgments in AJTs (although see Francis 2022; Schütze & Sprouse 2013 for discussion).



In summary, individual variation is not entirely consistent with either primary approach to islands.

- Some island violations provoke strong, consistent rejections.
- Others have large group effects accompanied by substantial inter- and intra-individual variation.
- Inter-individual variation does not correlate with working memory (or other individual characteristics).
- More work is needed to reach consensus on understanding variation in AJTs.

Questions?

Strong, consistent rejections are consistent with a structural view of islands, whereas substantial variation is not.

Our results are also inconsistent with resource-limitation accounts because individual variation is not correlated with other cognitive or individual factors, however.

Ultimately, both empirical and conceptual work is still needed to reach a consensus in experimental syntax on the treatment of (inter- and intra-individual) variation in AJTs.

References

Boeckx, Cedric. 2012. *Syntactic islands*. Cambridge: Cambridge University Press.

Chomsky, Noam. 1977. On wh-movement. In Peter W. Culicover, Thomas Wasow & Adrian Akmajian (eds.), *Formal syntax*, 71–132. New York: Academic Press.

Chomsky, Noam. 1986. *Barriers*. Cambridge, Mass: MIT Press.

Chomsky, Noam. 2001. Derivation by phase. In Michael J. Kenstowicz (ed.), *Ken Hale: A life in language*, 1–52. Cambridge, Mass.: MIT Press.

Citko, Barbara. 2016. Islands. Oxford: Oxford University Press. doi:10.1093/obo/9780199772810-0101.

Francis, Elaine. 2022. *Gradient acceptability and linguistic theory*. New York: Oxford University Press.

Gutiérrez-Bravo, Rodrigo. 2020. La sintaxis del español de México: Un esbozo. *Cuadernos de la ALFAL* 12(2). 44–70.

Häussler, Jana & Tom S Juzek. 2021. Variation in participants and stimuli in acceptability experiments. In Grant Goodall (ed.), *The Cambridge Handbook of Experimental Syntax*, 97–117. Cambridge University Press. doi:10.1017/9781108569620.005.

Hofmeister, Philip & Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language* 86(2). 366–415. doi:10.1353/lan.0.0223.

Izura, Cristina, Fernando Cuetos & Marc Brysbaert. 2014. Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicologica* 35(1). 49–66.

Juzek, Tom. 2016. Acceptability Judgement Tasks and Grammatical Theory. Oxford: University of Oxford Dissertation.

Kluender, Robert & Marta Kutas. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8(4). Routledge. 573–633. doi:10.1080/01690969308407588.

Kush, Dave, Terje Lohndal & Jon Sprouse. 2019. On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language* 95(3). 393–420. doi:10.1353/lan.2019.0051.

Liu, Yingtong, Elodie Winckel, Anne Abeillé, Barbara Hemforth & Edward Gibson. 2022. Structural, Functional, and Processing Perspectives on Linguistic Island Effects. *Annual Review of Linguistics* 8(1). 495–525. doi:10.1146/annurev-linguistics-011619-030319.

Michel, Daniel. 2014. Individual Cognitive Measures and Working Memory Accounts of Syntactic Island Phenomena. San Diego, Calif.: University of California, San Diego Dissertation.

Pañeda, Claudia & Dave Kush. 2022. Spanish embedded question island effects revisited: an experimental study. *Linguistics* 60(2). 463–504. doi:10.1515/ling-2020-0110.

Pañeda, Claudia, Sol Lago, Elena Vares, João Veríssimo & Claudia Felser. 2020. Island effects in Spanish comprehension. *Glossa* 5(1). 21. doi:10.5334/gjgl.1058.

Pham, Catherine, Lauren Covey, Alison Gabriele, Saad Aldosari & Robert Fiorentino. 2020. Investigating the relationship between individual differences and island sensitivity. *Glossa* 5(1). 94. doi:10.5334/gjgl.1199.

Rizzi, Luigi. 1990. *Relativized minimality*. Cambridge, Mass: MIT Press.

Ross, John Robert. 1967. Constraints on variables in syntax. Massachusetts Institute of Technology Dissertation.

Schütze, Carson T. 2020. Acceptability ratings cannot be taken at face value. In Samuel Schindler, Anna Drożdżowicz & Karen Brøcker (eds.), *Linguistic Intuitions: Evidence and Method*, 189–214. Oxford: Oxford University Press. doi:10.1093/oso/9780198840558.003.0011.

Schütze, Carson T. & Jon Sprouse. 2013. Judgment data. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 27–50. Cambridge: Cambridge University Press.

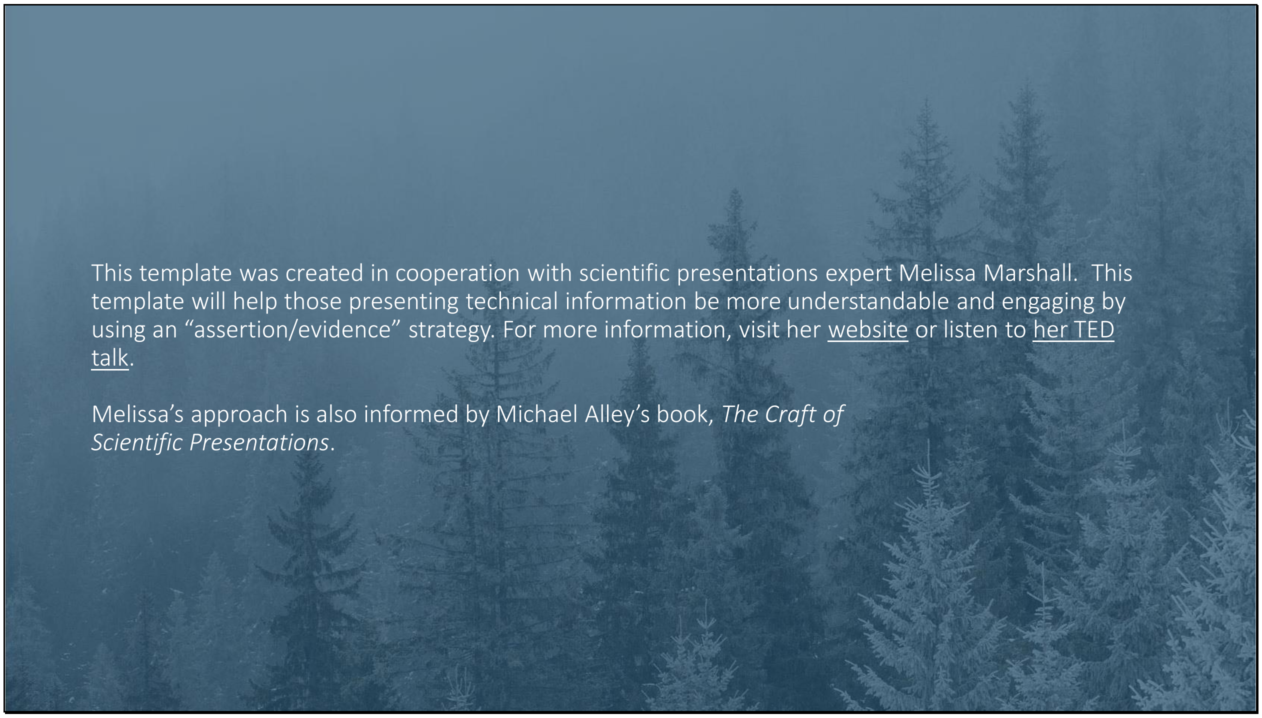
Sprouse, Jon, Ivano Caponigro, Ciro Greco & Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory* 34(1). 307–344. doi:10.1007/s11049-015-9286-8.

Sprouse, Jon, Matt Wagers & Colin Phillips. 2012. A test of the relation between working-memory capacity and syntactic island effects. *Language* 88(1). 82–123. doi:10.1353/lan.2012.0004.

Stigliano, Laura & Ming Xiang. 2021. Experimental Evidence on Island Effects in Spanish Relative Clauses. *Probus* 33(2). 271–296. doi:10.1515/prbs-2021-0008.

Szabolcsi, Anna & Terje Lohndal. 2017. Strong vs. Weak Islands. In Martin Everaert & Henk C. van Riemsdijk (eds.), *The Wiley Blackwell Companion to Syntax, Second Edition*, 1–51. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9781118358733.wbsyncom008.

Torrego, Esther. 1984. On Inversion in Spanish and Some of Its Effects. Linguistic Inquiry 15(1). 103–129. doi:10.2307/4178369.



This template was created in cooperation with scientific presentations expert Melissa Marshall. This template will help those presenting technical information be more understandable and engaging by using an “assertion/evidence” strategy. For more information, visit her website or listen to her TED talk.

Melissa’s approach is also informed by Michael Alley’s book, *The Craft of Scientific Presentations*.

Download a copy of this handout at https://tinyurl.com/Locality23.