A translation-matched, experimental comparison of three types of *wh*-island effects in Spanish and English

Claudia Pañeda[a,b], Sandra Villata[c], Dave Kush[d] and Jon Sprouse[e]

[a]Universidad de Oviedo (panedaclaudia@uniovi.es), [b]Universitat Oberta de Catalunya, [c]Università degli Studi di Enna "Kore" (sandra.villata@unikore.it), [d]University of Toronto (dave.kush@utoronto.ca), [e]New York University Abu Dhabi (jon.sprouse@nyu.edu)

Abstract. According to the historical empirical consensus in the field, *wh*-argument extraction from embedded *wh*-questions gives rise to island effects in English, but not in Spanish. This observation – which was important for the development of a parameters-based theory of cross-linguistic variation in islands – has recently been challenged by experimental studies showing *wh*-island effects in both languages. However, these studies typically employ different materials and experimental conditions between languages, limiting direct comparison. Our study addresses this limitation by testing *wh*-islands in both English and Spanish with translation-matched materials. We present twelve acceptability judgment experiments with approximately 100 participants per experiment. In each language, we examine *wh*-argument extraction from three *wh*-clause types (introduced by *whether*, *why* and *when*) under two matrix verb types (*know* and *ask*), amounting to six *wh*-islands that are relevant to assess the reported contrasts. We test (i) for the presence or absence of *wh*-island effects in the two languages, (ii) for a gradient contrast in effect size, and (iii) for evidence of increased individual variation in Spanish as compared to English. We find (i) that *wh*-island effects are present in both English and Spanish, (ii) that they are rather large in both languages and larger in Spanish for most *wh*-island types, and (iii) that Spanish does not show more individual variation in *wh*-island effects than English. Our results speak against the cross-linguistic contrast as originally proposed, suggesting that its use as evidence for theories that encode cross-linguistic variation in *wh*-island effects might need to be reconsidered.

Keywords: *wh*-islands; cross-linguistic variation; Spanish; English; acceptability judgments

1 Introduction

In a seminal paper on subject-verb inversion in Spanish, Torrego (1984) observed that Spanish and English differ with respect to *wh*-islands: extraction from embedded *wh*-questions in English gave rise to island effects, whereas extraction from embedded *wh*-questions in Spanish appeared to be possible, at least for certain types of *wh*-questions and matrix verbs (a complexity we review in Section 2). This observation was offered as evidence toward a theory of subject-verb inversion and successive cyclic movement, but it also became a critical example of cross-linguistic variation in island effects, and contributed to the development of the influential parameterized Subjacency theory, in which languages differ with respect to which phrasal categories act as bounding nodes (see also Rizzi 1982). Interestingly, recent experiment work has raised questions about the extent to which the contrast holds, as it has found *wh*-island effects not only in English, but also in Spanish. However, direct comparison is limited because these studies typically employ different materials and experimental conditions between languages. In this context, our empirical goal is to solidly establish the facts of cross-linguistic variation in *wh*-islands between Spanish and English, and our theoretical goal is to interpret those facts relative to five leading theories of island effects. To that end, we report twelve acceptability judgment experiments (six in each language) that used translation-matched materials to examine *wh*-extraction (i.e., extraction to form *wh*-questions) from three *wh*-clause types under two matrix verb types, amounting to six *wh*-islands that are relevant to the extraction patterns reported in Torrego (1984). These experiments were designed to (i) test for the presence or absence of *wh*-island effects in the two languages, (ii) test for a possible gradient contrast in effect size, in case the difference that Torrego observed is one of size rather than presence/absence, and (iii) test for evidence of increased individual variation in Spanish as compared to English (using both between-participant and within-participant measures), which might explain some of the variability in the observations reported in the literature. We address these questions with data from approximately 100 participants per experiment (approximately 1200 total), to ensure enough statistical power to detect small differences in effect sizes, and four tokens per condition per participant, to be able to examine within-participant variation.

Anticipating slightly, our results suggest (i) that *wh*-island effects are present for *wh*-extraction in both English and Spanish, (ii) that the effects are fairly large in both languages, and perhaps unexpectedly, larger in Spanish for most *wh*-island types; and (iii) that Spanish does not show increased individual variation based on either between-participant or within-participant measures. This establishes, unequivocally, that both Spanish and English have *wh*-island effects, at least for these three embedded *wh*-clause types. Because this is the opposite of the facts that have been assumed in the islands literature, this has consequences for all types of theories of island effects, including the five that we review here.

The rest of this article is organized as follows. Section 2 discusses Torrego's (1984) observation in more detail, including the structural (and potentially lexical) conditions that are reported to be necessary to extract a *wh*-word from a *wh*-island in (European) Spanish. Section 3 provides a brief introduction to five major theories of island effects and how they capture the reported cross-linguistic variation between Spanish and English: the Subjacency approach (Chomsky 1973) originally adopted by Torrego 1984, the Phase Impenetrability approach (Chomsky 2000; 2001), Relativized Minimality (Rizzi 1990; 2004), Information Structure approaches (Erteschik-Shir 1973; Goldberg 2006; Goldberg 2013; Abeillé et al. 2020), and processing resource-limitation approaches (Kluender & Kutas 1993; Hofmeister and Sag 2010). Section 4 reviews recent experimental work in both Spanish and English that has raised the possibility that extraction from *wh*-islands is not available for all speakers of Spanish. Section 5 describes the logic and design of our twelve experiments in detail. Section 6 presents the results with analyses for each of our three driving questions. Section 7 discusses the consequences of our results for the five theories of island effects, with a particular focus on cross-linguistic variation. Section 8 presents a brief conclusion.

2 The cross-linguistic contrast as described by Torrego (1984)

Wh-island effects are the unacceptability that arises when a *wh*-phrase is A′ moved from an embedded *wh*-clause. While counterexamples have been reported (e.g., Ross 1967), the historical consensus in the field is that there are *wh*-island effects for *wh*-extraction in English (1), but not in Spanish (2), (3).

(1) *What did he wonder where John put?

(Chomsky 1964: 47)

(2) ¿Qué dices que no te explicas por qué
what say.2SG.PRS that NEG 2SG.DAT explain.2SG.PRS why

Juan se habrá comprado?
Juan 3SG.DAT have.3SG.FUT buy.PTCP

'What do you say that you can't figure out why Juan may have bought for himself?'

(Torrego 1984: 115)

(3) ¿Qué diccionario no sabías si Celia había
what dictionary NEG know.2SG.PST whether Celia have.3SG.PST

devuelto ya?
return.PTCP yet

'Which dictionary didn't you know whether Celia had returned yet?'

(Torrego 1984: 115)

The contrast above is based on Torrego's (1984) seminal article on subject-verb inversion, but, in fact, her observations were more nuanced: She considered extraction from *wh*-islands to be readily available from embedded subject positions, but more constrained from embedded object positions, which was only deemed possible when the *wh*-island was introduced by a non-argument, like *si* ('whether/if'), *por qué* ('why'), or *cuándo* ('when').[1] Given that there is an independent comp-trace effect for extraction from embedded subject positions in English (Perlmutter 1968; Bresnan 1977), a well-

[1] In Torrego's analysis, extraction from these *wh*-islands is ultimately possible to the extent that they do not require subject-verb inversion. While she makes the generalization that inversion is not required in *wh*-islands introduced by non-arguments, she also points out that there could be variation depending on the non-argument. However, we set aside questions of the theory of subject-verb inversion (the focus of Torrego's paper) to focus solely on the empirical question of variation in *wh*-island effects.

controlled cross-linguistic contrast is only expected in sentences with extraction from *wh*-islands introduced by non-arguments.

The next section explains how the cross-linguistic contrast could be accounted for by five major theories of island effects.

3 Theories of island effects and cross-linguistic variation of *wh*-islands

In this section we provide a brief introduction to five major theories of island effects, focusing on how they can capture Torrego's (1984) observations about the cross-linguistic variation in *wh*-islands between English and Spanish. We discuss the Subjacency approach (Chomsky 1973) originally adopted by Torrego (1984), its modern descendant Phase Impenetrability (Chomsky 2000; 2001), Relativized Minimality (Rizzi 1990; 2004), Information Structure approaches (Erteschik-Shir 1973; Goldberg 2006; 2013; Abeillé et al. 2020), and processing resource-limitation approaches (Kluender and Kutas 1993; Hofmeister and Sag 2010). We focus on these five theories because each of them has been refined over decades of work with the goal of capturing all of the established facts about island effects. Given that our study is an investigation of one set of those facts, our results will have consequences for all of the existing theories of island effects – that is, each theory will need to be modified to capture any new facts that we discover. In other words, the goal of our study is exploratory (collecting data to construct or revise theories) rather than confirmatory (collecting data to select one theory over others). Our description of these theories will serve as a background to the discussion of the modifications motivated by our new results in Section 6.

The original Torrego (1984) study was situated within the Subjacency approach to island effects, which bans movement steps that cross more than one bounding node (Chomsky 1973; 1977). It was proposed that in English Inflectional Phrase (IP) is a bounding node, which causes extraction from *wh*-islands to violate Subjacency (Chomsky 1977): the reason is that there are two IPs between the base position of the extractee and its landing position, and they cannot be crossed in separate movement steps given that the intermediate Spec, Complementizer Phrase (CP) position that could provide an "escape hatch" is filled with the *wh*-word. The Subjacency approach to island effects was embedded within the Principles and Parameters approach to cross-

linguistic variation (Haegeman 1994; van Riemsdijk & Williams 1986). Rizzi (1982) and Torrego (1984) argued that the absence of *wh*-island effects in Italian and Spanish, respectively, could be captured if the choice of bounding node were parameterized, such that in these languages CP was a bounding node rather than IP. This parameterization would enable extraction from *wh*-islands to comply with Subjacency in Italian and Spanish, as there is only one CP node to cross between the base position of the extractee and the landing position.

Within the minimalist program (Chomsky 1995, et seq.), one prominent syntactic approach to island effects is the Phase Impenetrability approach (Chomsky 2000; 2001, and elaborated by many others). The critical idea is that there are special syntactic domains, called phases, that limit the application of syntactic operations, such that any given syntactic operation can only target two items if they are within the same phase or if one is within a phase and the other is within the "edge" of the next more deeply embedded phase (where "edge" is typically defined as the specifier or the head of a phase). It is easy to see how phases can give rise to something like *wh*-island effects – if the embedded question is a phase, and if the item that introduces the question is sitting in the specifier at the edge of the phase, then the extractee will not be able to move to the edge of the phase, and will not become available for movement to the matrix clause. Like all minimalist analyses, phases have a strong universalist component to them – phases derive from constraints on (syntactic) computational efficiency that limit operations to local domains (see Boeckx 2012; Citko 2014, and Müller 2021 for a review). This means that the definition of phase should not vary among languages, and instead any cross-linguistic variation in a phase impenetrability approach to *wh*-islands must be captured through differences in the lexical items that make up the *wh*-question (i.e., the Borer Conjecture; Borer 1984). One possibility would be for the C head in embedded questions in Spanish to license a second specifier position, as in the second COMP position proposed by Reinhart (1981) or the cP layer proposed by Nyvad et al. (2017). Another possibility would be for the *wh*-items that introduce embedded questions in Spanish to sit in a different structural position than

the *wh*-items in English, and crucially for that position to not be the phase edge (e.g., the cartographic approach to the left periphery in Rizzi 1997).[2]

A third prominent syntactic account of *wh*-island effects that is compatible with the assumptions of the minimalist program is Relativized Minimality (RM) (Rizzi 1990; 2004). RM holds that an extractee can only establish a dependency with its base position if there is no intervener that could engage in the same dependency, where intervener is defined as a constituent that shares relevant morphosyntactic features with the extractee (such as a *wh*-feature), c-commands the base position from the same type of position as the landing position (such as spec, CP), but crucially does not c-command the landing position (see also Friedmann et al. 2009; Belletti et al. 2012; Rizzi 2013; Atkinson et al. 2016; Villata et al. 2016). The *wh*-word introducing *wh*-islands meets the requirements of an intervener, causing extraction from *wh*-clauses to violate RM. In this framework, cross-linguistic variation in *wh*-island effects could be accounted for (i) if the intervener shared a movement feature with the extractee in one language and not in the other, or (ii) if the structural position of the intervener was the same type as the landing position in one language but not the other.

A fourth prominent approach to island effects is the Information Structure-based approach of Erteschik-Shir (1973), which has been more recently advocated by Goldberg (2006; 2013), Abeillé et al. (2020), and Cuneo & Goldberg (2023) (among others). Erteschik-Shir proposed that island effects arise when there is a clash between the information structure properties of the extractee (i.e., whether it is focused or backgrounded) and the information structure properties of the clause that contains the extractee. For example, given that *wh*-question formation is a focus operation, island effects should arise whenever the *wh*-item is extracted from a backgrounded clause. In her seminal dissertation, Erteschik-Shir (1973) proposes a number of tests to determine if a clause is backgrounded or focused, and reports a compelling alignment between the backgroundedness of clauses and island structures. This alignment has received some

[2] This is not an exhaustive list of the approaches to variation in phase impenetrability. But other approaches, like Rackowski & Richards' (2005) agreement-based approach and Müller's (2010) last-merged specifier approach, were not constructed to account for variation in *wh*-islands, and it is not clear that they could be extended to do so.

experimental support from studies like Cuneo & Goldberg (2023). Within information-structure-based theories, variation in the pattern of island effects across languages reduces to a question of the variation in the information structure properties of the clauses (and dependency types) in different languages. Erteschik-Shir in fact analyzes an example of variation in *wh*-islands between English and Danish in her dissertation, arguing that speakers of Danish who do not report *wh*-island effects are more able to treat embedded *wh*-questions as focused. Along the same lines, the cross-linguistic contrast between English and Spanish could be accounted for if Spanish speakers were more prone to treating embedded *wh*-questions as focused.[3]

A final approach is to view *wh*-island effects as resulting from processing difficulty. Under a maintenance theory of working memory, this difficulty may arise because processing the dependency and the embedded island structure exceeds the available working memory capacity, causing working memory overload (Kluender & Kutas 1993; Hofmeister & Sag 2010). Alternatively, under a cue-based theory of working memory, the difficulty may arise because the intervener is similar to the extractee in its featural composition and therefore interferes with the resolution of the dependency. There are two ways in which the intervener may interfere: (i) by hindering the encoding of the extractee in working memory when it is encountered at the landing position during left-to-right processing, or (ii) by hindering the retrieval of the extractee from working memory to resolve the dependency at the base position (Atkinson et al.

[3] We note for completeness that there are also semantic approaches to *wh*-island effects, which derive the unacceptability of the island effect from a semantic incompatibility between the *wh*-question operation and the semantics of the *wh*-operator in the embedded question (e.g., Szabolsci & Zwarts 1993; Abrusán 2014). We do not explore these here because they appear to be focused on the distinction between extractees that range over *individuals*, i.e., *wh*-arguments, and extractees that range over ordered sets (e.g., properties) or are non-referential, i.e., *wh*-adjuncts. We only test *wh*-arguments in this study. (There is also an empirical incompatibility in that these theories predict that extraction of *wh*-arguments should not give rise to *wh*-island effects, contrary to the findings in the experimental literature.)

2016; Villata et al. 2016; Keshev & Meltzer-Asscher 2019). From this perspective, cross-linguistic variation in *wh*-island effects could arise if *wh*-island sentences are easier to process in one language than the other. For example, it has been proposed that Spanish could show no or smaller *wh*-island effects than English because it has a richer morphology that provides more informative encoding and/or retrieval cues, facilitating processing (Ortega-Santos 2011).

We have discussed how Subjacency, Phase Impenetrability, Relativized Minimality, Information-based structure approaches and processing resource-limitation approaches can account for cross-linguistic contrasts in island effects like the one between English and Spanish originally observed by Torrego (1984). However, recent experimental work has found island effects in both English and Spanish, suggesting that the contrast may not hold for all speakers. This work and its limitations, which motivate our study, are reviewed in the next section.

4 Prior experimental work

Several experimental studies have assessed the presence of *wh*-island effects in English and Spanish using the 2×2 factorial design that has become standard in the island effects literature (4).

(4)

nonisland/matrix	Who __ thinks that John bought a car?
nonisland/embedded	What do you think that John bought __?
island/matrix	Who __ wonders whether John bought a car?
island/embedded	What do you wonder whether John bought __?

(Sprouse et al. 2016: 318)

Under this design, island effects are quantified as an interaction between two factors: STRUCTURE, which manipulates the structure of the embedded clause between a declarative (non-island) and a question (island), and POSITION, which manipulates whether the gap is in the matrix or embedded clause (see Sprouse 2007 and subsequent work). This design controls for the independent effects on acceptability of these two factors, as well as any other properties of the sentences that are distributed across the two levels of the factors. It isolates the island effect in the interaction term, driven by an unexpected low rating in the island/embedded condition (unexpected because it is lower than predicted by the linear sum of the independent effects of the two factors).

Thus, island effects are identified statistically as a 'superadditive' Position × Structure interaction. Island effects can be identified visually in a plot of the means of the four conditions as non-parallel lines arranged such that the island/embedded condition is lower than the other three conditions, as in the left and center panels of Figure 1. If the island/embedded condition has much lower acceptability than the other conditions, as in the left panel, this is an indication that there is a large island effect; if it only has slightly lower acceptability, as in the center panel, there is a small island effect. The right panel shows that the absence of island effects can be identified visually as parallel lines.

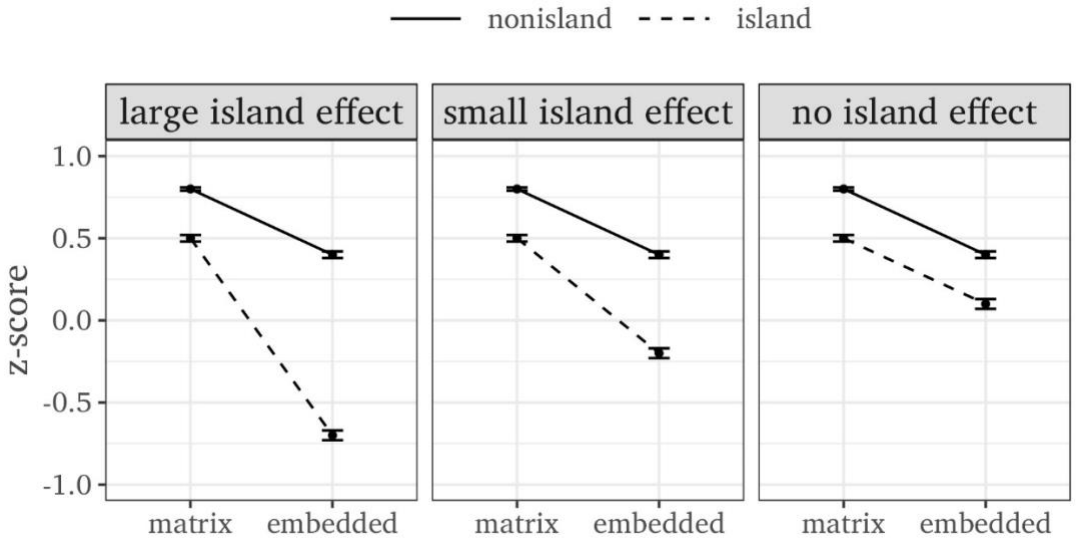

Figure 1: Example of how large, small and no island effects appear in plots of the means of the four conditions in the 2×2 design.

The size of the island effect is captured in the interaction term and can be calculated from the condition means with a differences-in-differences (DD) score, as in (5).

(5) DD score: (non-island/embedded – island/embedded) – (non-island/matrix – island/matrix)

Studies using the design in (4) in acceptability judgment tasks have repeatedly found *wh*-island effects in English, in line with the received view about this language. A summary of this prior work is shown in Table 1.

	Island	DD	Notes
***Wh*-questions with bare extractees**			
Almeida 2014	*whether*	~ 0.9	
Michel 2014	various verbs + *whether*	0.2	animate extractee
Sprouse 2007	*wonder whether*	0.18	ME
Sprouse 2007	*wonder whether*	0.15	ME, context
Sprouse 2007	*wonder wh-*	0.38	ME
Sprouse 2007	*wonder wh-*	0.33	ME, context
Sprouse et al. 2011, exp. 1	*wonder whether*	~ 0.65	ME
Sprouse et al. 2012, exp. 1	*wonder whether*	0.73	
Sprouse et al. 2012, exp. 2	*wonder whether*	0.59	ME
Sprouse et al. 2016, exp. 1	*wonder whether*	1.15	
Ortega-Santos et al. 2018	*know why*	1.36	context, subject (and animate) extractee
***Wh*-questions with complex extractees**			
Aldosari 2015	*wonder whether*	0.35	context
Sprouse et al. 2016	*wonder whether*	0.62	
Pham et al. 2020	*wonder whether*	NA	context
Relative clauses			
Sprouse et al. 2016	*wonder when / how / where / why*	0.4	
Sprouse et al. 2016	*wonder when / how / where / why*	0.35	
Left dislocation			
Almeida 2014	*wonder whether*	~ 0.2	

Table 1: Summary of prior studies testing *wh*-islands in English in *wh*-questions with bare and complex extractees, relative clauses and left dislocation configurations. We show which *wh*-islands were tested under which verb and the differences-in-differences (DD) score that indicates the island effect size (the sign ~ is used when neither the DD score nor the conditions means were reported in the study and had to be roughly estimated from the plots; NA is used when the DD score could not be estimated due to the absence of z-scores). All studies tested island effects with a 7-point scale task, no context and inanimate object extractees in the embedded conditions, unless otherwise

indicated in the "Notes" column (ME indicates that a magnitude estimation task was used).

As Table 1 shows, in most cases, *wh*-island effects have been investigated and observed in cases like (4), where a bare *wh*-word (e.g., *what*) is extracted from a *whether* clause to create a *wh*-question (Sprouse 2007; Almeida 2014; Michel 2014; Aldosari 2015; Sprouse et al. 2011; 2016; Pham et al. 2020). However, island effects have also been found when extracting complex *wh*-words (e.g., *which book*; Aldosari 2015; Sprouse et al. 2016; Pham et al. 2020), when extracting from other *wh*-clauses (Sprouse 2007; Sprouse et al. 2016; Ortega-Santos et al. 2018) and when creating relative clause dependencies (Sprouse et al. 2016). DD scores vary substantially across studies, ranging from 0.15 in Sprouse's (2007) *wonder whether* islands with object extractees to 1.36 in Ortega-Santos et al.'s (2018) *know why* islands with subject extractees (both involved *wh*-question configurations, bare extractees and a preceding context). We are aware of a single study where no *wh*-island effect (i.e., no statistical interaction) was observed, when island effects were tested in left dislocation configurations (Almeida 2014).

Interestingly, the design in (4) has also revealed the presence of *wh*-island effects in Spanish, in cases where no such effects are expected based on Torrego (1984), i.e., in cases of extraction from *wh*-clauses introduced by non-arguments like *si* ('whether'), *cuándo* ('when') or *por qué* ('why') (see (6) for an implementation of the design in Spanish and Table 2 for a summary of the Spanish studies).

(6) a. non-island/matrix

¿Quién	__	piensa	que	Rocío	vio	el	mensaje?
who	__	think.3SG.PRS	that	Rocío	see.3PL.PST	the.M	message

Who __ thought that Rocío saw the message?

b. non-island/embedded

¿Qué	piensas	que	vio	__	Rocío?
what	think.2SG.PRS	that	see.3PL.PST	__	Rocío

What do you think that Rocío saw __?

c. island/matrix

¿Quién __ se pregunta si Rocío
who __ REFL ask.3SG.PRS whether Rocío
vio el mensaje?
see.3PL.PST the.M message
Who __ wonders whether Rocío saw the message?

d. island/embedded

¿Qué te preguntas si Rocío vio __?
what REFL ask.2SG.PRS whether Rocío see.3PL.PST __
What do you wonder whether Rocío saw __?

(López-Sancio 2015: 10)

	Island	DD	Notes
***Wh*-questions with bare extractees**			
López-Sancio 2015	*preguntarse si* 'wonder whether'	~ 1.75	
Ortega-Santos et al. 2018	*saber por qué* 'know why'	1.15	
Pañeda et al. 2020	*preguntar si* 'ask whether'	NA	speeded task with binary judgments
Rodríguez & Goodall 2020	*preguntarse / necesitar saber / querer saber…* 'wonder / need to know / want to know…'		
	…si 'whether'	1.06	
		1.00	subject extractee
	…cuándo / dónde 'when / where'	0.87	
		0.71	subject extractee
***Wh*-questions with complex extractees**			
Pañeda & Kush 2022	*saber si* 'know whether'	0.22	context
	preguntar si 'ask whether'	0.38	context
	saber cuándo 'know when'	1.09	context
	preguntar cuándo 'ask when'	1.39	context
Relative clauses			
López-Sancio 2015	*preguntarse cómo / cuándo / por qué* 'wonder how / when / why'	~ 1.25	
Stigliano & Xiang 2021	*preguntar quién* 'ask who'	0.95	

Table 2: Summary of prior studies testing *wh*-islands in Spanish in *wh*-questions with bare and complex extractees and relative clauses. We show which *wh*-islands were tested under which verb and the differences-in-differences (DD) score that indicates the island effect size (the sign ~ is used when neither the DD score nor the condition means were reported in the study and had to be roughly estimated from the plots; NA is used when the DD score could not be estimated due to the absence of z-scores). All studies tested island effects with a 7-point scale task, no context and inanimate object extractees in the embedded conditions, unless otherwise indicated in the "Notes" column.

Just like in English, most of the Spanish studies have tested and identified *wh*-island effects in *wh*-question configurations and with bare extractees (López-Sancio

2015; Ortega-Santos et al. 2018; Pañeda et al. 2020; Rodríguez & Goodall 2020), but *wh*-island effects have also been observed in relative clause dependencies (López-Sancio 2015; Stigliano & Xiang 2021) and with complex extractees (Pañeda & Kush 2022). Some of these island effects could potentially be attributed to the presence of the embedding verb *preguntar*(*se*) ('to ask/wonder') (López-Sancio 2015; Pañeda et al. 2020; Stigliano & Xiang 2021), which independently prevents extraction according to Torrego (see also Suñer 1991). But *wh*-island effects have also been observed with *saber* ('to know'), which is not claimed to pose any such constraint (Ortega-Santos et al. 2018; Pañeda & Kush 2022). DD scores also vary greatly, ranging between 0.22 and 1.39 (obtained, respectively, in Pañeda & Kush's 2022 *know whether* and *ask when* islands, both tested in *wh*-configurations, with complex extractees and a preceding context).

Thus, previous experimental results suggest that currently spoken Spanish manifests *wh*-island effects even in the cases in which it was predicted not to do so, and that English and Spanish are more similar than previously thought, in that they both generally show *wh*-island effects, with similar (and significant) amounts of variation in island effect sizes. This similarity casts doubt on the cross-linguistic contrast often inferred from Torrego's (1984) observations.

Nonetheless, the conclusions that can be reached about the cross-linguistic contrast based on the experimental studies above are limited, for several reasons. First, those studies have often tested extraction out of a single type of *wh*-island, and it is unclear whether their results generalize to other cases. Notably, in most cases, island effects have been tested in *whether* islands under *ask*/*wonder*. However, *whether* may not behave in the same way as other *wh*-islands, due to a different featural composition or because it does not sit in the same projection within the Complementizer Phrase (Hernanz Carbó 2012; Pañeda & Kush 2022; Rizzi 2001). Similarly, *ask/wonder* may constrain extraction more than other verbs (Torrego 1984; Suñer 1991; Pañeda & Kush 2022). Second, previous studies have mostly tested the two languages separately, with different materials and, sometimes, under different conditions that are not comparable. For example, Pañeda & Kush's (2022) Spanish results cannot be compared to Sprouse et al.'s (2012) English results because the former were obtained with complex (or "d(iscourse)-linked") extractees, which have been claimed to independently reduce or

eliminate *wh*-island effects (Pesetsky 1987; see Szabolcsi & Lohndal 2017 for a review), while the latter were obtained with bare extractees. We are aware of one study that tested *wh*-island effects in both languages (Ortega-Santos et al. 2018), but this study only assessed *know why* islands and extraction was from the embedded subject position, which, as we indicated above, yields a comp-trace effect in English, making cross-linguistic comparisons difficult.

To better assess the cross-linguistic contrast, the current study tests *wh*-island effects in both languages in a wider range of syntactic configurations. To rule out items as a source of variation, we use the same translation-matched lexicalizations (see Ortega-Santos et al. 2018 and Abeillé et al. 2020 for a similar approach to cross-linguistic comparisons). Because Spanish and English are predicted to differ with regard to extraction from *wh*-clauses introduced by non-arguments, we test three such clauses —*whether*, *why* and *when* clauses. We chose to test all three to follow up on previous work, which observed differences between *whether* and *when* islands in Spanish (Pañeda & Kush 2022) and compared *wh*-island effects in the two languages with *why* clauses (Ortega-Santos et al. 2018). Because the *know*/*ask* contrast is predicted to affect *wh*-island effects (Torrego 1984; Suñer 1991; Pañeda & Kush 2022), we test both embedding verbs. We focus solely on bare *wh*-extractees as a uniform test case because they are expected to give rise to island effects in English (whereas complex or "d-linked" *wh*-extractees are sometimes claimed to obviate *wh*-islands; see Szabolcsi & Lohndal 2017 for a review). We not only address the binary question of whether island effects are present or absent in each language, but also the gradient question of whether they differ in size across languages: for instance, *wh*-island effects could be smaller in Spanish, supporting a weaker version of the cross-linguistic claim. In addition, given the relative uncertainty in the literature regarding *wh*-islands in Spanish (with contrasting judgments in the theoretical and the experimental literature, and also between and within experimental participants, see Pañeda & Kush 2022), we investigate whether there is increased individual variation in Spanish as compared to English (using both between-participant and within-participant measures), in case that is a potential source of the discrepancy between informal observations and experimental findings.

5 The design of our study

We ran twelve acceptability judgment experiments, six in English and six in Spanish, each examining one of three *wh*-island types (*whether* / *si, why* / *por qué* or *when* / *cuándo*) under one of two embedding verbs (*know* / *saber* or *ask* / *preguntar*). Island effects were examined in sentences with *wh*-extraction (e.g., *What did the politician ask when they would reject?*), in line with most previous studies on both languages (e.g., Sprouse et al. 2012; Ortega-Santos et al. 2018; Pañeda et al. 2020; Pham et al. 2020; see also Table 1 and Table 2). Extractees were bare (e.g., *What* rather than *which cake*), given that complex or "d-linked" fillers may independently reduce *wh*-island effects (Pesetsky 1987). The same item set was used in all experiments, varying only the words of interest. Between languages, translation-matched items were used. Each experiment was 55 items long, consisting of 16 target items (four tokens of each of the four conditions in the 2×2 factorial design for island effects, described in Section 2), 32 filler items distributed equally across the full range of acceptability from previous studies, and 7 burn-in items distributed equally across the range of acceptability that were presented at the beginning in a fixed order to avoid participants having temporary scale bias. The task used a 7-point rating scale. For each experiment, we recruited 112 participants using Prolific (www.prolific.co). In the following sections we discuss the design and methods in more detail.

5.1 *Participants*

We recruited 1344 participants in total through Prolific: 112 for each of the twelve experiments. They were paid $2.75 USD for their participation. To identify native speakers of US English or European Spanish, we used three tools. First, we used Prolific's prescreening tools to identify individuals who considered the target language their first language and had mostly lived in either the US or Spain before turning 18. Second, we asked all participants questions about where they lived from birth until age 13, and about the languages that were spoken in their home as children. Finally, we included two trials within the experiment where we asked participants to read a short description of an ethically challenging situation and write at least one complete sentence in the target language about how they would respond. These open-ended questions were included to identify both bots and potentially uncooperative or non-native participants (Chmielewski & Kucker 2020; Dennis et al. 2020). We excluded

participants from analysis if they reported not living in the target country, not speaking the target language (or speaking it as a non-dominant language) in their home, or if their responses to the morality trials appeared uncooperative or non-native to us, e.g., "mostly likely let my friend known about see he/she" (only six participants across the twelve experiments were excluded based on their responses to the morality trials). We also removed participants from analysis if they responded to four or more fillers with a rating that was more than 2 standard deviations away from the mean rating for that filler. The final sample sizes of each experiment are shown in Table 3.

Structure	English	Spanish
know whether	95	105
ask whether	96	104
know why	97	98
ask why	101	104
know when	99	97
ask when	101	101

Table 3: Number of participants that met the inclusion criteria and were included in the analysis for each experiment.

5.2 *Materials*

The materials for both languages followed the 2×2 factorial design for island effects described in Section 2, where island effects are identified statistically as a 'superadditive' Position × Structure interaction and measured by means of DD scores as in (5). Example (7) illustrates the 2×2 design for Spanish, with the alternative verbs and *wh*-phrases separated by slashes. The English design is illustrated in the translations.

(7) a. non-island/matrix

¿Quién __ pensaba que rechazarían la propuesta?
who __ think.3SG.PST that reject.3PL.COND the.F proposal
Who __ thought that they would reject the proposal?

b. non-island/embedded

¿Qué pensaba el político que rechazarían __?
what think.3SG.PST the.M politician that reject.3PL.COND __
What did the politician think that they would reject __?

c. island/matrix

¿Quién __ preguntó / quería saber
who __ ask.3SG.PST / want.3SG.PST know
si / por qué / cuándo rechazarían la propuesta?
whether / why / when reject.3PL.COND the.F proposal
Who __ asked / wanted to know whether / why / when they would reject the proposal?

d. island/embedded

¿Qué preguntó / quería saber el político
what ask.3SG.PST / want.3SG.PST know the.M politician
si / por qué / cuándo rechazarían __?
whether / why / when reject.3PL.COND __
What did the politician ask / want to know whether / why / when they would reject __?

For each language, we created sixteen item sets based on the four conditions. These sixteen sets were used in all experiments (with the island type and verb modifications). The English and Spanish sentences were translation-equivalents and as lexically-matched as possible. We did have to make several choices to ensure a clear test of island effects. First, the embedded subject in English was always an overt pronoun (either *they* or *you*), but a null pronoun in Spanish. We chose null pronouns in Spanish because we perceived them as more natural than overt pronouns in our sentences, and because their position relative to the verb is not observable, meaning that participants can posit it as preverbal or postverbal, as they prefer. We could have alternatively used an overt determiner phrase subject in a fixed position, but we decided not to do so because Torrego (1984) observed that there may be independent factors that influence the acceptability of the subject-verb vs verb-subject orderings, which in turn may be confounded with, or even interact with, island effects. While this is an interesting question in its own right, we abstract away from it here, allowing participants to posit the most acceptable subject position in their grammar, giving us a test of island effects alone. Second, in the *know* experiments, we used *want to know* / *querer saber* (rather than simply *know*) because *want to know* seems to highlight the interrogative nature of

the complement *wh*-clause, similar to *ask*. Third, in the embedded clause, we used verbs that we perceived to be transitively-biased to facilitate the interpretation of the *wh*-word as an embedded object. Finally, we presented embedded verbs in the conditional tense (e.g., *would reject*) to make a modifier interpretation of the *wh*-questions (particularly *cuándo* / *when*-islands) less likely (such an interpretation would make them an adjunct island rather than a *wh*-island). The full set of materials is available as Supplementary files (S1).[4]

In addition to the 16 experimental items, we selected 32 pre-tested fillers and seven pre-tested burn-in items per language that were distributed equally across the full range of acceptability from previous studies. The fillers and burn-in items were sentences from the theoretical syntactic literature —or modelled after its examples— that instanced different degrees of acceptability according to formal studies. For English, we took the items from Sprouse et al. (2013), who tested sentences on a 7-point scale and computed the most frequent score for each item. Based on that measure, we selected a set of sentences that evenly represented all seven ratings: there was one burn-in item and four or five fillers by rating. For Spanish, we took the sentences from Ortega-Santos (2020), who tested the acceptability of a number of sentences in Chilean, Venezuelan and Puerto Rican Spanish and provides the mean z-scores for each item. Given that our participants were speakers of yet another variety, we arbitrarily selected the Chilean variety as a reference and picked a set of sentences with a widespread distribution of z-scores: for the fillers, the range was −1.449 to 1.097 and the average

[4] While the Spanish and the English items were as similar as possible, some differences between the two languages inevitably remain, as an anonymous reviewer notes. For example, in Spanish, the subject DPs in the embedded conditions (e.g., *el político* 'the politician') had to be marked as masculine or feminine, while in English they had no gender marking. We do not think this affected the results, as the Spanish embedded conditions obtained similar mean ratings in items with masculine and feminine subjects (masculine non-island/embedded: mean = 0.369, SD = 0.610; feminine non-island/embedded: mean = 0.319, SD = 0.649; masculine island/embedded: mean = −0.975, SD = 0.688; feminine island/embedded: mean = −1.02, SD = 0.698).

was 0.108; for the burn-in items, the range was −1.545 to 1.108 and the average was 0.056.

5.3 *Presentation*

Items were distributed across four lists using a Latin Square procedure, such that participants rated each of the four conditions four times, and each time the item was from a unique item set (no repetitions of lexical items). Participants first saw instructions with three example items explicitly given ratings of 1, 4, and 7 to demonstrate the task. They then rated items themselves. The first seven items spanned the full range of acceptability based on our expectations, and they were presented in the same (pseudorandom) order for each participant. These items were not analyzed, as they were burn-in items presented to make sure that the participants saw the full range of possible acceptability prior to rating any trials that we would analyze. The next two items were two of the filler items (a very low and very high rating). The rest of the experiment contained the 16 experimental items for that particular list and 30 remaining fillers in a pseudorandom order such that there was at least one filler between two experimental items. The experiments were run using Qualtrics (Provo, Utah). It took participants about 10 minutes to complete an experiment.

6 Results

The 7-point scale data were z-score transformed by participant to eliminate common forms of scale biases prior to analysis. In all experiments, acceptability in the fillers comprised a wide range of ratings, as shown in Figure 2.

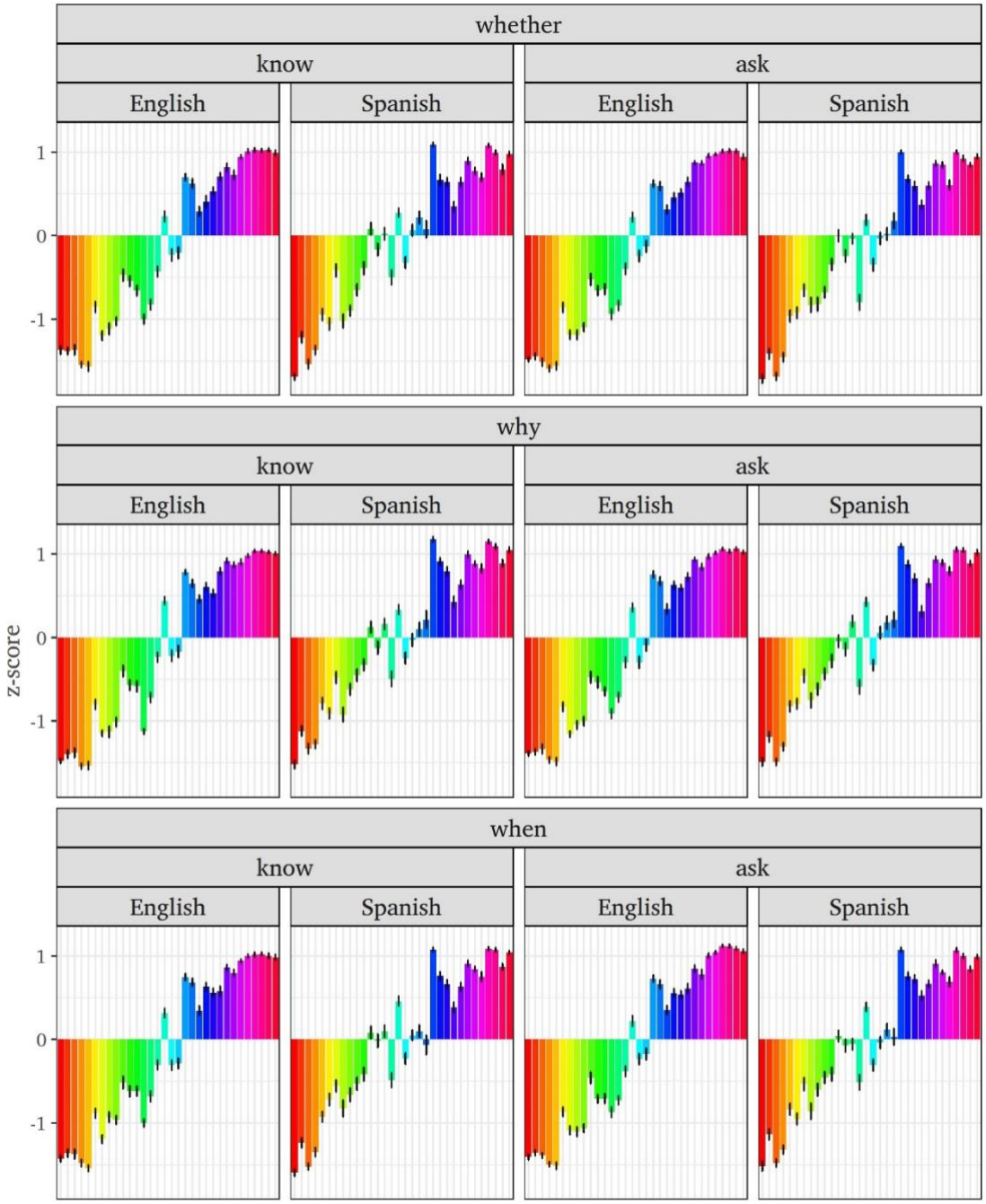

Figure 2: Filler acceptability by experiment. Each bar represents the mean acceptability of one filler. Error bars represent standard error.

We constructed linear mixed effects models for each of the questions of the study (described below). We then calculated two inferential statistics for the critical interaction terms: null hypothesis *p*-values using the lmerTest package (Kuznetsova et al. 2017), and Bayes factors (BF) using the BayesFactor package (Morey et al. 2022) in R (R Core Team 2023). We included BFs because they provide distinct information

to *p*-values – they represent the ratio of the probability of the data under the experimental hypothesis to the probability of the data under a null hypothesis. We interpreted a *p*-value less than .05 as statistically significant, a BF greater than 3 as meaningful evidence in favor of the presence of an interaction (i.e., the data is 3× more likely under the experimental hypothesis than under the null hypothesis), and a BF less than .33 as evidence against the presence of an interaction (i.e., the data is 3× more likely under the null hypothesis than under the experimental hypothesis). To check whether our BFs were robust to the choice of priors, we calculated them with the three different priors built-in to the BayesFactor package. In the text, we only report the BFs obtained with a medium width prior, but all three widths yield equivalent results unless otherwise indicated.

6.1 *The presence of island effects*

Our first question is whether each of the six island effects is present in the two languages. Figure 3 shows the interaction plots for each *wh*-island under each verb in the two languages, along with the differences-in-differences or DD scores, an estimate of the interaction term that we calculated as in (5). The mean z-scores by condition are shown in Table 4. To assess the presence of island effects statistically, we constructed linear mixed effects models crossing Structure × Position for each of the twelve island effects, with the maximal random effect structure that did not result in convergence failure. The presence of an island effect would show up as a significant Structure × Position interaction. Figure 3 also shows the *p*-values and Bayes factors for the interaction term in these models. The full results are shown in Table 5 for English and Table 6 for Spanish.

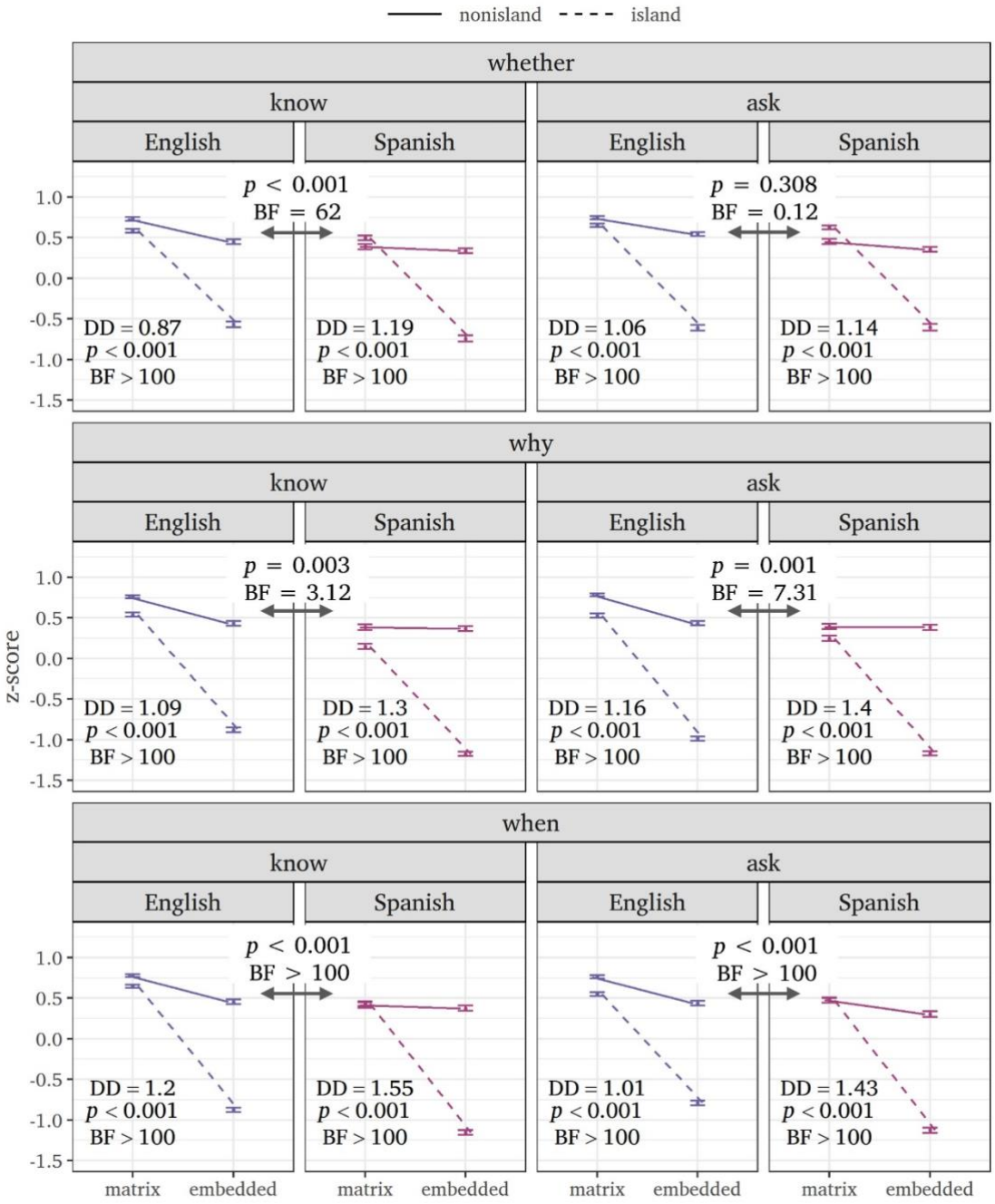

Figure 3: Interaction plots for *whether*, *why* and *when* islands under both *know* and *ask* in English and Spanish. For each island in each language, we show the differences-in-differences (DD) score, the *p*-value of the superadditive Structure × Position interaction indicating an island effect and the Bayes Factor (BF). Above the arrows, we also show the *p*-value and the BF for the three-way Structure × Position × Language interactions, which assess cross-linguistic differences in size (see Section 6.2).

	non-island matrix	non-island embedded	island matrix	island embedded
		English		
know whether	0.732 (0.450)	0.447 (0.573)	0.583 (0.494)	−0.568 (0.687)
ask whether	0.744 (0.378)	0.544 (0.477)	0.651 (0.415)	−0.612 (0.681)
know why	0.761 (0.342)	0.431 (0.581)	0.539 (0.498)	−0.879 (0.558)
ask why	0.781 (0.360)	0.433 (0.520)	0.527 (0.525)	−0.984 (0.530)
know when	0.777 (0.372)	0.453 (0.577)	0.646 (0.441)	−0.875 (0.540)
ask when	0.760 (0.405)	0.436 (0.570)	0.548 (0.521)	−0.789 (0.598)
		Spanish		
	non-island matrix	non-island embedded	island matrix	island embedded
know whether	0.390 (0.646)	0.336 (0.617)	0.495 (0.556)	−0.742 (0.744)
ask whether	0.451 (0.624)	0.358 (0.606)	0.624 (0.500)	−0.606 (0.842)
know why	0.386 (0.671)	0.366 (0.581)	0.147 (0.674)	−1.175 (0.536)
ask why	0.393 (0.685)	0.382 (0.644)	0.246 (0.663)	−1.168 (0.543)
know when	0.413 (0.661)	0.376 (0.597)	0.431 (0.585)	−1.153 (0.549)
ask when	0.477 (0.628)	0.304 (0.686)	0.472 (0.596)	−1.126 (0.619)

Table 4: Mean z-scores by language, condition, verb and island type. Standard deviations are shown within parentheses.

	Estimate	SE	*t*	*p*	BF
Know whether					
Intercept	0.299	0.028	10.710	< .001	
Structure	−0.291	0.019	−15.480	< .001	
Position	−0.359	0.020	−17.830	< .001	
Structure × Position	−0.217	0.018	−12.380	< .001	> 100
Ask whether					
Intercept	0.332	0.026	12.910	< .001	
Structure	−0.312	0.017	−17.980	< .001	
Position	−0.365	0.021	−17.410	< .001	
Structure × Position	−0.266	0.016	−16.880	< .001	> 100
Know why					
Intercept	0.213	0.024	8.800	< .001	
Structure	−0.383	0.020	−18.780	< .001	
Position	−0.438	0.020	−22.400	< .001	
Structure × Position	−0.272	0.021	−13.190	< .001	> 100
Ask why					
Intercept	0.189	0.019	10.100	< .001	
Structure	−0.418	0.018	−22.760	< .001	
Position	−0.465	0.021	−22.560	< .001	
Structure × Position	−0.291	0.018	−15.870	< .001	> 100
Know when					
Intercept	0.249	0.023	10.970	< .001	
Structure	−0.364	0.020	−18.220	< .001	
Position	−0.462	0.018	−25.620	< .001	
Structure × Position	−0.299	0.021	−14.440	< .001	> 100
Ask when					
Intercept	0.240	0.026	9.161	< .001	
Structure	−0.359	0.025	−14.555	< .001	
Position	−0.414	0.023	−18.393	< .001	
Structure × Position	−0.251	0.021	−11.892	< .001	> 100

Table 5: Results of the Structure × Position linear mixed models run on each of the English experiments and Bayes Factors (BF) for the Structure × Position interactions. The factors followed an effects coding scheme: Structure (non-island: −1, island: 1), Position (matrix: −1, embedded: 1). All models included random intercepts and Structure and Position slopes for participant and item. The Structure × Position interaction was included in the slopes whenever this converged.

	Estimate	SE	*t*	*p*	BF
***Saber si* 'know whether'**					
Intercept	0.120	0.038	3.137	.004	
Structure	−0.242	0.021	−11.563	< .001	
Position	−0.324	0.019	−16.637	< .001	> 100
Structure × Position	−0.296	0.014	−21.548	< .001	
***Preguntar si* 'ask whether'**					
Intercept	0.207	0.043	4.819	< .001	
Structure	−0.200	0.025	−7.866	< .001	
Position	−0.331	0.022	−14.750	< .001	> 100
Structure × Position	−0.285	0.019	−15.344	< .001	
***Saber por qué* 'know why'**					
Intercept	−0.069	0.031	−2.209	.038	
Structure	−0.445	0.024	−18.302	< .001	
Position	−0.336	0.030	−11.339	< .001	> 100
Structure × Position	−0.326	0.016	−20.245	< .001	
***Preguntar por qué* 'ask why'**					
Intercept	−0.036	0.034	−1.060	.301	
Structure	−0.425	0.027	−16.010	< .001	
Position	−0.357	0.025	−14.050	< .001	> 100
Structure × Position	−0.350	0.020	−17.410	< .001	
***Saber cuándo* 'know when'**					
Intercept	0.017	0.036	0.466	.645	
Structure	−0.376	0.024	−15.577	< .001	
Position	−0.406	0.021	−19.027	< .001	> 100
Structure × Position	−0.388	0.023	−16.677	< .001	
***Preguntar cuándo* 'ask when'**					
Intercept	0.032	0.031	1.021	.318	
Structure	−0.359	0.027	−13.381	< .001	
Position	−0.444	0.024	−18.353	< .001	
Structure × Position	−0.355	0.014	−25.330	< .001	> 100

Table 6: Results of the Structure × Position linear mixed models run on each of the Spanish experiments and Bayes Factors (BF) for the Structure × Position interactions. The factors followed an effects coding scheme: Structure (non-island: −1, island: 1), Position (matrix: −1, embedded: 1). All models included random intercepts and Structure and Position slopes for participant and item (except for *know whether*, where

only a Structure slope was included for item to ensure convergence). The Structure × Position interaction was included in the slopes whenever this converged.

In all cases, the island/embedded condition was much less acceptable than the other three conditions, resulting in the visual patterns indicative of superadditive Structure × Position interactions, which were confirmed statistically (all $p < .001$ and all BFs > 100). This suggests that all six island effects are present in both languages.

6.2 *The size of island effects*

Though there does not appear to be any cross-linguistic variation in the presence of island effects (a binary question), there could still be variation in the size of the island effects (a gradient question). For example, *wh*-island effects could be present but smaller in Spanish. To test this possibility, we constructed linear mixed effects models for each of the six island types, but crucially combined the results of the two languages. These models crossed Structure × Position × Language and included the maximal random effect structure that converged. A difference in island effect size between the two languages would show up as a significant three-way interaction. A summary of the three-way interaction effects is shown in Table 7. The full results of the models are available as Supplementary files (S2). The *p*-values and Bayes factors of the critical three-way interaction term are also shown in Figure 3 above the arrows.

	Estimate	SE	*t*	*p*	BF
know whether	−0.040	0.009	−4.259	< .001	62
ask whether	−0.009	0.009	−1.020	.308	0.12
know why	−0.027	0.009	−3.018	.003	3.12
ask why	−0.030	0.009	−3.325	.001	7.31
know when	−0.044	0.009	−5.174	< .001	> 100
ask when	−0.052	0.009	−5.693	< .001	> 100

Table 7: Summary of the critical Structure × Position × Language interaction effects obtained in the six Structure × Position × Language linear mixed models run on each island and verb combination. Bayes Factors (BF) for the interaction are also provided.

We find a statistically significant size difference for five out of the six island types by both null hypothesis testing ($p < .05$) and Bayes factor (BF > 3): *know whether*, *know why*, *ask why*, *know when*, and *ask when*. From these, all BFs are robust to prior widths except for *know why*, where the BF is inconclusive with a wide (BF = 1.65) and an

ultrawide prior (BF = 1.56). As for *ask whether*, it is not significant by *p*-value (p = .308) and shows evidence against a size difference by BF (BF = .12). It is important to note that the direction of the size differences we observed is opposite to the one that might be expected under a weaker version of the original cross-linguistic claim – Spanish island effects are in fact larger than English island effects by about .3 z-units on average. This suggests that while there is evidence of a gradient form of cross-linguistic variation for four or five out of six island types, it is not in line with the original observation.

6.3 *Individual variation*

Our third question is whether there is more individual variation in Spanish than English. We ask this because there can be individual variation, representing different idiolects, that is obscured by focusing on sample means. Idiolectal variation could explain the apparent discrepancy between our results and Torrego's (1984) observations. In other words, even if, overall, we find island effects, some speakers of Spanish may manifest no or smaller island effects while others show large island effects (i.e., between-participant variation). It is also possible that some speakers of Spanish show more variability within their own judgments than English speakers (i.e., within-participant variation), suggesting that they may have two grammars at their disposal. To be clear, we expect some amount of variation both between- and within-participants. That is the nature of behavioral studies (they are inherently noisy). The critical question is whether Spanish shows more variation than English in one or both of these ways.

To look for between-participant variation, we plot two sets of distributions: the island effect sizes calculated as by-participant DD scores (Figure 4), and the by-participant z-score means of the island/embedded conditions (Figure 5; the non-island/embedded conditions are also shown as a control; see Kush et al. 2018; 2019; Bondevik et al. 2021 for similar analyses). We can look for a visual pattern suggesting two or more populations of speakers, which would appear as a bimodal (or multimodal) distribution. We see no obvious sign of bimodality in the Spanish experiments, either in the DD scores or in the island/embedded conditions. Some signs of bimodality are instead observable in the English experiments (e.g., the *ask whether* DD scores and the *ask when* DD scores and island/embedded conditions).

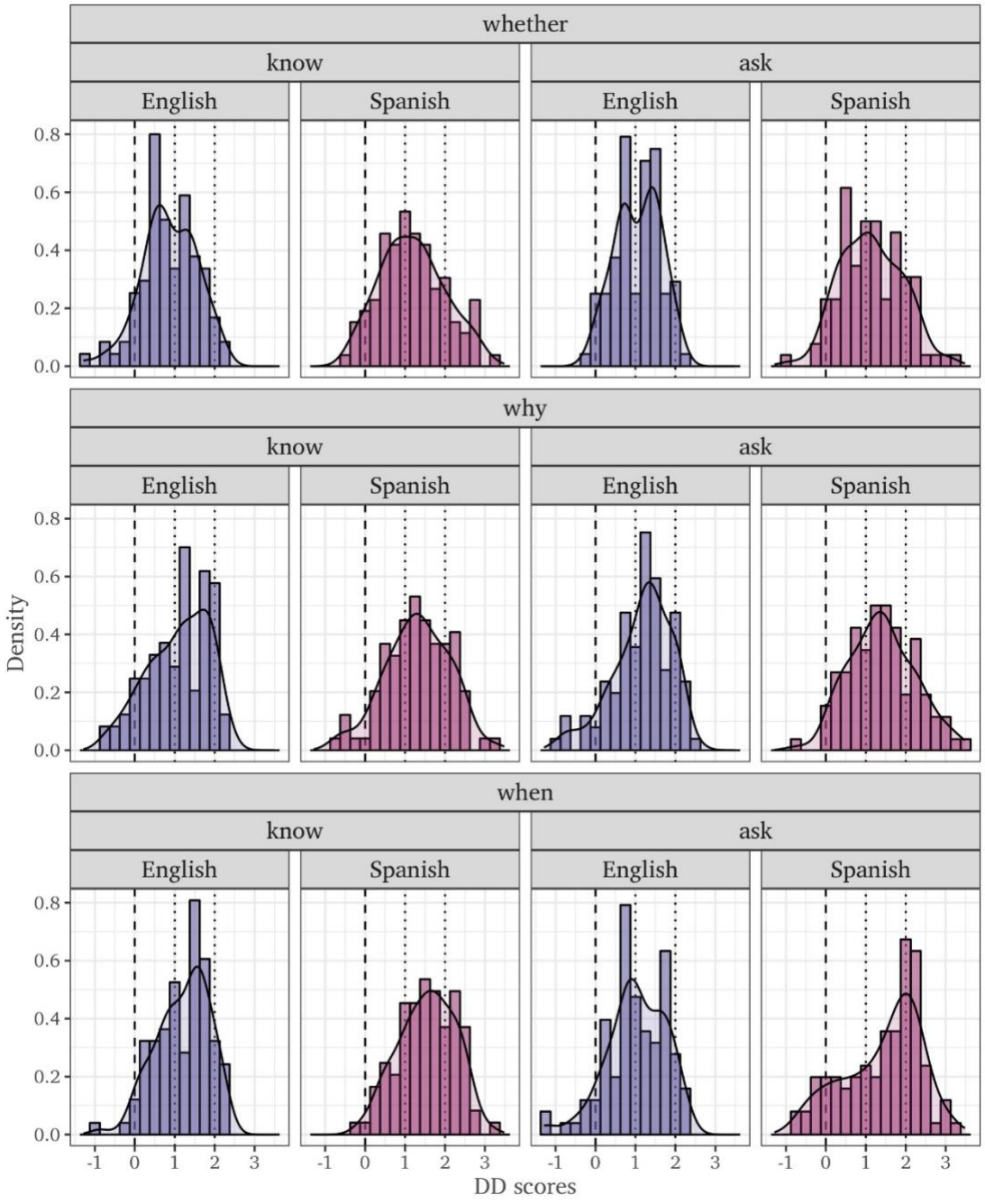

Figure 4: Distribution of by-participant DD scores in the English and Spanish experiments. The dashed vertical line marks the limit between DD scores higher than 0 (indicative of an island effect) and DD scores lower than 0 (indicative of no island effect).

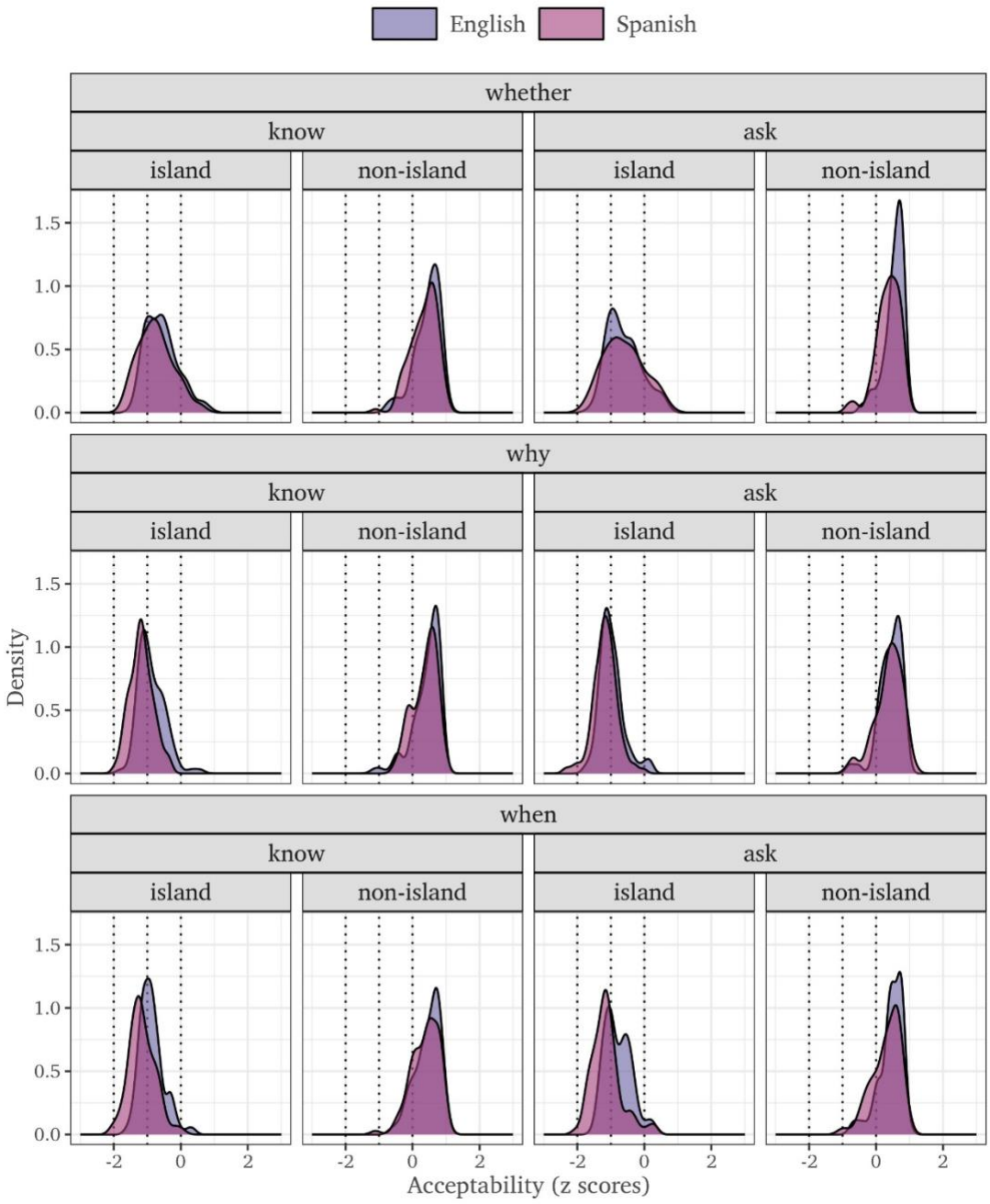

Figure 5: Distribution of by-participant z-score means in the island/embedded and non-island/embedded conditions of the English and Spanish experiments.

We tested for bimodality statistically using the multimode package (Ameijeiras-Alonso et al. 2021) in R (R Core Team 2023). We chose three tests that instantiate the most common approaches to identifying multimodality: the dip test (Hartigan & Hartigan 1985), the excess mass test (Müller & Sawitzki 1991; Cheng & Hall 1998;

Ameijeiras-Alonso et al. 2019), and the bandwidth test (Silverman 1981; Hall & York 2001). What we looked for is a statistically significant effect of multimodality in Spanish but not in English in either the DD scores or the individual condition z-scores. The full list of *p*-values for each of these tests is shown in Table 8. Crucially, there are no statistically significant effects for any of the Spanish islands under any of the tests, corroborating the visual inspection of Figure 4 and Figure 5, and suggesting that there is no evidence of two or more populations of speakers of Spanish in our studies. In English, there are significant effects in the *ask whether* DD scores and in the *know whether* and *ask when* island/embedded condition, both under the excess mass test (a significant effect is also observed in the English *ask when* non-island/embedded condition). If anything, these results suggest that English is more variable than Spanish, contrary to the hypothesis we set out to explore.

	Dip test		Excess mass test		Bandwidth test	
	English	Spanish	English	Spanish	English	Spanish
DD scores						
know whether	.256	.836	.060	.454	.308	.226
ask whether	.138	.410	**.026**	.120	.066	.600
know why	.370	.900	.118	.632	.516	.368
ask why	.748	.460	.220	.080	.506	.438
know when	.986	.994	.862	.908	.310	.874
ask when	.714	.998	.314	.966	.312	.774
island/embedded						
know whether	.302	.944	**.036**	.592	.124	.558
ask whether	.556	.810	.164	.382	.278	.540
know why	1	1	.994	.974	.238	.508
ask why	.998	.962	.936	.632	.214	.884
know when	.958	.786	.626	.318	.292	.400
ask when	.086	.864	**.008**	.398	.076	.316
non-island/embedded						
know whether	.622	.932	.150	.596	.358	.128
ask whether	.998	.506	.956	.142	.748	.116
know why	.986	.672	.832	.256	.344	.128
ask why	.670	.852	.186	.348	.088	.190
know when	.962	.438	.638	.126	.846	.138
ask when	.260	.976	.020	.746	.406	.322

Table 8: *P*-values obtained in three multimodality tests (the dip test, the excess mass test and the bandwidth test) run on the DD scores, the island/embedded conditions and the non-island/embedded conditions (which are shown as a control). *P*-values indicating significant results are bolded.

Within-participant variation is displayed in Figure 6 and Figure 7. Figure 6 plots each of the four ratings of the island/embedded condition for each participant: each horizontal line represents a participant; their individual ratings are represented by the colored dots and their mean rating is represented by the black circles. We categorized the four ratings of the island/embedded condition given by each participant as either below or above their mid-point of the scale (0). Then, we counted how many participants rated all four tokens below 0, how many rated three out of four below 0, etc. The end result is counts for five categories of speakers for each of the two

languages, which we then converted to proportions of the total sample for each experiment. These proportions are shown as a stacked bar plot in Figure 7. Visual inspection of the proportions in Figure 7 suggests that the two languages show similar amounts of within-participant variability.

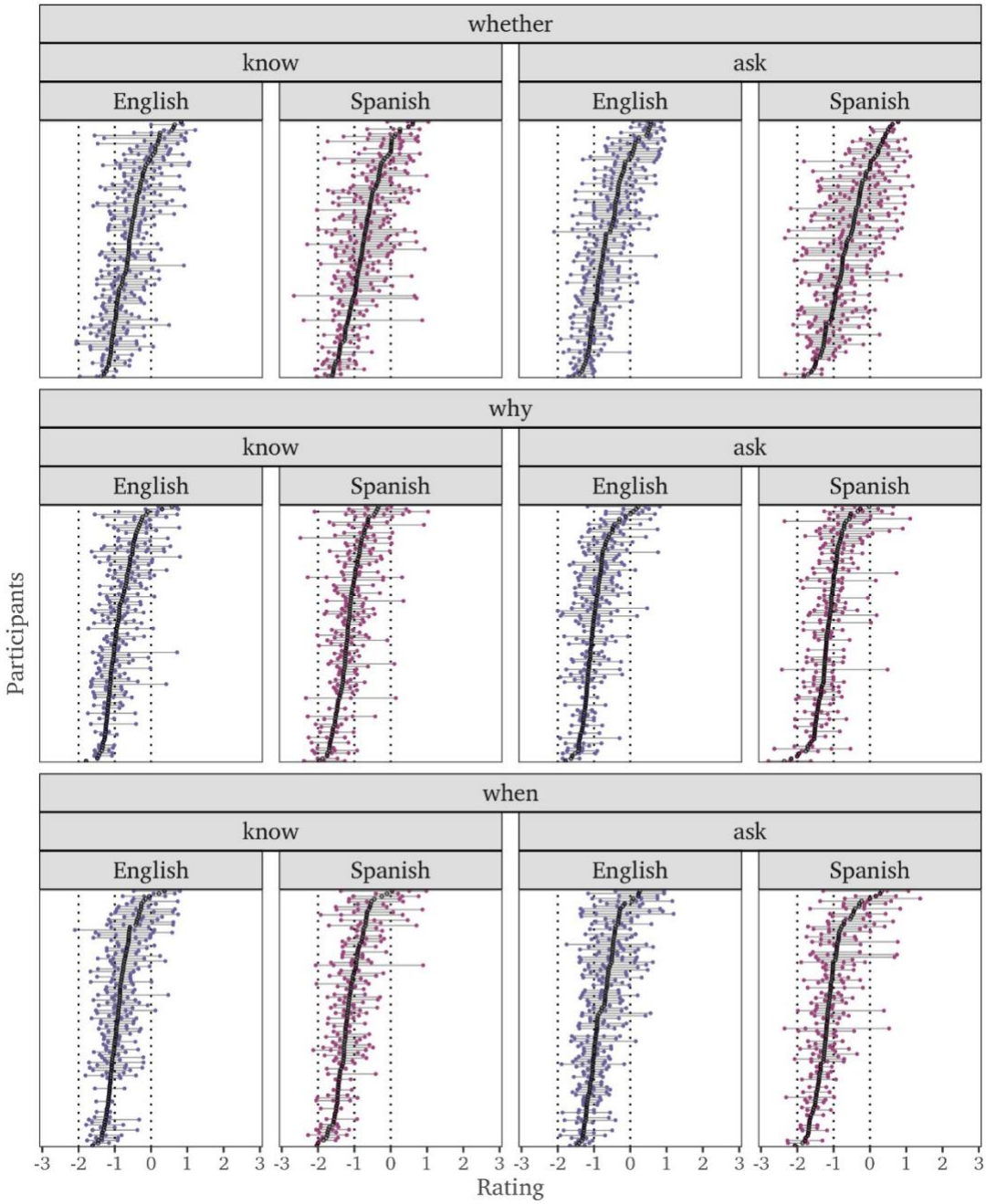

Figure 6: Distribution of by-participant z-scores in the island/embedded conditions of all the experiments. Each horizontal line represents a participant; the dots on each line indicate the location of each of the participant's observations on the z-score scale, and the black circles show the participants' mean (for each experiment, participants are ordered by their mean z-score).

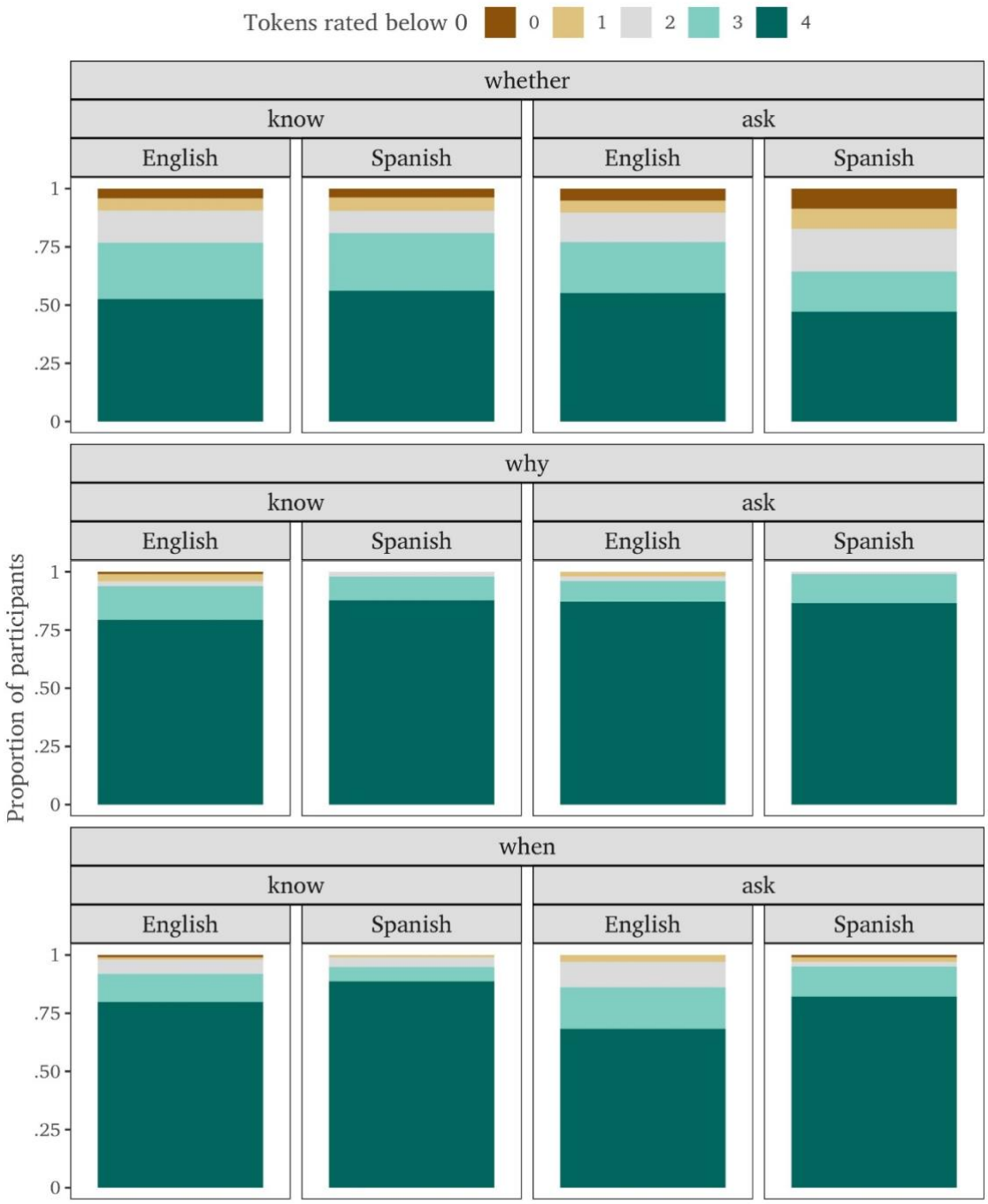

Figure 7: Proportion of participants that rated zero, one, two, three or four out of four tokens below their midpoint of the z-score scale (0) in each of the experiments.

We tested whether the counts of the five categories differed statistically between the two languages by using Fisher's exact test (on a 2×5 contingency table crossing language and the five count categories). We found that only *ask when* yielded a significant effect (p = .028), while the other five island types were non-significant (*know when*: p = .415; *ask whether*: p = .429; *know whether*: p = .922; *ask why*: p = .477; *know why*: p = .282). As we can see in Figure 7, the significant difference for *ask when* appears to be driven by English showing more within-participant variability (fewer participants with 4/4 ratings below 0) than Spanish. This runs contrary to the hypothesis that perhaps Spanish shows more within-participant variability.

Taken together, the results of these two analyses suggest that there is no more between- or within-participant variation in Spanish than in English, at least for the participants in our experiments.

6.4 *Post-hoc analyses*

A reviewer helpfully asked whether our data set could be explored for additional facts about island effects, in particular (i) whether there are differences in effect size between the island types within a language, and (ii) whether there are differences in individual variation between island types within a language. Though these questions are not part of our theoretical exploration (we are focused on cross-language, not within-language, differences), we are happy to provide these analyses for researchers who may be working on these questions.

To assess differences in effect size between island types, we constructed 2×2×2 linear mixed effects models for each possible pair of island types (within each verb), and we calculated the p-value of the three-way interaction term as an indicator of a significant difference in the size of the island effects. Table 9 reports these results for the twelve possible comparisons. Because this is a post-hoc exploration, we present the uncorrected p-values and three possible corrections: Dunn correction of every term in the models (81 comparisons; the most conservative correction); Dunn correction of the interaction terms alone (twelve comparisons; a less conservative correction); and false

discovery rate (Benjamini and Hochberg 2000) of the interaction terms (twelve comparisons; the least conservative correction). We use the Neyman-Pearson asterisk to indicate when the correction would yield significance at an alpha criterion of .05.

	p	Dunn 81	Dunn 12	FDR
English				
know				
whether – when	< .001	*	*	*
whether – why	.010	n.s.	n.s.	*
why – when	.084	n.s.	n.s.	n.s.
ask				
whether – when	.399	n.s.	n.s.	n.s.
whether – why	.100	n.s.	n.s.	n.s.
why – when	.019	n.s.	n.s.	*
Spanish				
know				
whether – when	< .001	*	*	*
whether – why	.127	n.s.	n.s.	n.s.
why – when	.001	n.s.	*	*
ask				
whether – when	< .001	*	*	*
whether – why	.001	n.s.	*	*
why – when	.803	n.s.	n.s.	n.s.

Table 9: Uncorrected *p*-values obtained in twelve Structure × Position × Island linear mixed models (one for each pair of islands within language and verb) and their interpretation as significant (*) or not significant (n.s.) under the Dunn correction of every term in the models (Dunn 81), the Dunn correction of the interaction terms alone (Dunn 12) and Benjamini and Hochberg's (2000) false discovery rate of the interaction terms (FDR).

We know of no theory that predicts differences in effect sizes between island types (or, differences in effect sizes more generally), so we leave this as information for other researchers to use if they are pursuing such a theory.

To assess differences in individual variation between island types within a language, we calculated Fisher's exact tests on 2×5 contingency tables comparing each

pair of island types (within language and verb) and the number of tokens that were rated below the midpoint of the scale (zero through four). Table 10 reports these results for the twelve possible comparisons. Because this is a post-hoc exploration, we present the uncorrected *p*-values and two possible corrections: Dunn correction of the twelve comparisons (the more conservative correction), and false discovery rate (Benjamini and Hochberg 2000) of the twelve comparisons (the less conservative correction).

	p	Dunn 12	FDR
English			
know			
whether-when	.001	*	*
whether-why	.001	*	*
why-when	.551	n.s.	n.s.
ask			
whether-when	.096	n.s.	n.s.
whether-why	< .001	*	*
why-when	.006	n.s.	*
Spanish			
know			
whether – when	< .001	*	*
whether – why	< .001	*	*
why – when	.477	n.s.	n.s.
ask			
whether – when	< .001	*	*
whether – why	< .001	*	*
why – when	.565	n.s.	n.s.

Table 10: Uncorrected *p*-values obtained for the Fisher's exact tests and their interpretation as significant (*) or not significant (n.s.) under the Dunn correction of the twelve tests (Dunn 12) and Benjamini and Hochberg's (2000) false discovery rate (FDR).

Again, we know of no theory that predicts differences in the amount of variation between island types, so these results are presented for informational purposes only. We also note that the categorization that we use (above/below the midpoint) means that island types with higher ratings for the island/embedded condition are more likely to

show variability. This likely explains why *whether* islands tend to show more variation than the other island types. So these results should be interpreted with caution. (This is less of a concern for the cross-linguistic comparison because we were comparing the same island type, which under default assumptions, should show similar absolute ratings.)

7 Discussion

7.1 *The empirical facts of* wh*-island effects in Spanish and English*

In a series of twelve acceptability judgment experiments, we examined three types of *wh*-island effects (*whether*, *why* and *when* island effects) under two embedding verbs (*know* and *ask*) in English and Spanish translation-matched sentences with *wh*-dependencies. The first empirical question we sought to address was whether *wh*-extraction showed island effects in these contexts not only in English but also in Spanish. The answer to this question appears to be yes – Spanish shows clear *wh*-island effects in all three *wh*-clauses under the conditions in which these effects were also observed in English.

Our findings replicate previous work showing that there are island effects in both English (Sprouse 2007; Sprouse et al. 2011; 2012; 2016; Almeida 2014; Michel 2014; Aldosari 2015; Ortega-Santos et al. 2018; Pham et al. 2020) and Spanish (López-Sancio 2015; Ortega-Santos et al. 2018; Pañeda et al. 2020; Rodríguez & Goodall 2020; Stigliano & Xiang 2021; Pañeda & Kush 2022), but they also extend it: previous studies usually tested fewer types of *wh*-islands and fewer verbs, they had a smaller sample size and, crucially, they generally examined the two languages separately, with different materials and, sometimes, under different conditions that hamper any cross-linguistic comparisons. In this context, our study significantly broadens our knowledge about how *wh*-islands compare in English and Spanish.

The second empirical question we sought to address was whether there is a gradient difference in effect size between *wh*-islands in English and Spanish. One possibility we considered was that the locus of cross-linguistic variation lies in effect size rather than the categorical presence/absence of island effects, perhaps with Spanish showing smaller island effects than English. What we found is that there is a difference in effect size for four or five out of the six island types that we tested, but it goes in an

unexpected direction – Spanish island effects tend to be slightly larger than English island effects (with translation-matched materials). Thus, not only are there *wh*-island effects in Spanish, but they also do not appear to be mild, in contrast to what previous work has suggested (Pañeda et al. 2020; Pañeda & Kush 2022). We note that the experimental findings supporting this possibility were obtained under different conditions that may impact effect sizes. For instance, Pañeda & Kush (2022) obtained very small *whether* island effects (DD = 0.22, 0.38), but their sentences did not contain bare extractees, like ours, but rather complex or "d-linked" extractees, which are known to reduce *wh*-island effects (Pesetsky 1987; Goodall 2015; Atkinson et al. 2016; Villata et al. 2016).[5] Similarly, Pañeda et al.'s (2020) island effect sizes were based on binary judgments obtained in a speeded task, in contrast to our effect sizes, which were based on 7-point scale judgments obtained in an untimed task. Because of these differences, our study is not directly comparable to those previous ones.

An anonymous reviewer asks if we can propose a theory for the surprising observation that island effects are larger in Spanish than English. While the experimental literature has cataloged a number of results showing minor variation in the island effect sizes across languages (Sprouse & Villata 2021 for a review), we know of no theories of differences in effect sizes for island effects (or any other acceptability judgment effect) that we can use to interpret the contrast. While there are frameworks that can capture differences in effect sizes (e.g., Keller's 2000 Linear Optimality Theory or Featherston's 2005 Decathlon model), our understanding is that these frameworks do not currently predict those differences, but are rather at an earlier stage: namely, they are setting the constraint values (based on judgment studies) necessary to make such predictions in the future. We are therefore reluctant to propose an entirely new theory based on this surprising result.

That said, we can rule out some possible explanations. For example, one possibility would be that island effects are equal in size in the two languages, but the

[5] Interestingly, Pañeda & Kush (2022) also tested *when* island effects, and their effect sizes (*know when*: 1.09, *ask when*: 1.39) were more similar to ours (*know when*: 1.20, *ask when*: 1.43), despite their using complex extractees. Future work should examine whether complex extractees affect different *wh*-islands differentially.

English island/embedded condition hit the floor of the scale, yielding an underestimation of the effect size. However, the English island/embedded conditions are all at or less than −1 z-scores on average, while some of the English fillers are rated on average as low as −1.5. This suggest that the island/embedded conditions did not hit the floor, and therefore that the English effect is not underestimated.

Another possibility is that the Spanish violation conditions were perceived as less acceptable than their English counterparts due to differences in the acceptability of the fillers that we used for each language, which were not translation-matched. Filler sentences can act as a benchmark against which the experimental conditions are evaluated (Cowart 1997). Thus, if the fillers were less acceptable in English than in Spanish, this might have caused the island/embedded conditions to be perceived as more acceptable in English than Spanish. We do not think this was the case, though, because, impressionistically, the fillers seem to span a similar range of acceptability in both languages (Figure 2).

Finally, a reviewer wonders whether the larger island effects in Spanish could be related to cross-linguistic differences regarding the interpretation of the embedded third person plural subject that appears in half of the items (overt *they* in English and its null counterpart in Spanish). The difference is that, in the matrix conditions, this subject might be interpreted as co-referential with the extractee *who* in English but not in Spanish. The reviewer notes that co-referentiality could help process the long-distance dependency, making the sentences easier to process and more acceptable in English than in Spanish. If English *they* facilitated processing and increased the acceptability of the sentences due to co-referentiality, the subset of the English items that contained *they* should be more acceptable than the other half of the items, where the embedded subject was *you* and co-referentiality with *who* was not possible. This was not the case: overall, the English *they* and *you* items obtained very similar ratings in the two relevant matrix conditions (non-island/matrix *they*: mean = 0.749, SD = 0.402; non-island/matrix *you*: mean = 0.770, SD = 0.368; island/matrix *they*: mean = 0.567, SD = 0.494; island/matrix *you*: mean = 0.597, SD = 0.479). Thus, the presence of overt *they* in English did not affect ratings, and therefore it is unlikely that the larger island effects in Spanish are attributable to the contrast between English *they* and its Spanish null counterpart.

The final empirical question that we sought to address was whether there is increased individual variation in Spanish as compared to English (both between- and within-participants), as this could be a potential explanation for some of the variability in the observations reported in the literature on Spanish. We find no evidence of greater between- or within-participant variability in Spanish (and, in fact, the only island that showed increased variability was in English). This provides additional support for the similarity between the two languages. The question whether there could be more variability in Spanish than English was reasonable given the contrast between Torrego's (1984) observations and recent experimental findings, and given that experimental studies have found increased between and within-participant variability, both in other languages (e.g., Norwegian; see Kush et al. 2018; 2019; Kobzeva et al. 2022; Kush & Dahl 2022) and in Spanish *when wh*-islands with "d-linked" fillers (Pañeda & Kush 2022). However, the Spanish speakers in our study do not show more variation than the set of English speakers. This finding helps to underscore that the island effects that we observed are likely a robust part of Spanish for these speakers. On that note, it seems valuable for future work to continue to probe individual variation because it can potentially reveal more information on factors that govern acceptability (cognitive, dialectal, etc.) or help to establish that the effect is indeed robust.

7.2 *Theoretical implications*

We have shown that *wh*-extraction gives rise to island effects in both English and Spanish and that these effects are not smaller in Spanish. The interpretation of our results and their implications for different theories depend on two different assumptions: (i) whether one assumes a *single* source for *wh*-island effects, and (ii) whether one assumes a *common* source for the effects in English and Spanish.

If one assumes a single, shared source for *wh*-island effects, then our findings simplify the theories that we considered above by removing the need to account for variation in Spanish. For the Subjacency approach, this would entail that the critical bounding node for Spanish is IP, similar to English. For the Phase-Impenetrability approach, this would entail that the C head in Spanish embedded questions licenses a single specifier position that hosts the *wh*-items that introduce embedded questions, just as in English (and that the position is, crucially, a phase edge). For Relativized Minimality, our results would entail that the *wh*-interveners that we tested in Spanish

all share a critical movement feature with the extracted *wh*-items, and that the structural position of the intervener is the same type as the landing position of the extracted *wh*-items. For the Information-Structure-based approach, our results suggest that the embedded *wh*-questions that we tested are all likely to be considered backgrounded by speakers of Spanish (and therefore would show as such under the backgroundedness/focus diagnostics that Erteschik-Shir 1973 identifies). For WM-related processing-difficulty approaches, our results suggest that the encoding and/or retrieval cues in Spanish are similar to those in English, such that the processing of island structures while processing a long-distance dependency creates an overload in the WM system. All of these accounts predict that *wh*-island effects should be present across comparable configurations in Spanish and English beyond the sentences we tested. Whether there is such cross-linguistic uniformity is an empirical question for future research.

Relaxing the single source assumption while maintaining the common source assumption would admit more flexibility regarding the possibility of cross-linguistic variation. Under a multiple constraint approach, the fact that Spanish and English exhibit island effects entails that the test sentences violate *at least one* shared constraint, but still allows variation among any remaining constraints included in the model. The abstract logic holds regardless of the set of constraints one adopts, but the implications vary depending on the constraints chosen by the analyst. We offer a hypothetical example where a multiple constraint approach would lead to different conclusions than a single constraint one. Suppose that Subjacency exists alongside a universal semantic/Information-Structural constraint that independently blocks *wh*-extraction from embedded questions. In this scenario, our test sentences would be ruled out in English and Spanish even if the two languages differed in their bounding nodes. Thus, the presence of island effects in both languages would not necessarily correspond to alignment in underlying grammar – a conclusion that might be welcomed by proponents of parameterized Bounding or Phase Theory. Importantly, the multiple constraint model in the above scenario predicts that if the contribution of the shared constraint(s) was neutralized or factored out, cross-linguistic differences would emerge: *wh*-island effects would persist in English, but not be found in Spanish. Once again, we leave testing these predictions to future experiments. Before moving on, it is important to note that

a multiple constraint approach in which both/all constraints are acquired from evidence during language acquisition would likely create a difficult learning problem for the child, as it would not be clear how to apportion responsibility across the constraint set. As such, the multiple constraint approach might only be tractable if one or more of the constraints is universal (as in the hypothetical example above).

Finally, we could relax the assumption that the source of the *wh*-island effects is the same in both languages. It is possible that the English and Spanish *wh*-island effects arise for different reasons. This style of reasoning has been called the *eclectic approach* by Chaves & Putnam (2020). An example of such an approach can be found in Christensen et al. (2013), who observed low acceptability for extraction out of a *wh*-structure in Danish (about 1.2 z-units). They argue based on evidence from an fMRI experiment that the low acceptability is driven by sentence processing difficulty rather than a grammatical constraint, in contrast to English (e.g., Sprouse et al. 2012; Yoshida et al. 2014). Eclectic approaches allow for more degrees of freedom than common source approaches. Theories of the acquisition of eclectic approaches to islands would likely entail the child tracking multiple pieces of evidence (perhaps across dependency types) to determine the sources of the constraints.

Broadly, our results suggest new research directions for two of the major overarching theoretical questions surrounding island effects: the extent of the cross-linguistic variation of island effects, and the source of island effects. For cross-linguistic variation, our results join a growing number of experimental studies that have shown that *wh*-island effects do in fact exist in languages that were thought to not have them, such as Brazilian Portuguese (Almeida 2014), Italian (Sprouse et al. 2016), Norwegian (Kush et al. 2018; 2019), and Mandarin (Chen 2024). Italian and Spanish together formed the initial empirical argument for a highly constrained theory of cross-linguistic variation, instantiated at the time through parameters (Rizzi 1982; Torrego 1984), whereas Scandinavian languages raised the possibility that languages with *wh*-movement could exhibit even fewer island effects (Engdahl 1982), and Mandarin raised the possibility that the syntactic operations underlying *wh*-in-situ, such as covert movement (Huang 1982) and/or unselective binding (Nishigauchi 1986; Pesetsky 1987; Cheng 1991; Tsai 1994), could also be immune to island effects. The results that have been building within the experimental syntax literature challenge all of these

conclusions. This raises the question of what exactly the empirical contours of cross-linguistic variation are in (*wh*- and other) island effects. Though there has been progress on this front in recent years (see Sprouse & Villata 2021 for a review up to that date), the field has still only tested a small number of languages compared to the number that have contributed to the development of previous theories of cross-linguistic variation of island effects. We see our study as a small contribution to this effort; and we see our results as demonstrating the value in systematically re-testing the languages that have contributed to these theories. While our results and the results of the empirical studies listed above have uncovered island effects where there were presumed to be none, they do not conclusively determine what set of factors bear causal responsibility for the effects themselves. Future work should systematically explore competing predictions of proposals already on the market and explore new proposals to account for effects that traditional accounts have difficulty explaining.

8 Conclusion

In a series of twelve acceptability judgment experiments, we examined *wh*-extraction from three types of embedded *wh*-questions (*whether*, *why* and *when* island effects) under two embedding verbs (*know* and *ask*) in both Spanish and English sentences. Our goal was to explore the original observation by Torrego (1984) that Spanish does not show *wh*-island effects in sentences with object *wh*-extraction from embedded *wh*-questions introduced by non-arguments. We found: (i) *wh*-island effects for both languages for all six island types tested, (ii) larger island effects for Spanish compared to English for most of the island types tested, and (iii) no evidence of additional between- or within-participant variation for Spanish compared to English. These results run contrary to both the original, binary version of the cross-linguistic claim and a plausible gradient variant. Our findings replicate previous experimental work showing that there are island effects in both languages (Sprouse 2007; Sprouse et al. 2011; 2012; 2016; Almeida 2014; Michel 2014; Aldosari 2015; López-Sancio 2015; Ortega-Santos et al. 2018; Pañeda et al. 2020; Pham et al. 2020; Rodríguez & Goodall 2020; Stigliano & Xiang 2021; Pañeda & Kush 2022), and also extend it to a wider range of relevant island types, using relatively large samples of participants, and translation-matched materials. Our results suggest that the European Spanish spoken by the participants that volunteered for our study is clearly a *wh*-island language, and therefore that the use of

European Spanish as evidence for theories that encode cross-linguistic variation in *wh*-island effects may need to be reconsidered.

Abbreviations

COND: conditional; CP: complementizer phrase; BF: Bayes Factor; DD: differences-in-differences; d-linked: discourse-linked; F: feminine; fMRI: functional magnetic resonance imaging; IP: inflectional phrase; M: masculine; PL: plural; PST: past; RM: Relativized Minimality; SE: standard error; SG: singular.

Data availability

The data and data analysis script are available at https://osf.io/xztvy/?view_only=274bd9e9dca44c498caf5c743bcec656

Author's contributions

Claudia Pañeda: Conceptualization; Methodology; Software; Formal analysis; Investigation; Writing – Original Draft; Writing - Review & Editing; Visualization; Project administration. **Sandra Villata:** Conceptualization; Methodology; Software; Writing - Review & Editing. **Dave Kush**: Conceptualization; Writing - Review & Editing. **Jon Sprouse**: Conceptualization; Methodology; Formal analysis; Writing - Review & Editing; Supervision; Funding acquisition.

References

Abeillé, Anne & Hemforth, Barbara & Winckel, Elodie & Gibson, Edward. 2020. Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition* 204. 104293. DOI: https://doi.org/10.1016/j.cognition.2020.104293

Abrusán, Márta. 2014. *Weak island semantics*. New York, NY: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199639380.001.0001

Aldosari, Saad Mohammed. 2015. *The role of individual differences in the acceptability of island violations in native and non-native speakers*. Lawrence, KS: University of Kansas dissertation.

Almeida, Diogo. 2014. Subliminal *wh*-islands in Brazilian Portuguese and the consequences for syntactic theory. *Abrallin* 13(2). 55–93. DOI: https://doi.org/10.5380/rabl.v13i2.39611

Ameijeiras-Alonso, José & Crujeiras, Rosa M. & Rodríguez-Casal, Alberto. 2019. Mode testing, critical bandwidth and excess mass. *Test* 28. 900–919. DOI: https://doi.org/10.1007/s11749-018-0611-5

Ameijeiras-Alonso, José & Crujeiras, Rosa M. & Rodríguez-Casal, Alberto. 2021. multimode: An R package for mode assessment. *Journal of Statistical Software* 97(9). 1–32. DOI: https://doi.org/10.18637/jss.v097.i09

Atkinson, Emily & Apple, Aaron & Rawlins, Kyle & Omaki, Akira. 2016. Similarity of *wh*-phrases and acceptability variation in *wh*-islands. *Frontiers in Psychology* 6. 2048. DOI: https://doi.org/10.3389/fpsyg.2015.02048

Belletti, Adriana & Friedmann, Naama & Brunato, Dominique & Rizzi, Luigi. 2012. Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua* 122(10). 1053–1069. DOI: https://doi.org/10.1016/j.lingua.2012.02.007

Benjamini, Yoav & Hochberg, Yosef. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics* 25(1). 60–83. DOI: https://doi.org/10.3102/10769986025001060

Bondevik, Ingrid & Kush, Dave & Lohndal, Terje. 2021. Variation in adjunct islands: The case of Norwegian. *Nordic Journal of Linguistics* 44(3). 223–254. DOI:

https://doi.org/10.1017/S0332586520000207

Boeckx, Cedric. 2012. *Syntactic islands*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139022415

Borer, Hagit. 1984. *Parametric syntax: Case studies in Semitic and Romance languages*. Berlin: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110808506

Bresnan, Joan. 1977. Variables in the theory of transformations. In Culicover, Peter W. & Wasow, Thomas & Akmajian, Adrian (eds.), *Formal syntax*, 157–196. New York, NY: Academic Press.

Chaves, Rui P. & Putnam. Michael T. 2020. *Unbounded dependency constructions: Theoretical and experimental perspectives*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198784999.001.0001

Chen, Xu. 2024. *Wh-island effects in Chinese: A formal experimental study*. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/la.282

Cheng, Lisa L.-S. 1991. *On the typology of wh-questions*. Cambridge, MA: MIT dissertation.

Cheng, Ming-Yen & Hall, Peter. 1998. Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society* 60(3). 579–589. DOI: https://doi.org/10.1111/1467-9868.00141

Chmielewski, Michael & Kucker, Sarah C. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11(4). 464–473. DOI: https://doi.org/10.1177/194855061987514

Chomsky, Noam. 1964. *Current issues in linguistic theory*. The Hague: Mouton.

Chomsky, Noam. 1973. Conditions on Transformations. In Anderson, Stephen & Kiparsky, Paul (eds.), *A Festschrift for Morris Halle*, 232–286. Holt, UK: Rinehart and Winston.

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

Chomsky, Noam. 1995. *The minimalist program.* Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/9780262527347.001.0001

Chomsky, Noam. 2000. Minimalist inquiries: The framework. In Martin, Roger & Michaels, David & Uriagereka, Juan (eds.), *Step by step. Papers in Minimalist*

Syntax in honor of Howard Lasnik, 89–153. Cambridge, MA: MIT Press.

Chomsky, Noam. 2001. Derivation by phase. In Kenstowicz, Michael (ed.), *Ken Hale: A life in language*, 1–52. Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/4056.003.0004

Christensen, Ken Ramshøj & Kizach, Johannes & Nyvad, Anne Mette. 2013. Escape from the island: Grammaticality and (reduced) acceptability of wh-island violations in Danish. *Journal of Psycholinguistic Research* 42. 51–70. DOI: https://doi.org/10.1007/s10936-012-9210-x

Citko, Barbara. 2014. *Phase Theory: An introduction.* Cambridge, UK: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139644037

Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

Cuneo, Nicole & Goldberg, Adele E. 2023. The discourse functions of grammatical constructions explain an enduring syntactic puzzle. *Cognition* 240. 105563. DOI: https://doi.org/10.1016/j.cognition.2023.105563

Dennis, Sean A. & Goodson, Brian Matthew & Pearson, Chris. 2020. Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting* 32(1). 119–134. DOI: https://doi.org/10.2308/bria-18-044

Engdahl, Elisabet. 1982. Restrictions on unbounded dependencies in Swedish. In Engdahl, Elisabet & Ejerhed, Eva (eds.), *Readings on unbounded dependencies in Scandinavian languages*, 151–174. Stockholm: Almqvist and Wiksell International.

Erteschik-Shir, Nomi. 1973. *On the nature of island constraints*. Cambridge, MA: MIT dissertation.

Featherston, Sam. 2005. The Decathlon Model of empirical syntax. In Kepser, Stephan & Reis, Marga (eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives*, 187–208. Berlin: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110197549.187

Friedmann, Naama & Belletti, Adriana & Rizzi, Luigi. 2009. Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua* 119(1).

67–88. DOI: https://doi.org/10.1016/j.lingua.2008.09.002

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199268511.001.0001

Goldberg, Adele E. 2013. Backgrounded constituents cannot be "extracted." In Sprouse, Jon & Hornstein, Norbert (eds.), *Experimental syntax and island effects*, 221–238. Cambridge, UK: Cambridge University Press.

Goodall, Grant. 2015. The D-linking effect on extraction from islands and non-islands. *Frontiers in Psychology* 5. 1493. DOI: https://doi.org/10.3389/fpsyg.2014.01493

Hernanz Carbó, María Lluïsa. 2012. Sobre la periferia izquierda y el movimiento: El complementante "si" en español [On the left periphery and movement: The complementizer "si" in Spanish]. In Brucart, José María & Gallego, Ángel J. (eds.), *El movimiento de constituyentes* [Constituent movement], 151–171. Madrid: Visor.

Haegeman, Liliane. 1994. *Introduction to Government and Binding Theory*, 2nd ed. Hoboken, NJ: Wiley Blackwell.

Hall, Peter & York, Matthew. 2001. On the calibration of Silverman's test for multimodality. *Statistica Sinica* 11(2). 515–536. URL: https://www.jstor.org/stable/24306875

Hartigan, John A. & Hartigan, Pamela M. 1985. The dip test of unimodality. *The Annals of Statistics* 13(1). 70–84. DOI: https://doi.org/10.1214/aos/1176346577

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

Huang, C.-T. James. 1982. *Logical relations in Chinese and the theory of grammar*. Cambridge, MA: MIT dissertation.

Keshev, Maayan & Meltzer-Asscher, Aya. 2019. A processing-based account of subliminal *wh*-island effects. *Natural Language and Linguistic Theory* 37(2). 621–657. DOI: https://doi.org/10.1007/s11049-018-9416-1

Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality.* Edinburgh: University of Edinburgh dissertation.

Kluender, Robert & Kutas, Marta. 1993. Subjacency as a processing phenomenon.

Language and Cognitive Processes 8(4). 573–633. DOI: https://doi.org/10.1080/01690969308407588

Kobzeva, Anastasia & Sant, Charlotte & Robbins, Parker T. & Vos, Myrte & Lohndal, Terje & Kush, Dave. 2022. Comparing island effects for different dependency types in Norwegian. *Languages* 7(3). 197. DOI: https://doi.org/10.3390/languages7030197

Kush, Dave & Dahl, Anne. 2022. L2 transfer of L1 island-insensitivity: The case of Norwegian. *Second Language Research* 38(2). 315–346. DOI: https://doi.org/10.1177/0267658320956704

Kush, Dave & Lohndal, Terje & Sprouse, Jon. 2018. Investigating variation in island effects: A case study of Norwegian *wh*-extraction. *Natural Language and Linguistic Theory* 36(3). 743–779. DOI: https://doi.org/10.1007/s11049-017-9390-z

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

Kuznetsova, Alexandra & Brockhoff, Per B. & Christensen, Rune H. B. 2017. lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. DOI: https://doi.org/10.18637/jss.v082.i13

López-Sancio, Sergio. 2015. *Testing syntactic islands in Spanish*. Vitoria-Gasteiz: University of the Basque Country master thesis.

Michel, Dan. 2014. *Individual cognitive measures and working memory accounts of syntactic island phenomena*. San Diego, CA: University of California dissertation.

Morey, Richard D. & Rouder, Jeffrey N. & Jamil, Tahira & Urbanek, Simon & Forner, Karl & Ly, Alexander. 2022. BayesFactor: Computation of Bayes Factors for common designs (Version 0.9.12-4.5). URL: https://cran.r-project.org/web//packages//BayesFactor/BayesFactor.pdf

Müller, Dietrich Werner & Sawitzki, Günther. 1991. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association* 86(415). 738–746. DOI: https://doi.org/10.2307/2290406

Müller, Gereon. 2010. On deriving CED effects from the PIC. *Linguistic Inquiry*

41(1). 35–82. DOI: https://doi.org/10.1162/ling.2010.41.1.35

Müller, Gereon. 2021. Constraints on grammatical dependencies. In Allott, Nicholas & Lohndal, Terje & Rey, Georges (eds.), *A companion to Chomsky*, 190–209. Hoboken: Wiley.

Nishigauchi, Taisuke. 1986. *Quantification in syntax*. Amherst: University of Massachusetts dissertation.

Nyvad, Anne Mette & Christensen, Ken Ramshøj & Vikner, Sten. 2017. CP-recursion in Danish: A cP/CP-analysis. *The Linguistic Review* 34(3). 449–477. DOI: https://doi.org/10.1515/tlr-2017-0008

Ortega-Santos, Iván. 2011. On Relativized Minimality, memory and cue-based parsing. *Iberia: An International Journal of Theoretical Linguistics* 3(1). 35–64. URL: https://revistascientificas.us.es/index.php/iberia/article/view/101

Ortega-Santos, Iván. 2020. Dialect distance and data assessment in Chilean, Venezuelan and Puerto Rican Spanish. In Rogers, Brandon M.A. & Figueroa Candia, Mauricio A. (eds.), *Lingüística del castellano chileno / Chilean Spanish Linguistics: Estudios sobre variación, innovación, contacto e identidad / Studies on variation, innovation, contact, and identity*, 151–171. Wilmington, DE: Vernon Press.

Ortega-Santos, Iván & Reglero, Lara & Franco, Jon. 2018. *Wh*-islands in L2 Spanish and L2 English: Between poverty of the stimulus and data assessment. *Fontes Linguae Vasconum* 126. 435–471. URL: https://revistas.navarra.es/index.php/FLV/article/view/1404/14

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

Pañeda, Claudia & Lago, Sol & Vares, Elena & Veríssimo, João & Felser, Claudia. 2020. Island effects in Spanish comprehension. *Glossa: A Journal of General Linguistics* 5(1). 21. DOI: https://doi.org/10.5334/gjgl.1058

Perlmutter, David M. 1968. *Deep and surface structure constraints in syntax*. New York, NY: Holt, Rinehart and Winston.

Pesetsky, David. 1987. *Wh*-in-situ: Movement and unselective binding. In Reuland, Eric J. & ter Meulen, Alice G. B. (eds.), *The representation of (in)definiteness*,

98–129. Cambridge, MA: MIT Press.

Pham, Catherine & Covey, Lauren & Gabriele, Alison & Aldosari, Saad & Fiorentino, Robert. 2020. Investigating the relationship between individual differences and island sensitivity. *Glossa: A Journal of General Linguistics* 5(1). 1–17. DOI: https://doi.org/10.5334/gjgl.1199

Prolific. 2023. London, UK. Retrieved from https://www.prolific.co

Qualtrics. 2023. Provo, UT. Retrieved from https://www.qualtrics.com

R Core Team. 2023. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rackowski, Andrea & Richards, Norvin. 2005. Phase edge and extraction: A Tagalog case study. *Linguistic Inquiry* 36(4). 565–599. DOI: https://doi.org/10.1162/002438905774464368

Reinhart, Tanya. 1981. A second Comp position. In Belletti, Adriana & Brandi, Luciana & Rizzi, Luigi (eds.), *Theory of Markedness in Generative Grammar*, 517–557. Pisa: Scuola Normale Superiore.

Rizzi, Luigi. 1982. Violations of the *wh-* island constraint and the Subjacency condition. In *Issues in Italian syntax*, 49–76. Dordrecht, Netherlands: Foris. DOI: https://doi.org/10.1515/9783110883718.49

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

Rizzi, Luigi. 1997. The fine structure of the left periphery. In Haegeman, Liliane (ed.), *Elements of Grammar. Kluwer International Handbooks of Linguistics*, 281–337. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-011-5420-8_7

Rizzi, Luigi. 2001. On the position "Int (errogative)" in the left periphery of the clause. In Cinque, Guglielmo & Salvi, Giampaolo (eds.), *Current studies in Italian syntax: Essays offered to Lorenzo Renzi* (North-Holland linguistic series 59), 287–296. New York: Brill. https://doi.org/10.1163/9780585473949_016

Rizzi, Luigi. 2004. Locality and left periphery. In Belletti, Adriana (ed.), *Structures and beyond*, 223–251. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780195171976.003.0008

Rizzi, Luigi. 2013. Locality. *Lingua* 130. 169–186. DOI:

https://doi.org/10.1016/J.LINGUA.2012.12.002

Rodríguez, Alejandro & Goodall, Grant. 2020. On the universality of *wh*-islands: Experimental evidence from Spanish. In *50th Linguistic Symposium on Romance Languages*, Austin: University of Texas.

Ross, John Robert. 1967. *Constraints on variables in syntax*. Cambridge, MA: MIT dissertation.

Silverman, Bernard W. 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society* 43(1). 97–99. DOI: https://doi.org/10.1111/j.2517-6161.1981.tb01155.x

Sprouse, Jon. 2007. *A program for experimental syntax: Finding the relationship between acceptability and grammatical knowlege*. College Park, MD: University of Maryland dissertation.

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

Sprouse, Jon & Fukuda, Shin & Ono, Hajime & Kluender, Robert. 2011. Reverse island effects and the backward search for a licensor in multiple *wh*-questions. *Syntax* 14(2). 179–203. DOI: https://doi.org/10.1111/j.1467-9612.2011.00153.x

Sprouse, Jon & Schütze, Carson T. & Almeida, Diogo. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua* 134. 219–248. DOI: https://doi.org/10.1016/j.lingua.2013.07.002

Sprouse, Jon & Villata, Sandra. 2021. Island effects. In Goodall, Grant (ed.), *The Cambridge Handbook of Experimental Syntax*, 227–257. Cambridge, UK: Cambridge University Press. DOI: https://doi.org/10.1017/9781108569620.010

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

Stigliano, Laura & Xiang, Ming. 2021. Experimental evidence on island effects in Spanish relative clauses. *Probus* 33(2). 271–296. DOI: https://doi.org/10.1515/PRBS-2021-0008

Suñer, Margarita. 1991. Indirect questions and the structure of CP: Some consequences. In Campos, Héctor & Martínez-Gil, Fernando (eds.), *Current Studies in Spanish Linguistics*, 283–312. Washington, D.C.: Georgetown University Press.

Szabolcsi, Anna & Lohndal, Terje. 2017. Strong vs. weak islands. In Everaert, Martin & van Riemsdijk, Henk (eds.), *The Wiley Blackwell companion to syntax*, 1–51. Wiley Blackwell. DOI: https://doi.org/10.1002/9781118358733.wbsyncom008

Szabolcsi, Anna & Zwarts, Frans. 1993. Weak islands and an algebraic semantics for scope taking. *Natural language semantics* 1(3). 235–284. URL: https://link.springer.com/article/10.1007/BF00263545

Torrego, Esther. 1984. On inversion in Spanish and some of its effects. *Linguistic Inquiry* 15(1). 103–129. URL: https://www.jstor.org/stable/4178369

Tsai, Wei-Tien Dylan. 1994. On nominal islands and LF extraction in Chinese. *Natural Language and Linguistic Theory* 12(1). 121–175. DOI: https://doi.org/10.1007/BF00992747

Van Riemsdijk, Hank & Williams, Edwin. 1986. *Introduction to the Theory of Grammar*. Cambridge, MA: MIT Press.

Villata, Sandra & Rizzi, Luigi & Franck, Julie. 2016. Intervention effects and Relativized Minimality: New experimental evidence from graded judgments. *Lingua* 179. 76–96. DOI: https://doi.org/10.1016/j.lingua.2016.03.004

Yoshida, Masaya & Kazanina, Nina & Pablos, Leticia & Sturt, Patrick. 2014. On the origin of islands. *Language, Cognition and Neuroscience* 29(7). 761–770. https://doi.org/10.1080/01690965.2013.788196