Kim, Boyoung & Goodall, Grant. 2022. The island/non-island distinction in long-distance extraction: Evidence from L2 acceptability. *Glossa: a journal of general linguistics* 7(1). pp. 1–42. DOI: https://doi.org/10.16995/glossa.5857

Open Library of Humanities

The island/non-island distinction in long-distance extraction: Evidence from L2 acceptability

Boyoung Kim, KAIST, KR, boyoung612@gmail.com
Grant Goodall, University of California, San Diego, US, ggoodall@ucsd.edu

Experimental studies regularly find that extraction out of an embedded clause ("long-distance extraction") results in a substantial degradation in acceptability but that the degradation is much greater when the embedded clause is an island structure. We explore these two facts by means of a series of acceptability experiments with L1 and L2 (L1 Korean) speakers of English. We find that the L2 speakers show greater degradation than L1 speakers for extraction out of non-islands, even though the two groups behave very similarly for extraction out of islands. Moreover, the L2 degradation with non-islands becomes smaller and more L1-like as exposure to the language increases. These initially surprising findings make sense if we assume that speakers must actively construct environments in which extraction out of embedded clauses is possible and that learning how to do this takes time. Evidence for this view comes from cross-linguistic variation in long-distance extraction, long-distance extraction in child English, and lexical restrictions on long-distance extraction. At a broader level, our results suggest that long-distance extraction does not come "for free" once speakers have acquired embedded clauses and extraction.

In experimental studies of sentence acceptability, commonly known as "experimental syntax," perhaps the most robust and widespread result is the degradation associated with long-distance *wh*-extraction. That is, with lexically and structurally matched sentences such as in (1), the presence of a *wh*-dependency that crosses a clause boundary, as in (1b), leads to a significant drop in acceptability relative to a sentence without such a dependency, as in (1a).

(1) a. Who thinks that Mary saw you?
b. Who do you think that Mary saw __ ?

This basic result has been found across many studies and across many languages (e.g., Cowart,1997; Alexopoulou & Keller 2007; Keffala 2011; Sprouse et al. 2012; Rodríguez & Goodall 2020; Fanselow 2021; Goodall 2021).

On the one hand, this result seems very surprising. Both (1a) and (1b) are standardly considered fully grammatical, so to the extent that acceptability is linked to grammaticality, one would not expect the two sentence types to differ significantly. Long-distance extraction does lead to a large drop in acceptability when it is out of an island, of course, but that is not the situation in (1b). In fact, the *that*-clause as in (1b) is the canonical example of an environment in which long-distance extraction is thought to be perfectly well-formed.

On the other hand, it is well known that long-distance *wh*-dependencies, even in non-island environments such as (1b), pose special burdens for the processor (see Fodor 1978; Wanner & Maratsos 1978) and much subsequent work), reflecting the possible activation of processes such as storing the *wh*-phrase filler in working memory, predicting the upcoming gap, retrieving the filler and integrating the filler into the gap position. In reading time studies, activities such as these are inferred from increased reading time at certain crucial regions in the sentence ((Frazier & Clifton 1989; Chen et al. 2005; Gibson 1998; 2000), while in ERP studies, they are suggested by the presence of certain characteristic ERP components (Kluender & Kutas 1993). In acceptability studies, the fact that the dependency induces such a large degradation suggests that these processing difficulties have an effect on acceptability as well, even in the absence of any grammatical violation. This effect does not seem to be detectable through traditional means of collecting acceptability judgments, such as fieldwork or introspection, but it is easily detectable in acceptability experiments and yields very striking results.

The robust degradation associated with long-distance *wh*-dependencies out of non-island environments has thus been abundantly documented and can be reasonably tied, at least in part, to known difficulties that these dependencies pose for the processor. It has also been abundantly documented, however, that a much larger degradation occurs with extraction out of island environments, as in the samples in (2).

(2) a. Who did you wonder [whether Mary saw __]? (*whether*-island)
b. Who did you ask [who saw __]? (*wh*-island)
c. Who did you scream [when Mary saw __]? (adjunct island: *when*)
d. Who did you scream [because Mary saw __]? (adjunct island: *because*)
e. Who did you scream [after Mary saw __]? (adjunct island: *before/after*)

The source of this additional degradation is still a matter of debate, but there is little doubt that there is a significant distinction between extraction out of *that*-clauses, as in (1b), and extraction out of island clauses, as in (2) (see, e.g., Sprouse & Hornstein 2013; Sprouse & Villata 2021).

We aim to shed light on this distinction by examining how it operates in second-language (L2) speakers. We begin by asking whether L2 speakers show the type of degradation that L1 speakers display with long-distance extraction. The answer to this question is not currently known, as there have been relatively few studies of L2 speakers in the experimental syntax tradition, and none that address this question specifically (but see Ortega-Santos et al. 2018).[1] As for processing difficulties that might induce such a degradation, the previous literature offers a mixed picture as to what we might expect. Many studies suggest that long-distance *wh*-dependencies are difficult for L2 speakers in a way that is very similar to what has been found for L1 speakers. These studies (e.g., Dallas & Kaan 2008; Aldwayan et al. 2010; Omaki & Schulz 2011; Pliatsikas & Marinis 2013; Kim et al. 2015; Cunnings 2017; Pliatsikas et al. 2017) suggest that L2 speakers actively search for a gap once they have processed the *wh*-filler, using lexical properties of the verb and structural properties of the clause to detect likely gap positions. These are the types of processes that are assumed to lead to a cost in terms of acceptability for L1 speakers, so if these studies of L2 processing are correct, then we would expect to find the same type of degradation among L2 speakers. In fact, it may even be that long-distance dependencies impose a greater burden in an L2 than they do in an L1, in that some studies find that L2 speakers are less efficient and/or less reliable in their ability to resolve these dependencies (Love et al. 2003; Felser and Roberts 2007; Dallas et al. 2013; Leal et al. 2017; Jessen et al. 2017), so it could be that the degradation among L2 speakers would be even greater than it is among L1 speakers. Other studies, however, argue that L2 speakers do not employ the same mechanisms as L1 speakers when dealing with filler-gap dependencies. In particular, some have argued, based on reading-time measures, that L2 speakers make use of "shallow structure" strategies in which they avoid, at least partially, a full syntactic

[1] Note that there is a long tradition of using the Grammaticality Judgment Task (GJT) in L2 research (for recent overviews, see Kim & Nam 2017 and Plonsky et al. 2020), but it seems unlikely that this method would be able to capture the effect under discussion here, as GJT does not typically use a factorial design, lexical matching, counterbalancing, or the type of response method that is common in the experimental syntax literature (see Ionin 2021 and Kim 2015 for discussion). To our knowledge, no one has attempted to use a traditional GJT to try to detect this effect with either L1 or L2 populations.

parse of the sentence (Marinis et al. 2005; Felser et al. 2012). If this type of account is correct, one might expect to find only a relatively small degradation with long-distance extraction in L2, given that the full range of processing resources used in L1 in these cases would not need to be marshalled here.

We also ask whether L2 speakers show the even greater degradation expected with extraction out of island contexts. Here there is already an extensive existing literature suggesting that L2 speakers do clearly distinguish between islands and non-islands in terms of extraction (Martohardjono 1993; White & Juffs 1998; Belikova & White 2009; Aldwayan et al. 2010; Kim & Goodall 2011; Omaki & Schulz 2011; Kim et al. 2015; Kim & Goodall 2016; Johnson et al. 2016; Goodall 2022), but it is not known what the size of the degradation is, how it compares with the degradation in cases like (1), and how the difference in the amount of degradation between islands and non-islands compares to that observed in L1 speakers.

Examining L2 speakers for these types of degradation turns out to be of interest because of two facts that we will uncover. First, we will see that L2 speakers do show significant degradation in acceptability with regard to extraction out of *that*-clauses. This is in accord with the idea from the literature that L1 and L2 speakers engage in similar mechanisms when processing these structures (and casts doubt on suggestions that L2 speakers do only a shallow parse). In fact, the degradation is significantly larger among L2 speakers than among L1 speakers, and the amount of L2 degradation decreases and approaches the amount of L1 degradation as the length of exposure to the L2 increases. Second, L2 speakers also show a very large degradation with extraction out of islands, but in this case, the size of the degradation in L1 and L2 populations is very similar. Putting these two facts together, we see that L1 and L2 speakers both display degradation with long-distance extraction, in both non-island and island contexts, but with a difference: L1 speakers show a sharp difference in the amount of degradation between non-islands and islands, but L2 speakers do not, at least initially. L2 speakers show L1-like degradation for islands, but they show a similar amount of degradation for non-islands as well.

Put more simply, our results show that L2 speakers initially treat non-island clauses as if they were islands in terms of the size of the degradation associated with long-distance extraction. Only as their length of exposure to the language increases do they gradually begin to treat these clauses as true non-islands and does a clear distinction between island and non-island environments begin to emerge for them. This surprising result suggests that long-distance extraction does not arise immediately once simple extraction and complementation are in place, but rather takes time to develop, a conclusion that receives independent support from what is known about long-distance extraction out of non-islands more generally. All embedded clauses are islands at first, under this view, and it is only through exposure that speakers learn to do long-distance extraction.

We base our conclusions on a series of sentence acceptability experiments with both L1 and L2 speakers. In section 1, we present the results of an experiment that compares sentences with a long-distance dependency out of a *that*-clause to matched sentences without the dependency. Both groups of speakers show a degradation in this case, but it is significantly larger for the L2 speakers. In section 2, we present five experiments that examine extraction out of island environments. Here too, both groups show degradation, but the L2 speakers do not show increased degradation over the L1 speakers here, suggesting that when they do so with *that*-clauses, it is not because they are simply worse than L1 speakers at processing long-distance dependencies in general. If that were the case, we would expect to see an L1 vs. L2 difference for islands like what we see for non-islands. In section 3, we bring the results of all the experiments together and explore the implications of our primary finding, that for L2 speakers, the island/non-island distinction seems to emerge as exposure to the language increases.

1 Experiment 1

As we saw earlier, the fact that long-distance extraction leads to significant degradation has been extensively documented in sentence acceptability experiments with L1 speakers. It is not known, however, if L2 speakers show the same effect. Given the research results discussed above suggesting that L1 and L2 speakers process long-distance dependencies in very similar ways, one would expect that L2 speakers would also exhibit degradation in this case, but if L2 speakers use very different strategies for processing these dependencies, as has also been suggested in the literature, then we would not necessarily expect to find a similar amount of degradation. In our first experiment, we address this issue directly by analyzing participants' responses to long-distance dependencies in comparison to matched sentences without such dependencies in both L1 and L2 populations. This will allow us to determine whether degradation occurs in the two populations and if so, whether there is a difference in size.

1.1 Participants

60 L1 speakers of English and 63 L2 speakers of English (L1 Korean) participated in the experiment. The L2 group was divided into three sub-groups based on Age of Arrival (AoA) in the U.S.: 1–5, 6–10, and 11–14. We take AoA in our study to be a proxy for amount of exposure to the L2, with AoA 1–5 as the group with the largest amount of exposure and AoA 11–14 as the group with the smallest. We do this for two reasons. First, all three groups have very similar mean current ages, so AoA is highly (and negatively) correlated with length of residence. Second, all of the participants in our study are beyond the critical age of 17.4 years, which is when, assuming the findings in Hartshorne et al. (2018), the rate of learning substantially declines.

Even for participants with the same length of residence, then, those in the AoA 1–5 group will have had more years of highly efficient learning than those in the AoA 6–10 group, who in turn will have had more than those in the AoA 11–14 group.

After the experiment session, all participants took an English proficiency test created for this experiment consisting of a multiple-choice vocabulary section (21 questions) and two cloze passages (5 sentences in the first and 3 sentences in the second, with 7 items in each passage). A one-way ANOVA, with proficiency score as a dependent variable and group (i.e. native and the three L2 groups) as a between-subjects factor showed a significant effect of group (F (3, 119) = 12.89, p < .0001). A Pearson correlation analysis between proficiency scores and AoA also revealed a significant negative correlation between the two (r = –.33, N = 63, p = .009, two tails). Further details about the participants are given in **Table 1**.

Group		**L1**	**AoA 1–5**	**AoA 6–10**	**AoA 11–14**
N		60	19	22	22
AoA	Mean (SD) Range		1 year (1.8) 1–5	8 years (1.5) 6–10	12 years (.09) 11–14
Current age	Mean (SD) Range	21 years (2.7) 18–36	20 years (2.1) 18–25	21 years (2.4) 18–30	22 years (4) 18–37
Length of Residence in the U.S.	Mean (SD) Range		19 years (2.8) 14–25	13 years (2.2) 9–20	10 years (4.1) 7–25
Proficiency test scores	Mean (SD) Range	80.8% (4.1) 68.6–88.6	78.3% (5.2) 71.4–85.7	78.2% (5.3) 65.7–82.9	71.7% (9.9) 51.4–82.9

Table 1: Participant Information.

1.2 Materials and methods

All experimental stimuli in this experiment were *wh*-questions with embedded *that*-clauses, but they differed in terms of the length of the *wh*-dependency. In the short-dependency condition, as in (3a), the gap is within the matrix clause, but in the long-dependency condition, as in (3b), the gap is within the embedded clause.

(3) a. Who _ thought [that Lisa bothered Mary]?
    b. Who did Mary think [that Lisa bothered _]?

Crucially, (3a) and (3b) are of essentially equal length and have the same number of clauses, but only (3b) contains a non-vacuous *wh*-dependency. Participants saw 5 tokens of each of these two conditions. Given discussions in the literature (e.g., Cowart 1997; Sprouse & Almeida 2017;

Goodall 2021), 5 tokens per condition would appear to yield more than enough statistical power to detect a linguistically significant effect, if there is one to be found. Five lexically matched sets of stimuli as in (3) were created and distributed into 2 lists using a Latin Square procedure. 2 additional lists were created by reversing the order of stimuli. Each list also contained 95 additional stimuli (50 items from other sub-experiments and 45 fillers) that spanned the full range of acceptability and acted as fillers for this experiment. Each participant thus saw 105 stimuli in total.

Participants were instructed to rate each sentence on a scale from 1 ("bad") to 7 ("good") based on how the sentence sounded to them. They were instructed not to analyze the sentence, but to give their first reaction to it. The stimuli were presented in written form on a computer screen, with the rating scale showing as a set of evenly-spaced radio buttons. Participants had no time limits (see Goodall (2021) for discussion of the speed at which participants typically rate sentences).

1.3 Results

Raw acceptability ratings were transformed to z-scores and are displayed in **Figure 1**, for L1 and L2, and in **Figure 2**, where the L2 results are broken down by AoA group.

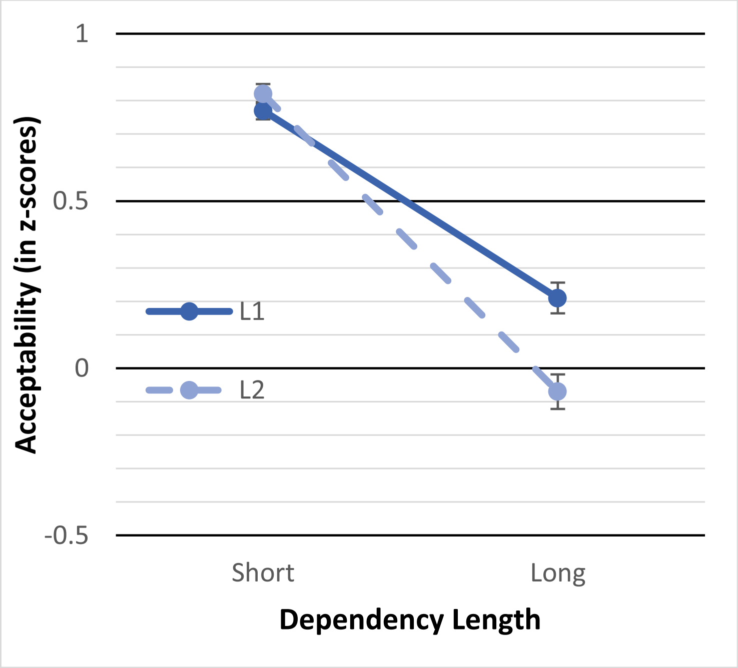

Figure 1: Mean z-scores for L1 and L2 in Experiment 1.

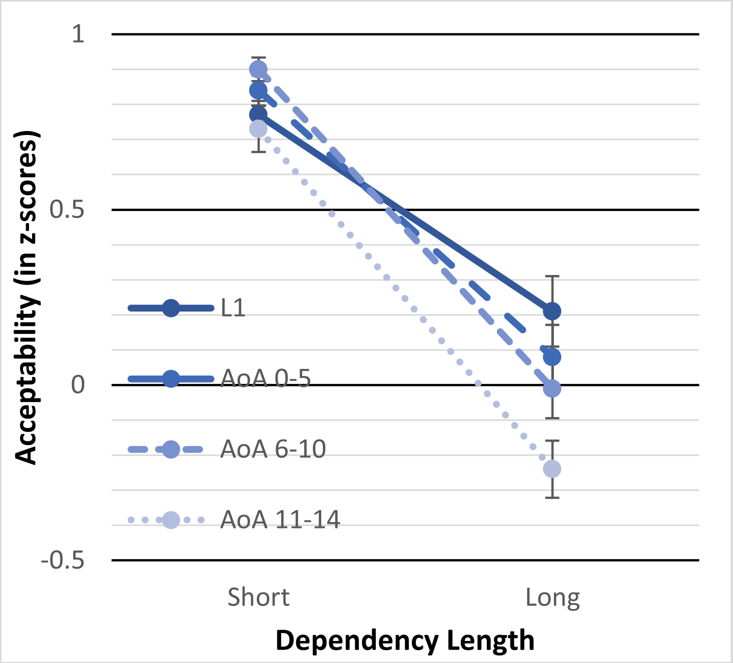

Figure 2: Mean z-scores for L1 and three AoA groups in Experiment 1.

A linear mixed effects model was run, using the lmer function in the lme4 package for R (Bates et al. 2015). All p-values were calculated by Satterthwaite approximation, using the lmerTest package (Kuznetsova et al. 2017). First, to see whether each group shows the expected distance effect, z-scores of native speakers and L2 speakers were entered into a linear mixed effects model, with the fixed factor *dependency distance* (a short dependency, as in (3a), vs. a long dependency, as in (3b)) and the random factors of participants and items as well as the maximal random effect structure. As represented in **Table 2**, the results revealed a significant main effect of *dependency distance* in all groups, in line with the well-established finding for L1 in the previous literature. Notably, this effect is found not just in the L2 speaker group as a whole (as well as the L1 group), but in each of the AoA sub-groups when examined individually. All groups thus show a significant degradation in response to a long-distance *wh*-dependency.

To see whether the distance effect differs in size between L1 and L2, the model was run again with *dependency distance* and *group* (L1 vs. L2) as fixed factors, participants and items as random intercepts, and by-participant and by-item random slopes for *dependency distance* and a by-item random slope for *group*. The model revealed significant main effects of *dependency distance* (Estimate $= -0.067$, SE $= 0.103$, $t = 8.637$, $p < 0.001$) and *group* (Estimate $= 0.273$, SE $= 0.106$, $t = 2.585$, $p = 0.019$), as well as significant interaction of the two factors (Estimate $=$

–0.329, SE = 0.124, *t* = –2.665, *p* = 0.019). To see whether this difference in the size of the distance effect is significant not just for L1 vs. L2, but also for all four groups (L1 and the three AoA sub-groups for L2), one model was run with an interaction between *dependency distance* and *group* (four groups), while a second was run without this interaction. The *anova* function showed a significant difference between these two models ($p < 0.001$), suggesting that the interaction between *dependency distance* and *group* is significant.

	Estimate	**SE**	***t***	***p***
L1 speakers				
distance	0.564	0.143	3.945	0.002
All L2 speakers				
distance	0.892	0.104	8.540	<0.001
AoA 1–5				
distance	0.754	0.155	4.849	<0.001
AoA 6–10				
distance	0.901	0.151	5.959	<0.001
AoA 11–14				
distance	0.984	0.15	6.532	<0.001

Table 2: Linear Mixed Effects results for Experiment 1.

The size of the degradation associated with the presence of a long-distance dependency can be calculated by subtracting the mean of the condition with this dependency, as in (3b), from the mean of the condition without it, as in (3a). The results are displayed in **Figure 3** for L1 and L2. A one-way between-subjects ANOVA with the independent variable of group (L1 and L2) and the dependent variable of the effect size yielded a significant main effect of group (F(1,121) = 18.334, p < .0001). **Figure 4** shows the results broken down by AoA subgroups. A one-way between-subjects ANOVA with the independent variable of group (L1, AoA 1–5, AoA 6–10, AoA 11–14) and the dependent variable of the effect size yielded a significant main effect of group (F(3,119) = 7.04, p < .0001). In addition, a regression analysis showed a significant correlation between AoA and effect size (R^2 = .13, F (1, 121) = 18.056, p < .0001, ß = 0.36, *t* = 4.249) (for L1, AoA was set as 0).

In summary, the results show that the presence of a long-distance *wh*-dependency causes a significant decline in acceptability in all groups but that the size of this decline varies significantly by group, with the decline becoming greater as AoA increases.

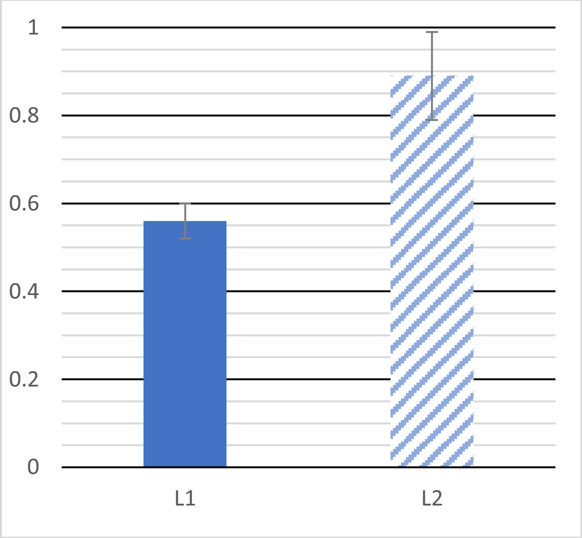

Figure 3: Size of degradation associated with long-distance extraction for L1 and L2.

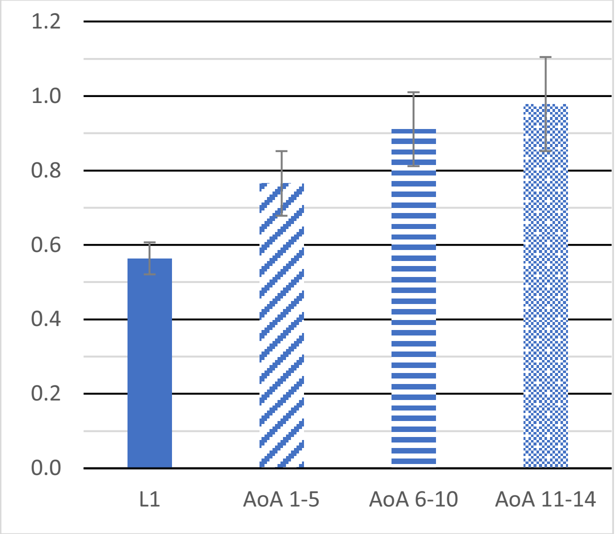

Figure 4: Size of degradation associated with long-distance extraction for L1 and three AoA groups.

1.4 Discussion

The most basic question that Experiment 1 addresses is whether L2 speakers exhibit the degradation associated with long-distance extraction that has been very well attested with L1 speakers in formal sentence acceptability experiments. The results of the experiment show that they do, in that long-distance dependencies are significantly less acceptable than matched sentences with short dependencies for both L1 and L2 participants, as expected. Given this, we can now ask whether this degradation in acceptability is of the same size in the two groups. The answer is clearly negative: The degradation for the L2 speakers is significantly larger than it is for the L1 speakers. This supports the view that L2 speakers encounter some difficulty in performing long-distance dependencies that L1 speakers do not. Moreover, Experiment 1 found a significant correlation between AoA and the size of the decline associated with long-distance extraction. That is, the later the AoA, the more difficulty the speaker has with long-distance dependencies. It thus appears that increased exposure to the language leads to a more native-like level of degradation in acceptability (i.e., less degradation) for long-distance extraction.

Experiment 1 shows us clearly that long-distance filler-gap dependencies induce degradation in acceptability for both L1 and L2 speakers, but that L2 speakers seem to encounter more difficulties than L1 speakers do. It does not tell us much about what these difficulties might consist of, however. For that, we turn to the series of experiments presented in the next section.

2 Experiments 2–6

The results of Experiment 1 suggest that L2 speakers face additional difficulty with long-distance dependencies beyond that experienced by L1 speakers. Broadly speaking, this additional difficulty could have one of two sources. First, it could stem from something about the participants' ability to deal with filler-gap dependencies in general. As we saw in the introduction, it may be that L2 speakers are generally less efficient in the way they resolve these dependencies than L1 speakers, so one might expect this L2 disadvantage to become apparent in the amount of degradation in acceptability. This view might be even more likely since the L2 speakers in Experiment 1 are all L1 speakers of Korean. This language does not use extraction in *wh*-questions, so it could be that *wh*-dependencies still pose a challenge to these speakers, leading to decreased acceptability. Caution is in order, since the L2 participants in the experiment have lived in an English-speaking environment for many years (minimum Length of Residence = 7 years; average Length of Residence = 16 years) and even in their native Korean, filler-gap dependencies are common in other contexts, but still, it could be that the participants' L1 contributes to a generally reduced ability to manage filler-gap dependencies.

Alternatively, it could be that the difficulty experienced by L2 participants in Experiment 1 is not so much about filler-gap dependencies in general, but more about the specific structure tested. That is, these speakers might have the general capacity to manage filler-gap dependencies, but still have difficulty in positing a gap in the specific environment of a complement *that*-clause, which is where the L1/L2 difference was seen.

In Experiments 2-6, we tease apart these two possibilities by testing extraction across a range of other structures. As in Experiment 1, we will compare sentences that have a long-distance dependency to matched sentences that do not. The structures tested will all be island structures. We know that L1 speakers find extraction out of such structures very difficult, whether this is due to effects of processing or grammar (see Phillips 2013 for discussion), and that this difficulty is manifested in acceptability experiments in terms of a very large degradation (significantly larger than the degradation associated with extraction out of a *that*-clause; see Sprouse et al. 2012; Sprouse & Villata 2021). For L2 speakers, a degradation of this type has not been documented to the same extent, but we expect it to occur, given the findings in the literature that L2 speakers appear to be sensitive to island constraints (Martohardjono 1993; White & Juffs 1998; Belikova & White 2009; Aldwayan et al. 2010; Omaki & Schulz 2011; Kim et al. 2015; Johnson et al. 2016; Kim & Goodall 2016; Boxell & Felser 2017).

The main questions that we address in these experiments concern the size of the degradation and any differences in that regard between the L1 and L2 participants. If the L2 speakers have a reduced ability (relative to the L1 speakers) to handle filler-gap dependencies in general, we then expect to see the same effect here that we did in Experiment 1. That is, the L2 participants should show a significantly larger degradation in acceptability than the L1 participants when faced with extraction out of these island structures. This is because L2 speakers would have the same difficulty with extraction out of islands that L1 speakers do, but this would be in addition to their baseline difficulty with long-distance extraction in general. We should also expect to find a correlation between AoA and degree of degradation, as we did in Experiment 1, where we saw that increased exposure to the language seemed to lead to increased ability to handle long-distance extraction. As their baseline ability to do long-distance extraction improves, we would expect to see a decrease in the amount of degradation for extraction out of islands, until it approaches that of L1 speakers.

If, on the other hand, the difficulty for L2 participants in Experiment 1 has to do with the specifics of extraction out of complement *that*-clauses and does not stem from some generalized difficulty with long-distance extraction, then we should expect that in these new experiments, L1 and L2 speakers would have about equal amounts of degradation, and for the L2 speakers, there should be no correlation between AoA and the size of the degradation.

2.1 Participants

All the experiments reported here were administered as sub-experiments within a single larger experiment, so the participants in Experiments 2–6 were the same as in Experiment 1.

2.2 Materials and methods

In each of Experiments 2–6, there were two conditions. As in Experiment 1, both were biclausal *wh*-questions, with either a short dependency (i.e., gap in the matrix clause) or a long dependency (i.e., gap in the embedded clause). Unlike in Experiment 1, the embedded clauses in Experiments 2–6 were all islands, so at least for L1 participants, we expect the long dependency condition to be substantially degraded. (4) – (8) show the type of island structure used in each experiment and give sample stimuli of the two conditions in each.

(4) Experiment 2: *whether*-island
  a. Who __ wondered [whether Lisa bothered Mary]?
  b. Who did Mary wonder [whether Lisa bothered __]?

(5) Experiment 3: *wh*-island (*who*)
  a. Who __ heard [who bothered Mary]?
  b. Who did Mary hear [who bothered __]?

(6) Experiment 4: adjunct island with a *when*-clause
  a. Who __ screamed [when Lisa bothered Mary]?
  b. Who did Mary scream [when Lisa bothered __]?

(7) Experiment 5: adjunct island with a *because*-clause
  a. Who __ screamed [because Lisa bothered Mary]?
  b. Who did Mary scream [because Lisa bothered __]?

(8) Experiment 6: adjunct island with a clause headed by *before* or *after*
  a. Who __ screamed [after Lisa bothered Mary]?
  b. Who did Mary scream [after Lisa bothered __]?

Stimuli for these conditions were created and distributed into lists in the same way as in Experiment 1. The task for the participants was also the same, since all six experiments were administered as sub-experiments within a single, larger experiment.

2.3 Results

Results were transformed to z-scores; means for the L1 and L2 groups are presented in **Figures 5–9**.

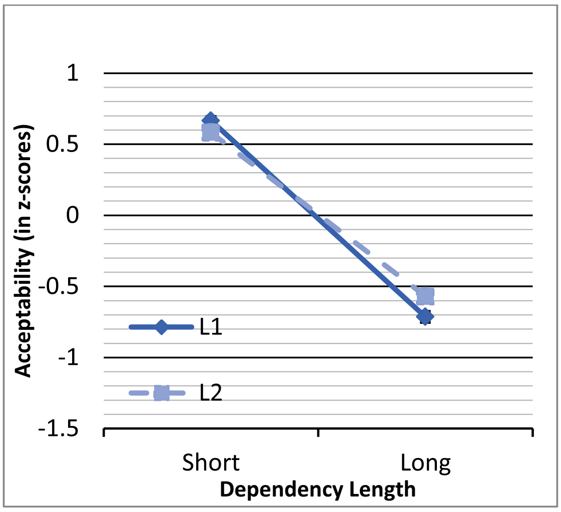

Figure 5: Mean z-scores for L1 and L2 in Experiment 2 (*whether*-island).

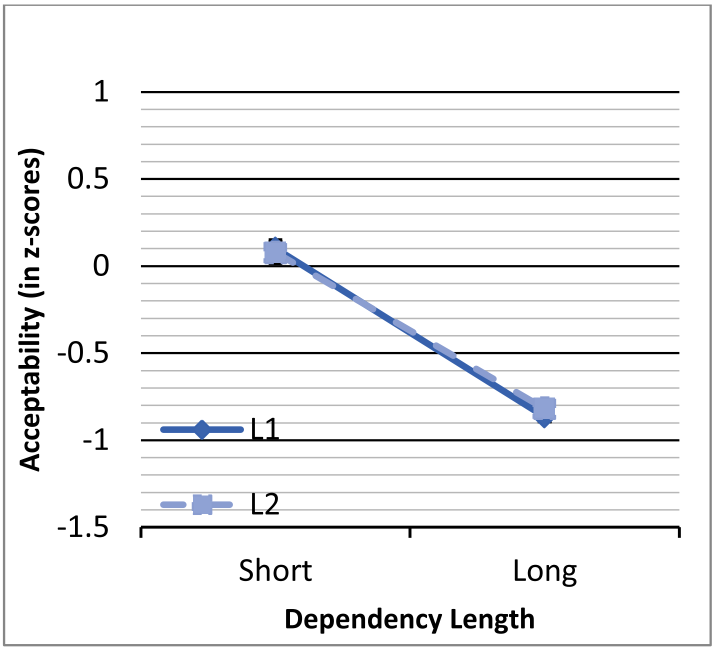

Figure 6: Mean z-scores for L1 and L2 in Experiment 3 (*wh*-island).

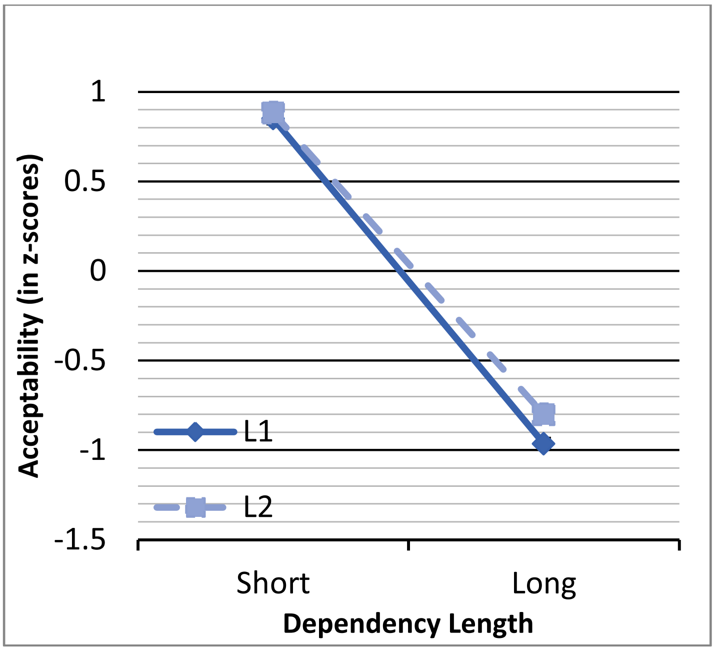

Figure 7: Mean z-scores for L1 and L2 in Experiment 4 (adjunct island: *when*).

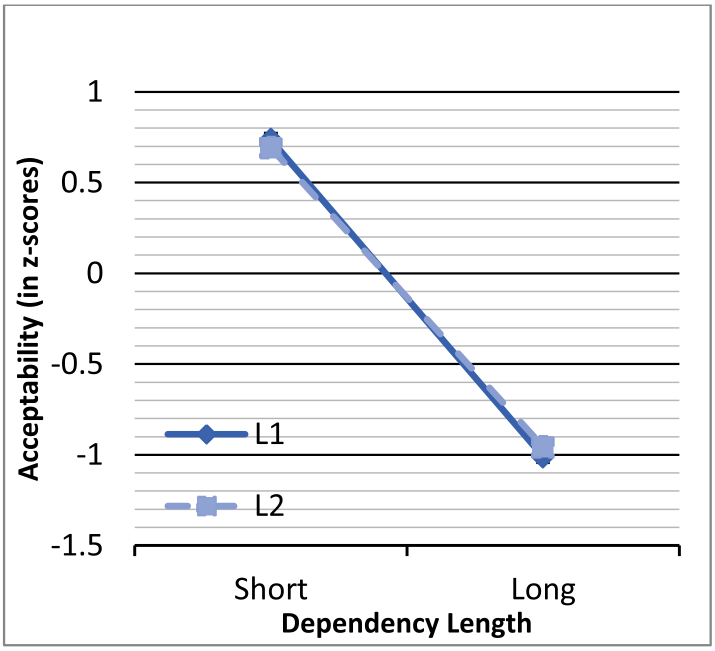

Figure 8: Mean z-scores for L1 and L2 in Experiment 5 (adjunct island: *where*).

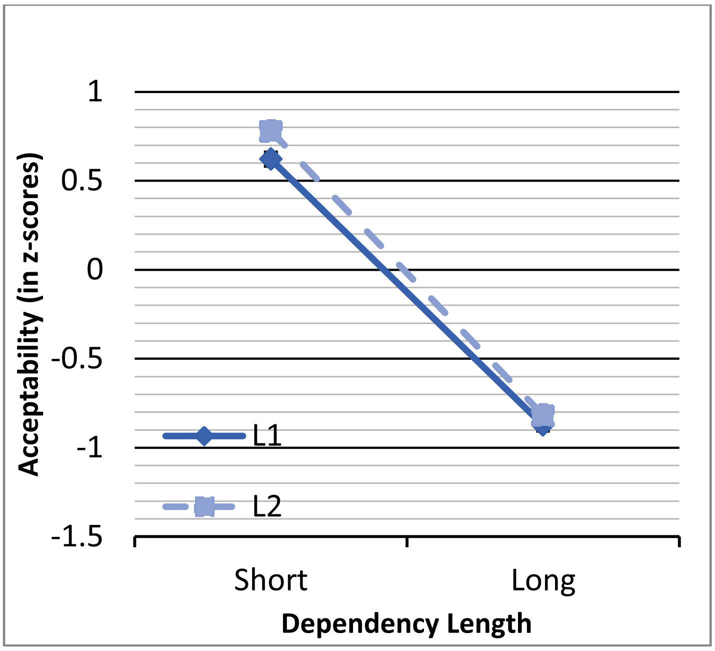

Figure 9: Mean z-scores for L1 and L2 in Experiment 6 (adjunct island: *before/after*).

Statistical analyses were run as in Exp. 1 and the results are summarized in **Tables 3–7**. As expected, the presence of a long-distance dependency resulted in a significant decline in acceptability in all five experiments for both the L1 group and the L2 group, as well as when the L2 group is broken down into the three AoA groups.

	Estimate	**SE**	***t***	***P***
L1 speakers				
distance	1.380	0.067	20.57	< 0.001
All L2 speakers				
distance	1.156	0.083	13.809	< 0.001
AoA 1–5				
distance	1.270	0.139	9.084	< 0.001
AoA 6–10				
distance	1.286	0.107	11.925	< 0.001
AoA 11–14				
distance	0.942	0.103	9.074	< 0.001

Table 3: Linear Mixed Effects results for Experiment 2 (*whether*-island).

	Estimate	**SE**	***t***	***p***
L1 speakers				
distance	0.965	0.107	8.951	< 0.001
All L2 speakers				
distance	0.894	0.071	12.56	< 0.001
AoA 1–5				
distance	0.918	0.106	8.604	< 0.001
AoA 6–10				
distance	1.050	0.099	10.59	< 0.001
AoA 11–14				
distance	0.734	0.118	6.193	< 0.001

Table 4: Linear Mixed Effects results for Experiment 3 (*who*-island).

	Estimate	**SE**	*t*	*p*
L1 speakers				
distance	1.81542	0.04153	43.71	< 0.001
All L2 speakers				
distance	1.67713	0.09220	18.190	< 0.001
AoA 1–5				
distance	1.76960	0.10929	16.19	< 0.001
AoA 6–10				
distance	1.82447	0.10327	17.667	< 0.001
AoA 11–14				
distance	1.4590	0.1709	8.537	< 0.001

Table 5: Linear Mixed Effects results for Experiment 4 (*when*-island).

	Estimate	**SE**	***t***	***p***
L1 speakers				
distance	1.747	0.065	26.59	< 0.001
All L2 speakers				
distance	1.65	0.076	21.50	< 0.001
AoA 1–5				
distance	1.637	0.115	14.24	< 0.001
AoA 6–10				
distance	1.844	0.100	18.41	< 0.001
AoA 11–14				
distance	1.474	0.116	12.70	< 0.001

Table 6: Linear Mixed Effects results for Experiment 5 (*because*-island).

To see whether the decline in acceptability varies between L1 and L2, we ran the model again with *dependency distance* and *group* (L1 and L2) as fixed factors, participants and items as random intercepts, and by-participant and by-item random slopes for dependency distance and a by-item random slope for group. As shown in **Table 8**, the main effect of distance was significant in all experiments, but the main effect of *group* and its interaction with *dependency distance* were not significant, except in Experiment 2 (*whether*-island). In that experiment, the effect for *dependency distance* was smaller for L2 than for L1 and the interaction between *dependency distance* and *group* was significant.

	Estimate	SE	*t*	*p*
L1 speakers				
distance	1.494	0.156	9.521	< 0.001
All L2 speakers				
distance	1.588	0.130	12.216	< 0.001
AoA 1–5				
distance	1.62	0.133	12.127	< 0.001
AoA 6–10				
distance	1.705	0.170	9.993	< 0.001
AoA 11–14				
distance	1.452	0.178	8.132	< 0.001

Table 7: Linear Mixed Effects results for Experiment 6 (*before/after*-island).

	Estimate	SE	*t*	*p*
Experiment 2				
distance	1.156	0.084	13.718	< 0.001
group	–0.14	0.072	–1.947	0.065
distance*group	0.223	0.082	2.712	0.008
Experiment 3				
distance	0.896	0.071	12.486	< 0.001
group	–0.048	0.063	–0.765	0.460
distance*group	0.069	0.103	0.667	0.515
Experiment 4				
distance	1.676	0.089	18.811	< 0.001
group	–0.169	0.081	–2.09	0.058
distance*group	0.139	0.087	1.591	0.128
Experiment 5				
distance	1.651	0.069	23.844	< 0.001
group	–0.054	0.053	–1.004	0.321
distance*group	0.095	0.0825	1.163	0.247
Experiment 6				
distance	1.59	0.127	12.499	< 0.001
group	–0.06367	0.07002	–0.909	0.376
distance*group	–0.09574	0.12453	–0.769	0.454

Table 8: Linear Mixed Effects results for L1 and L2 groups together in Experiments 2–6.

Figures 10–14 show the results with the L2 participants divided into the AoA groups.

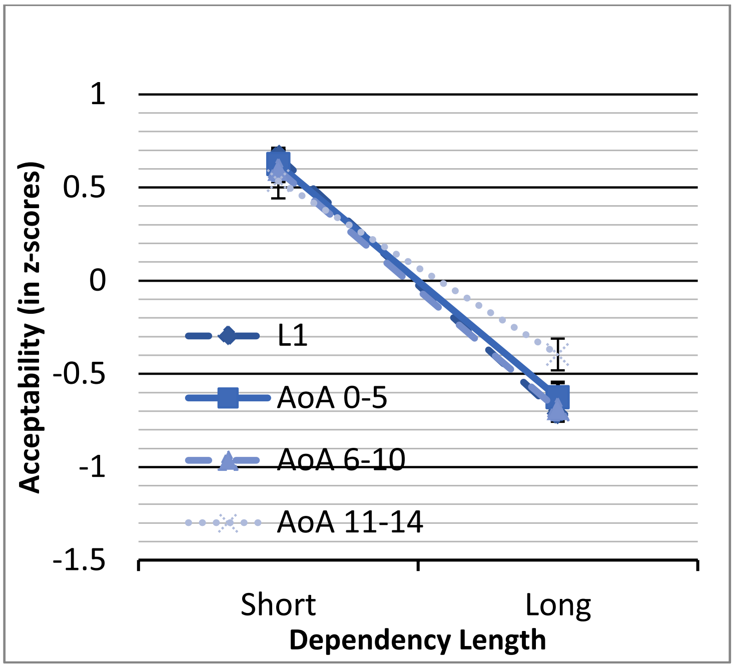

Figure 10: Mean z-scores for L1 and three AoA groups in Experiment 2 (*whether*-island).

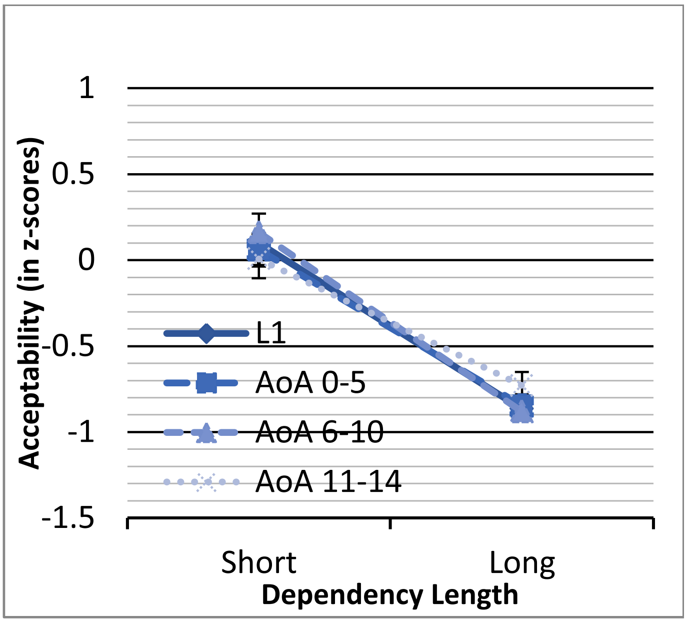

Figure 11: Mean z-scores for L1 and three AoA groups in Experiment 3 (*wh*-island).

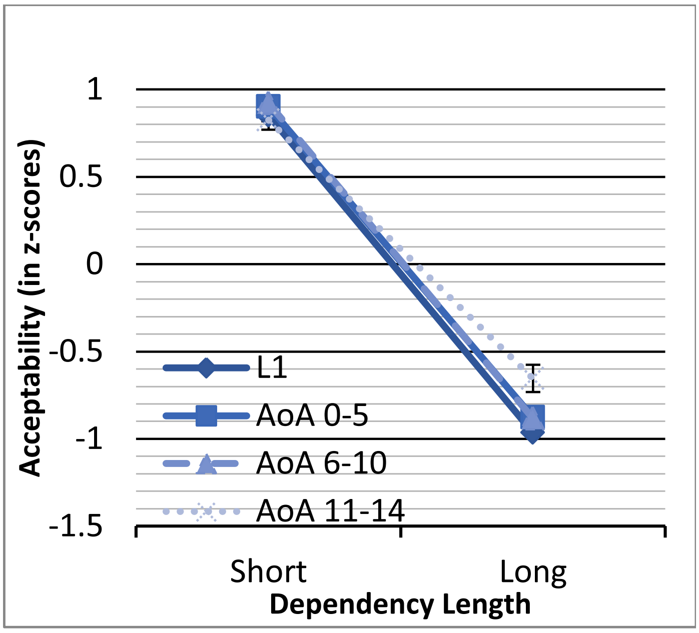

Figure 12: Mean z-scores for L1 and three AoA groups in Experiment 4 (adjunct island: *when*).

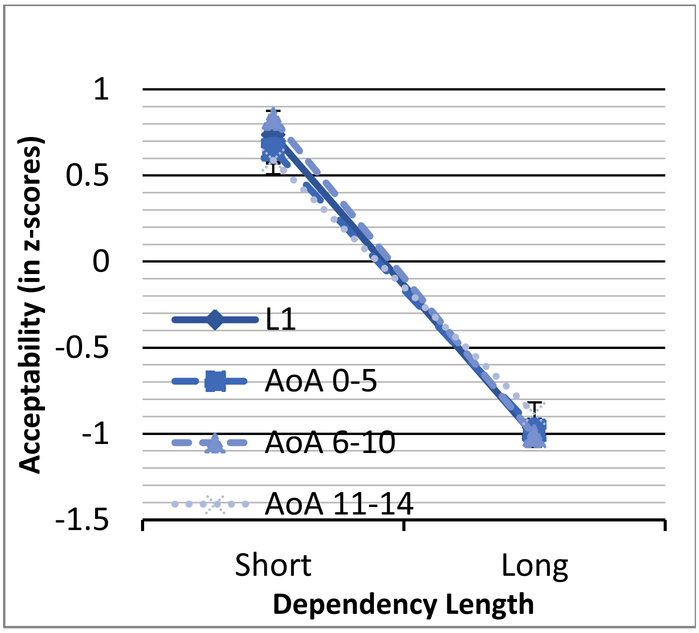

Figure 13: Mean z-scores for L1 and three AoA groups in Experiment 5 (adjunct island: *where*).

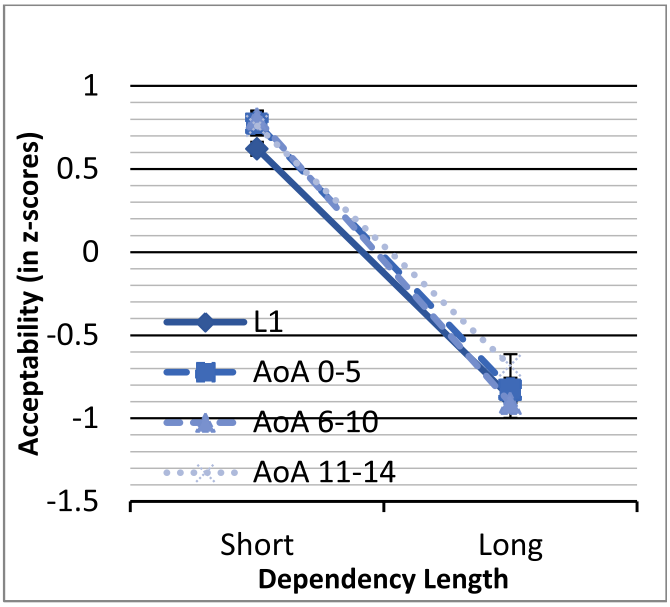

Figure 14: Mean z-scores for L1 and three AoA groups in Experiment 6 (adjunct island: *before/after*).

To see whether the decline in acceptability varies across all four groups (L1 and the three AoA sub-groups for L2), one model was run with an interaction between *dependency distance* and *group* (four groups) and another without this interaction. This procedure was done for each experiment separately. The *anova* function showed a significant difference between the two models in Experiments 2, 4, and 5 (Experiments 2, and 4: $p = 0.001$; Experiment 5: $p = 0.039$), but not in Experiments 3 and 6 (Experiment 3: $p = 0.124$; Experiment 6: $p = 0.053$), suggesting a significant interaction between *dependency distance* and *group* in the former cases, but not in the latter ones. As we will see, however, even in the cases where there is an interaction, there is no pattern of the degradation increasing as AoA increases in the manner of the results from Exp. 1.

As in Exp. 1, the size of the degradation induced by the presence of a long-distance dependency may be calculated by subtracting the mean of the long-distance dependency condition from that of the matching condition without this dependency. These effect sizes are shown in **Figures 15–19** for the L1 and L2 groups in each experiment.

One-way ANOVAs with the between-subjects variable of group (L1 and L2) and the dependent variable of the effect size show that the size of the degradation is significantly larger for L1 than for L2 in Experiment 2 (F (1, 121) = 8.17, $p = .005$), but that there is no significant difference between L1 and L2 in Experiments 3–6 (Exp. 3: F (1, 121) = .773, $p = .381$; Exp. 4: F (1, 121) = 3.616, $p = .06$; Exp. 5: F (1, 121) = 1.451, $p = .231$; Exp. 6: F (1, 121) = 1.16, $p = .284$). In none of the cases examined here is the degradation significantly larger for L2 than for L1.

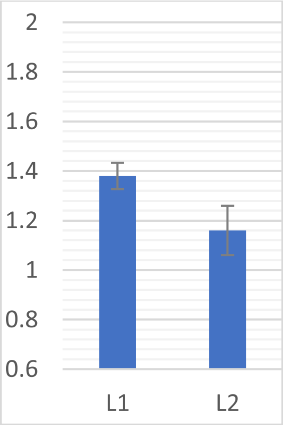

Figure 15: Size of degradation with *whether*-islands in Exp. 2.

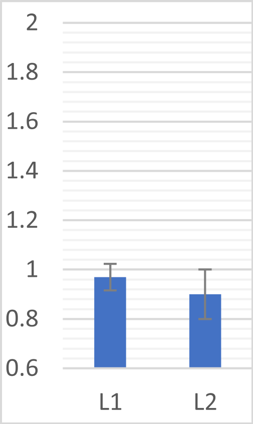

Figure 16: Size of degradation with *who*-islands in Exp. 3.

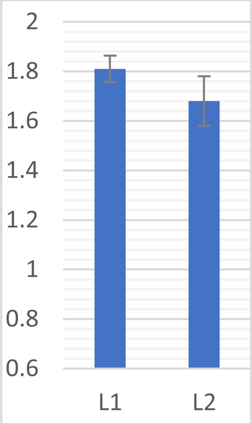

Figure 17: Size of degradation with adjunct islands (*when*) in Exp. 4.

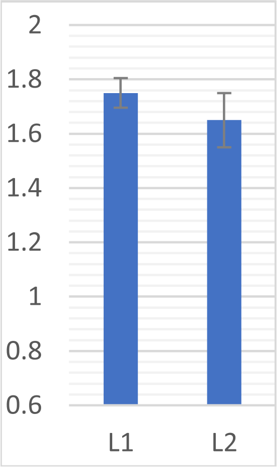

Figure 18: Size of degradation with adjunct islands (*where*) in Exp. 5.

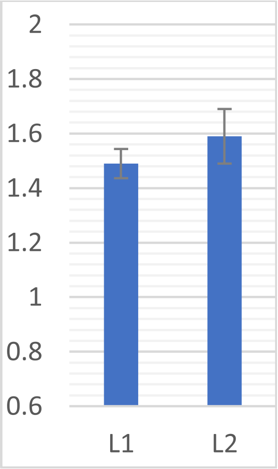

Figure 19: Size of degradation with adjunct islands (*before/after*) in Exp. 6.

A more fine-grained view of the degradation associated with the long-distance dependency is shown in **Figure 20**, where the L2 group is broken down into three AoA subgroups.

One-way between-participants ANOVAs with the independent factor of group (L1, AoA 1–5, AoA 6–10, AoA 11–14) and the dependent variable of the size of the degradation show a significant difference in the size of degradation between four groups in Experiments 2, 4 and 5 (Exp. 2:F (1, 119) = 5.796, *p* = .001; Exp. 4: F (1, 119) = 4.677, *p* = .004; Exp. 5: F (1, 119) = 3.003, *p* = .033), and no such difference in Experiments 3 and 6 (Exp. 3:F (1, 119) = .194, *p* = .114; Exp. 6: F (1, 119) = 1.651, *p* = .181)). Post-hoc Bonferroni tests show that the difference was mainly significant between L1 and AoA 11–14 and between AoA 1–5 and

AoA 11–14 (p <.05). A regression analysis shows a significant correlation between the size of the degradation and AoA in Experiment 2 (R^2 = .087, F (1, 121) = 11.574, p =.001, ß = –.295, t = – 3.302) and Experiment 4 (R^2 = .089, F (1, 121) = 11.797, p =.001, ß = –.298, t = – 3.435), but not in Experiment 3 (R^2 = .016, F (1, 121) = 1.909, p =.17, ß = –.125, t = – 1.382), Experiment 5 (R^2 = .021, F (1, 121) = 2.588, p =.11, ß = –.145, t = – 1.609), and Experiment 6 (R^2 < .0001, F (1, 121) = .001, p =.977, ß = .003, t = .029). In none of these experiments, however, does the size of degradation consistently increase as AoA increases, unlike what we saw in Exp. 1.

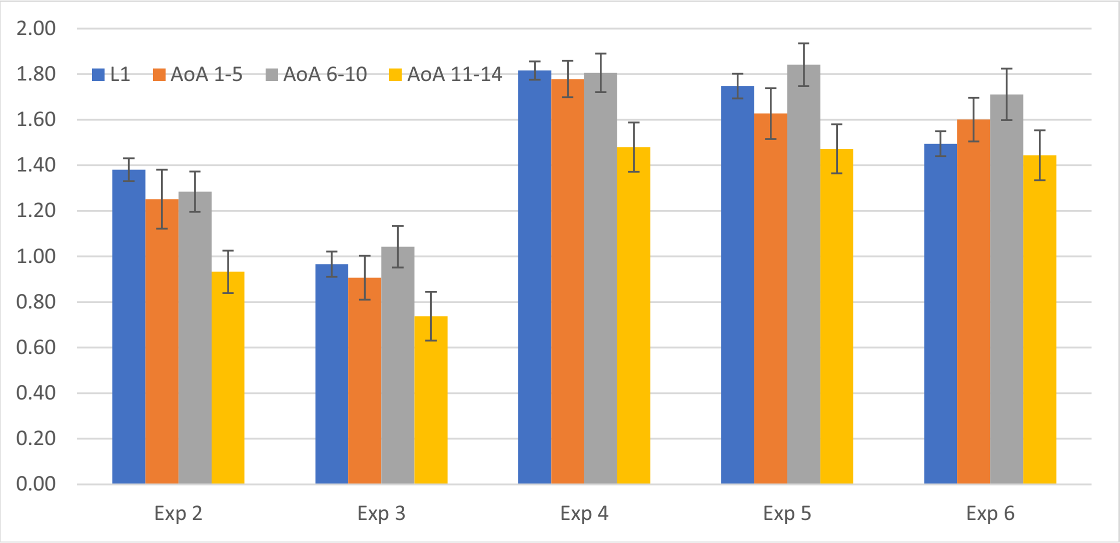

Figure 20: Size of degradation associated with long-distance extraction for L1 and three AoA groups in Exp. 2–6.

Unlike what we saw in Exp. 1, then, in Exp. 2–6 the degradation induced by a long-distance dependency is not consistently larger for L2 speakers than it is for L1. One possible explanation for the lack of greater degradation for L2 is that there could be a floor effect here. That is, perhaps island violations are so low in acceptability that the L2 participants did not have enough room at the bottom of the scale to register the true size of the degradation. Though initially plausible, this explanation is not supported by the results. **Figures 21–24** show the island-violating condition (i.e., the long dependency condition) in relation to all of the filler items in Experiment 6 (used here as a sample; Experiments 2–5 are very similar). For all four sub-groups, the island-violating condition is distant from the worst of the fillers. This suggests there is no floor effect artificially raising the mean for this condition; if participants had felt that acceptability was lower, they could have given it a lower score.

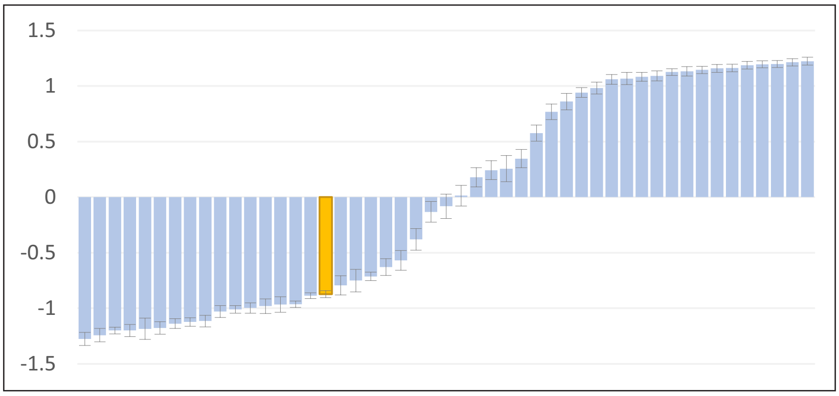

Figure 21: Mean acceptability of adjunct island violation (in gold) in relation to filler items in Exp. 6: L1 participants.

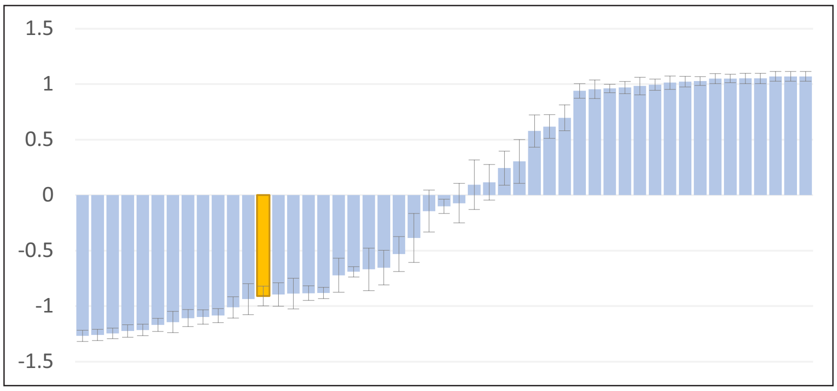

Figure 22: Mean acceptability of adjunct island violation (in gold) in relation to filler items in Exp. 6: AoA 0–5 participants.

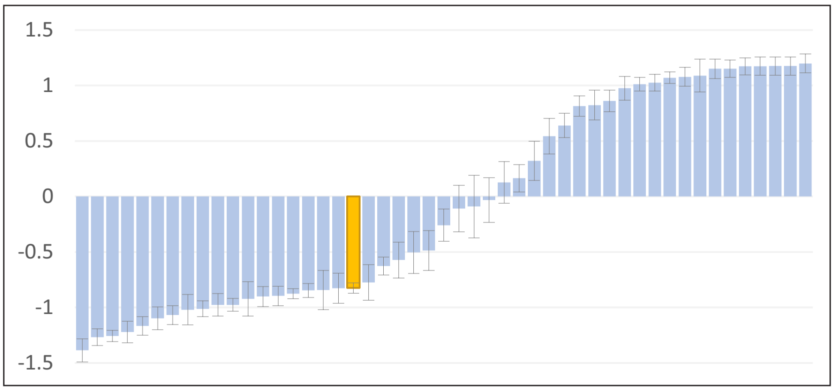

Figure 23: Mean acceptability of adjunct island violation (in gold) in relation to filler items in Exp. 6: AoA 6–10 participants.

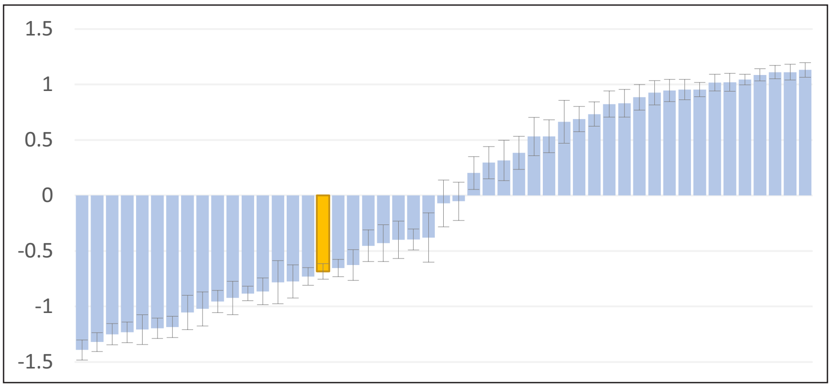

Figure 24: Mean acceptability of adjunct island violation (in gold) in relation to filler items in Exp. 6: AoA 11–14 participants.

2.4 Discussion

In general, then, for the structures examined in Experiments 2–6, the L1 and L2 participants behaved in a remarkably similar fashion. Both showed significant declines in acceptability when extraction occurred out of any of the five island structures tested, and the degradation for the L2 group was never significantly larger than for the L1 group. In addition, there was no consistent increase in the size of the degradation as AoA increased. Among other things, all of this suggests that L1 and L2 speakers are generally very similar in the way that they process long-distance dependencies. The structures tested here all require speakers to deploy fine-grained grammatical knowledge, as well as storage, retrieval, and integration of the filler at the gap site, and we see no sign here that L2 speakers perform this task in a substantially different way than L1 speakers.

3 General discussion

In all six of the experiments presented above, we measured the acceptability of long-distance extraction across a range of structures for both L1 and L2 speakers. For all speakers and for all structures, we found that the long-distance dependency triggered a significant degradation in acceptability, replicating what has been found in many other studies for L1, but documenting this for the first time for L2. What is particularly interesting, though, is the difference in results between Experiment 1, on the one hand, and Experiments 2–6 on the other. In Experiment 1, we tested extraction out of complement *that*-clauses, a case that is standardly assumed to be grammatically licit. Both L1 and L2 speakers showed the expected degradation when a long-distance extraction was present, relative to matched sentences without such a dependency, but the degradation was significantly larger for the L2 speakers. Moreover, there was a significant correlation between the size of the degradation and AoA: as the latter increased, so did the former. In Experiments 2–6, we tested extraction out of island structures, i.e., embedded clauses that are known to result in a much larger degradation for extraction than the *that*-clauses that were tested in Experiment 1, reflecting the standard assumption that gaps in these structures are unacceptable to some degree. In these cases, the L1 and L2 speakers largely tracked together, and there was not the same correlation between the size of the degradation and AoA.

Put simply, the L1 speakers showed a sharp difference between *that*-clauses (in Experiment 1) and islands (in Experiments 2–6) with regard to the size of the degradation induced by the long-distance dependency, but L2 speakers did not. For L1 speakers, extraction from *that*-clauses resulted in an average degradation of .56, in z-score units, while the smallest counterpart among the islands (*wh*-islands) was .97, with the other islands ranging up to 1.81 (adjunct islands with *when*), as seen in **Figure 4** and **Figure 20**. For L2 speakers, on the other hand, extraction from *that*-clauses caused a degradation of .89, statistically indistinguishable from the degradation of .90 for *wh*-islands. As we have seen, the size of the degradation with *that*-clauses is correlated with AoA, so as AoA decreases, extraction from *that*-clauses becomes more and more distinct from extraction from islands. Given that in our study, it is reasonable to take AoA as a proxy for amount of exposure to the language (see the discussion in the introduction), it appears that as the amount of exposure to the language increases, extraction from *that*-clauses becomes more native-like and more distinct from extraction from islands.

Descriptively, then, the L2 speakers treat complement *that*-clauses as islands, in the sense that extraction from them causes a degradation in acceptability on a par with the classical islands. As L2 speakers accumulate more years of exposure to the language, though, they treat *that*-clauses less like islands and more in line with the behavior of L1 speakers.

Why would L2 speakers do this? That is, why would they treat complement *that*-clauses as islands, but then gradually stop doing this over time? We cannot give a definitive answer here, but we can point to ways in which this result is not as surprising as it might seem at first glance. In what follows, we will see facts from three domains, cross-linguistic variation in long-distance extraction, long-distance extraction in child English, and lexical restrictions on long-distance extraction, all of which suggest that long-distance extraction is not simply the default option once speakers have acquired embedded clauses and *wh*-extraction. Rather, speakers must actively construct environments in which extraction out of embedded clauses is possible, and learning how to do this takes time (see Culicover & Jackendoff 2005; Pearl & Sprouse 2013 for additional discussion on this point).

3.1 Cross-linguistic variation in long-distance extraction

Given a familiarity with English, it is easy to have the impression that any language which permits complement clauses and extraction of *wh*-phrases will also allow extraction out of complement clauses, resulting in long-distance extraction. This is not the case, however, as many languages do not readily allow long-distance extraction even though they appear to have all the needed components for doing so.

Russian is one such language (see Ross 1967 and much subsequent discussion), in that extraction out of indicative complement clauses with an overt complementizer is very degraded. Russian allows (and in fact, generally requires) extraction of *wh*-phrases, as seen in (9), and it allows indicative complement clauses, as seen in (10) (data from Khomitsevich 2008).

(9) Kogo Maša ljubit _ ?
who Masha loves
'Who does Masha love _ ?'

(10) Ivan znaet, čto Maša ljubit Petra.
Ivan knows that Masha loves Peter
'Ivan knows that Masha loves Peter.'

When we try to do extraction as in (9) out of an embedded clause as in (10), however, speakers report a very substantial decline in acceptability, as seen in (11) (from Khomitsevich 2008).

(11) ?* Kogo ty dumaeš, čto Ivan priglasil _ ?
who you think that Ivan invited
'Who do you think that Ivan invited _ ?'

There is also a decline in these circumstances in English, as we have observed, but the decline in Russian appears to be much greater, as many speakers find sentences such as (11) very unacceptable, and the difference between (10) and (11) can be easily detected by individual speakers even without the use of a formal acceptability experiment.

Cases such as this suggest very strongly that for L1 acquisition, at least, children need some direct input to know that long-distance extraction is possible. Simply having evidence that extraction is possible, as in (9), and that complement clauses are possible, as in (10), is not enough, since if it were, speakers would accept sentences like (11). Instead, they seem to need some positive evidence that long-distance extraction is possible in environments of this type. Children exposed to English get this evidence, but those exposed to Russian apparently do not.

In addition, many languages that allow long-distance extraction nonetheless offer ways to avoid it. "Partial *wh*-movement," for instance, occurs in many languages as an alternative to long-distance *wh*-extraction (Fanselow 2006). The *wh*-phrase appears at the left edge of the embedded clause, even though it is understood as having scope over the entire sentence. In Indonesian (Saddy 1991), for instance, long-distance extraction is possible, as in (a), but it is also possible to have partial extraction, as in (b) (gap sites, including the one necessitated by successive cyclicity, are indicated by "_").

(12) a. Siapa yang Bill tahu [_ Tom cintai _]
who FOC Bill knows Tom loves

b. Bill tahu [siapa yang Tom cintai _]
Bill knows who FOC Tom loves

As has often been pointed out (see Fanselow 2006 and references cited there), structures such as (12b) seem odd at first, because the *wh*-phrase appears neither in its argument position (as object of *cintai* 'love'), nor in its scope position (as it does in (12a)). The fact that so many languages allow the option of (12b), though, suggests that (12a) is not as ideal a structure as it might seem, and the fact that speakers avail themselves of (12b) even though (12a) is also available suggests something similar at the level of production.

"Partial *wh*-movement" languages thus show that even when long-distance extraction is allowed, the language may also provide a way to avoid it and in that case, speakers makes use of this option. Children exposed to these languages do hear direct positive evidence for the existence of long-distance extraction, but they nonetheless still sometimes refrain from using it.

Further evidence that children are cautious in adopting long-distance extraction comes from languages that give the appearance of allowing long-distance extraction, but actually do not. Sundanese, another Austronesian language of Indonesia, is one such example (Davies & Kurniawan 2013). The *wh*-question in (13), for instance, looks like a typical long-distance filler-gap structure.

(13) Naon nu di-sangka ku Ahmad [(nu) di-sumput-keun ku Dédén]?
what rel pv-suspect by Ahmad rel pv-hide-keun by Deden
'What did Ahmad suspect that Deden hid?'

There is a *wh*-word at the left edge of the sentence, *naon* 'what', that is understood as an argument of the main verb of the embedded clause, *sumput* 'hide. Davies & Kurniawan (2013) show that treating this as a filler-gap (A') structure would seem to be incorrect, though, since several characteristic properties of filler-gap dependencies are not present. One such property is "reconstruction," in which the filler behaves as if it occupied the gap position. This property may be seen, for example, by embedding a reflexive inside the *wh*-phrase filler. One would expect such a configuration to be ill-formed, since reflexives must be in a specific structural relationship with an antecedent that would not be possible in this case, but it is nonetheless possible. A sample from English is given in (14).

(14) [Which picture of herself] do you think [that Mary has purchased _] ?

The reflexive *herself* here is able to take *Mary* as its antecedent only because of reconstruction, which allows it to be treated as if it were in the gap position. That reconstruction is specifically a property of filler-gap structures may be seen in (15), which is much less acceptable than (14).

(15) ?*[This picture of herself] seems to have been purchased by Mary.

(14) and (15) are similar in that in both, *herself* is in a phrase that precedes *Mary*, but only in (14) is *herself* within the filler in a filler-gap structure. The crucial piece of evidence regarding Sundanese, then, is that the equivalent of (14) is not acceptable:

(16) *Gambar dirina sorangan nu mana nu di-sangka nu di-tingal-i ku Asép?
picture self rel which rel pv-suspect rel pv-look-i by Asep
'Which picture of self$_i$ is it suspected that Asep$_i$ looked at?

In (16), *dirina sorangan* 'self' is within the *wh*-phrase, so if this were a standard filler-gap (A') structure, we would expect it to be able to take *Asép* as its antecedent.

Because of evidence of this type, Davies and Kurniawan conclude that *wh*-questions as in (13) in Sundanese do not involve the same mechanism that produces filler-gap structures as in English. If correct, it means that when children encounter *wh*-questions as in (13), they do not opt for what might seem to be the most straightforward analysis, involving long-distance extraction of the *wh*-phrase and a filler-gap structure. Instead, they take a more conservative route, finding a way to generate the string in (13) using only mechanisms that are independently needed (as Davies and Kurniawan show is possible, though space prevents discussion here). This is important for our present concerns, because it suggests that children adopt long-distance extraction only if there is no alternative. In Sundanese, there is an alternative, so the mechanism of long-distance extraction goes unused, resulting in the lack of reconstruction effects, among other things, as in (16).

Together, the facts discussed in this section suggest that learners do not assume that long-distance extraction is possible, and even if they discover that it is possible, speakers will sometimes avoid it if the language makes that option available. For L2 acquisition, then, it makes sense, especially for learners whose L1 does not make use of long-distance *wh*-extraction, to be very cautious about concluding that long-distance extraction is possible in the new language. They would presumably need a lot of evidence in the input to reach this conclusion, yet analyses suggest that such input is relatively infrequent. Phillips (2013), for instance (see also Pearl & Sprouse 2013), shows that well under 10% of the *wh*-questions in a written corpus involve long-distance extraction. The basic scenario that our experiments have portrayed here, in which learners do become close to native-like in their ability to process long-distance extraction, but only after many years of exposure, makes sense given this background. L2 learners do get the input that they need to realize that long-distance extraction is possible, but it takes time for this input to have an impact.

An open question at this point is whether we would get the same type of results with learners whose L1 has long-distance extraction. A reasonable first guess would be that such learners would show the same island status for *that*-clauses that we have seen here, but that they would begin to be able to handle long-distance dependencies more quickly than the participants in our experiments, whose L1 does not use extraction for *wh*-questions.

3.2 Long-distance extraction in child English

Given what we have seen so far, we would expect children to take a relatively long time to fully acquire long-distance extraction in a language like English. As discussed above, children do not begin with the assumption that long-distance extraction will be possible and even if it is, they seem to be conservative in reaching this conclusion, so it is reasonable to expect a prolonged period of development.

Studies that have been done on children acquiring long-distance extraction in English largely bear out this expectation (see Roeper & de Villiers 2011 for an overview). Thornton (1991), for instance, found that in an elicited production task, 9 out of 20 children (ages 2;10 to 5;5) used a type of "partial movement," as in (17), or "*wh*-copying," as in (18), in place of adult long-distance extraction, as in (13).

(17) What do you think who's in that can?

(18) Who do you think who's in the box?

(19) Who do you think _ is in that can?

(17) is "partial movement" in the sense that *who* actually has scope over the entire sentence (i.e., has the same meaning as (19)), despite its surface position within the embedded clause. The "*wh*-copying" case in (18) is similar, but the higher *wh*-word matches the one in the embedded clause. In both of these cases, children are producing a type of sentence that is not in the adult input, but that allows them to avoid producing straightforward long-distance *wh*-dependencies.

McDaniel et al. (1995) studied similar structures using an acceptability experiment with 32 child participants (ages 2;11 to 5;7). They found an acceptance rate of 36% for sentences as in (17) and (18), compared to a rate of 0% with adult controls. Here too, we see children's willingness to avoid long-distance dependencies, even if this means accepting structures that are not present in the input.

In an elicited production study with 4-year-old German children, Grohe, Schulz & Muller (2011) show that the children strongly prefer the equivalents of (17) and (18) over the long-distance equivalent in (19). Sentences like (17) and (18) are possible in at least some varieties of German, but the fact that children have a strong preference for them again suggests that they are avoiding long-distance extraction.

These findings are in accord with the general thrust of the experimental results we have obtained here. Children do not seem to treat long-distance extraction as the default option in their L1 and often take years to fully adopt it, so it should not be surprising that L2 learners behave similarly, finding long-distance extraction difficult at first and then slowly adapting to it. Our results from Experiment 1 show the difficulty that L2 speakers have with long-distance extraction and how the difficulty diminishes with continued exposure, but other studies have shown that L2 speakers adopt strategies very similar to what we have just seen that children do. Schulz (2011), for instance, provides particularly striking evidence that L1 Japanese / L2 English speakers produce and accept "partial movement" questions Like (17), even though this structure is not possible in either Japanese or English (see Wakabayashi & Okawara 2003; Yamane 2003; Gutierrez 2005; for additional evidence, and Slavkov (2015) for general discussion). The picture that emerges applies both to L1 and L2: learners have difficulty with long-distance extraction, so they find reasonable alternative for expressing the intended meaning, whether these alternatives are found in the input or not.

3.3 Lexical restrictions on long-distance extraction

Our discussion so far has assumed that complement *that*-clauses, in contrast to a variety of other embedded clause types, allow extraction and are thus not islands. At a first level of approximation, this is a useful generalization, but as Ross (1967) noted, the reality is more fine-grained, in that properties of the matrix verb determine whether or not extraction is allowed out of the *that*-clause. Verbs which allow this are traditionally known as "bridge" verbs (Erteschik 1973; Erteschik-Shir 2006). These include *think*, *say*, and *decide*, as seen in (20).

(20) a. Who do you think [that Mary saw _]?
b. What did they say [that the movie is about _]?
c. Who did the committee decide [that we should invite _]?

As for "non-bridge" verbs (i.e., those that do not allow extraction), these are usually grouped in two types: factive verbs, which presuppose the truth of the complement clause, and manner-of-speaking verbs, which differ from *say* in giving some specific information about how the speaking was done. The examples with factive verbs in (21) and with manner-of-speaking verbs in (22) show how these are less tolerant of extraction than bridge verbs.

(21) a. ?Who do you regret [that Mary saw _]?
b. ??What did they realize [that the movie was about _]?
c. ??Who did the committee find out [that we invited _]?

(22) a. ?Who did you yell [that Mary saw _]?
b. ??What did they whisper [that the movie was about _]?
c. ??Who did the committee murmur [that we should invite _]?

Of the two types of non-bridge verbs, factives have been more extensively studied, often in the broader context of negative islands and *wh*-islands. Collectively, these islands (known as "weak islands" in the literature) yield much sharper results when the *wh*-phrase is a manner or degree phrase, as seen with factives in (23), but there is still a perceptible effect with argument *wh*-phrases, as in (21).

(23) a. *How do you regret [that Mary behaved _]?
b. *How short did they realize [that the movie was _]?
c. *How warm did the committee find out [that the rooms were _]?

One influential type of analysis (see, e.g., Fox & Hackl 2006; Abrusán 2014; Dayal 2016) has been to attribute the ill-formedness of questions like (21) and (23) to the idea that since the embedded clause is presupposed, the set of all possible answers is also presupposed. This set will contain some pairs of answers that cannot both be true (e.g., "the movie was 10 minutes long" and "the movie was 20 minutes long" in (23b)), so the question will inevitably lead to presupposition failure and unacceptability. Other weak islands are approached similarly under this analysis.[2]

Another influential analysis of factive islands as in (21) and (23) does not assimilate them to the more general case of weak islands, but instead subsumes them under a broad and independently motivated condition on long-distance extraction that requires that the matrix and embedded clauses be able to be grouped into a single event (Truswell 2011). Truswell argues that non-presupposed complement clauses, such as those in (20), satisfy this condition, but presupposed complement clauses, such as the factive complements in (21) and (23), do not, thus predicting a degraded status for these latter cases.

We will not attempt to decide on the correct analysis of factive islands here, but the above discussion should make clear that there is more than one analysis available that is well developed and well motivated. The manner-of-speaking cases, as in (22), have received less attention in the literature, but it is possible that the analysis of factive islands can be extended to include them (see Ambridge & Goldberg 2008 and Huang et al. (2022) for discussion).

This overview of extraction from complement *that*-clauses has a number of interesting implications for our concerns here. Most importantly, it is clear that the non-island status of *that*-clauses is not a fact about morphosyntax. The embedded clauses in (20)-(22) are all the same in their morphological properties, and presumably in their syntactic properties, but they clearly differ in the extent to which they allow extraction. It thus cannot be that learners, at least in L1, associate the non-island status of bridge verb complements to their morphology, since if they did, the island facts as in (21)-(23) would not arise. What learners are paying attention to

[2] The smaller island effect in (21) is less often discussed, but presumably a similar type of analysis can be extended to this case.

instead is whether the question is semantically well-formed (and in particular, whether it leads to contradictory presuppositions) or whether the extraction is compatible with the event structure, depending on which analysis turns out to be correct. If the speaker is successful in creating non-contradictory presuppositions or a single event structure for both clauses, as would be the case with the bridge verb complements as in (20), then extraction is possible, but otherwise it is not.

The general picture that we are left with is that speakers must be able to do something (i.e., create a particular type of representation) in order to perform long-distance extraction. *That*-clause complements of bridge verbs are not simply cases where extraction is allowed because nothing prevents it, but rather cases where the speaker is able to successfully carry out an operation that makes the extraction possible (see Momma (2022) for differences in sentence production between complement clauses out of which there has been long-distance extraction and simple complement clauses without extraction, as well as Moulton 2015 for more general discussion of clausal complementation and extraction). From this viewpoint, our finding from Experiment 1 in which L2 speakers treat bridge verb complements as islands seems unsurprising. Being able to extract out of an embedded clause requires the speaker to access detailed information about the lexical semantics of the matrix verb and use this to construct a detailed semantic representation of the *wh*-question, so it is reasonable that the path to being able to do this quickly and reliably would be very protracted. The general finding from the literature that in both L1 and L2, learners often avoid long-distance extraction, also makes sense under this view. If the speaker cannot do what is required to enable long-distance extraction, other strategies for expressing the question will be deployed.

3.4 Interim conclusion

We have now seen some facts about long-distance extraction across a variety of languages, in both L1 and L2 acquisition, and across a variety of complement *that*-clauses in English, and there are two general conclusions. First, it is clear that extraction out of *that*-clauses is either prohibited or avoided in a number of languages, for many children (even in a language like English), and for some complement types in English. Second, it seems that when speakers allow extraction out of a *that*-clause, it is not just because nothing prevents them from doing so, but because they have performed a specific operation that permits it. We have seen this in languages in which all the components of long-distance extraction are present, but it nevertheless does not occur, in children who have likewise learned all of the components of long-distance extraction, but nonetheless avoid it, and in *that*-clauses in English that do not allow long-distance extraction, even though they seem to be morphologically and syntactically identical to those that do. These conclusions are relevant to our concerns here, because they make our basic findings from Experiment 1 more understandable. If learning how to do long-distance extraction requires learning something new, on top of learning how to do extraction and how to embed clauses, then it makes sense that

L2 learners will have difficulty doing it, and we would expect this difficulty to be reflected in acceptability. In addition, it makes sense that given enough years of input, the difficulty for L2 learners would gradually decrease, and this too should be reflected in acceptability. Both of these expectations were met in our results from Experiment 1: L2 learners showed a large degradation for extraction from *that*-clauses, on a par with islands, but this degradation decreased as AoA decreased (see Kim & Goodall (in press) for additional evidence that L2 learners have difficulty with extraction from *that*-clauses).

4 Conclusion

We have presented two major sets of empirical findings in this paper. First, we have shown that L2 speakers, like L1 speakers, exhibit significant degradation in acceptability in the presence of a long-distance *wh*-dependency out of a *that*-clause, but that the effect is much larger in L2 than in L1 and is comparable in size to the effect observed in islands. In addition, the degradation in L2 appears to get smaller as the speaker's amount of exposure to the language increases. Second, we have shown that both L1 and L2 speakers exhibit significant degradation when there has been extraction out of an island. In this case, though, the size of the degradation is very similar across the two groups and we do not see any decrease in degradation as AoA decreases.

Descriptively, then, it seems that L2 speakers treat *that*-clauses as islands, but that these clauses become less island-like as the speaker accumulates more experience with the language. This scenario is puzzling under the traditional view that extraction out of *that*-clauses is freely available simply because nothing prevents it, and it is even more puzzling given that L2 speakers otherwise track L1 speakers very closely in their reactions to classical islands. It makes much more sense, however, if long-distance extraction is not available "for free," but in fact requires speakers to perform operations that go beyond simply being able to do extraction and have embedded clauses, such as ensuring that there are no contradictory presuppositions or that the two clauses can be grouped together as a single event. L1 speakers have learned how to do this for *that*-clauses, though apparently at some cost in terms of acceptability, but not all L2 speakers have done the same. Put simply, then, *that*-clauses are island-like for L2 speakers because they have not yet learned how to make them non-islands. They do begin to learn how to do this, but it appears to be a very gradual, years-long process.

This view of complement clauses, in which their non-island status must be learned, is compatible with a number of suggestions from the literature (e.g., Culicover & Jackendoff 2005; Pearl & Sprouse 2013) and finds support from various aspects of long-distance extraction, including cross-linguistic variation, child language, and lexical restrictions, but the L2 evidence that we have seen here seems particularly striking, in that we can observe *that*-clauses becoming less and less island-like as speakers gain additional experience with the language.

The current study is also of interest for two methodological reasons. First, it shows that L1 and L2 can be profitably studied and compared using the techniques of formal sentence acceptability experiments from "experimental syntax." These techniques have been used very widely and usefully in adult L1 studies in recent years, but they have had less of an impact on the L2 literature. The experiments conducted here suggest that these techniques have a lot to offer in the area of L2. Second, the current study demonstrates how evidence from L2 can play a role in helping to decide issues of broader theoretical interest. In this particular case, we have seen how the behavior of L2 speakers in acceptability experiments is puzzling under standard views of complement clauses, in which their non-island status is assumed to be true by default, and argues instead for a view in which speakers gradually learn to perform the mechanisms needed for long-distance extraction, thus in effect gradually turning complement clauses into non-islands.

Acknowledgements

We dedicate this article to the memory of Bill Davies, teacher to the first author, classmate of the second, and esteemed friend of both. For their many comments and suggestions, we are also grateful to the anonymous reviewers and to audiences at the Annual Meeting of the Linguistic Society of America and the CUNY Conference on Human Sentence Processing, and to the members of the Experimental Syntax Lab and Syntax and Semantics Babble at UC San Diego. Special thanks also to Savithry Namboodiripad, Dave Kush and Adam Morgan for organizing the symposium on Experimental Approaches to Cross-linguistic Variation in Island Phenomena, at which some of this work was presented.

Competing interests

The authors have no competing interests to declare.

References

Abrusán, Marta. 2014. Weak island semantics. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/acprof:oso/9780199639380.001.0001

Aldwayan, Saad & Fiorentino, Robert & Gabriele, Alison. 2010. Evidence of syntactic constraints in the processing of *wh*-movement: a study of Najdi Arabic learners of English. In Van Patten, Bill & Jegerski, Jill (eds.), Research in Second Language Processing and Parsing, Amsterdam: John Benjamins Publishing Company, 65–86. DOI: https://doi.org/10.1075/lald.53.03ald

Alexopoulou, Theodora & Keller, Frank. 2007. Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, 110–160. DOI: https://doi.org/10.1353/lan.2007.0001

Ambridge, Ben & Goldberg, Adele E. 2008. The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics* 19(3). 357–389. DOI: https://doi.org/10.1515/COGL.2008.014

Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67. 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Belikova, Alyona & White, Lydia. 2009. Evidence for the fundamental difference hypothesis or not? Island constraints revisited. *Studies in Second Language Acquisition* 31(2). 199–223. DOI: https://doi.org/10.1017/S0272263109090287

Boxell, Oliver & Felser, Claudia. 2017. Sensitivity to parasitic gaps inside subject islands in native and non-native sentence processing. *Bilingualism: Language and Cognition* 20(3). 494–511. DOI: https://doi.org/10.1017/S1366728915000942

Chen, Evan & Gibson, Edward & Wolf, Florian. 2005. Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language* 52(1). 144–169. DOI: https://doi.org/10.1016/j.jml.2004.10.001

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

Culicover, Peter W. & Jackendoff, Ray. 2005. Simpler Syntax. New York: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199271092.001.0001

Cunnings, Ian. 2017. Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition* 20(4). 659–678. DOI: https://doi.org/10.1017/S1366728916000675

Dallas, Andrea & DeDe, Gayle & Nicol, Janet. 2013. An Event-Related Potential (ERP) Investigation of Filler-Gap Processing in Native and Second Language Speakers. *Language Learning* 63(4). 766–799. DOI: https://doi.org/10.1111/lang.12026

Dallas, Andrea & Kaan, Edith. 2008. Second Language Processing of Filler-Gap Dependencies by Late Learners. *Language and Linguistics Compass* 2(3). 372–388. DOI: https://doi.org/10.1111/j.1749-818X.2008.00056.x

Davies, Wiliiam D. & Kurniawan, Eri. 2013. Movement and Locality in Sundanese Wh-Questions. *Syntax* 16(2). 111–147. DOI: https://doi.org/10.1111/j.1467-9612.2012.00174.x

Dayal, Veneeta. 2016. Questions. Oxford Surveys in Semantics and Pragmatics. Oxford University Press. Oxford, UK.

Erteschik, Nomi. 1973 On the Nature of Island Constraints. Ph.D. dissertation, MIT.

Erteschik-Shir, Nomi. 2006. Bridge Phenomena. In Everaert, Ma and van Riemsdijk, Henk (eds.), *The Blackwell companion to syntax*, 284–294. Oxford: Blackwell. DOI: https://doi.org/10.1002/9780470996591.ch10

Fanselow, Gisbert. 2006. Partial *wh*-movement. In Everaert, Martin & van Riemsdijk, Henk (eds.), *The Blackwell companion to syntax*, vol. III, 437–492. Oxford: Blackwell. DOI: https://doi.org/10.1002/9780470996591.ch47

Fanselow, Gisbert. 2021. Acceptability, grammar and processing. In Goodall, Grant (ed.), The Cambridge Handbook of Experimental Syntax. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781108569620.006

Felser, Claudia & Cunnings, Ian & Batterham, Claire & Clahsen, Harald. 2012. The timing of island effects in nonnative sentence processing. *Studies in Second Language Acquisition* 34(1). 67–98. DOI: https://doi.org/10.1017/S0272263111000507

Felser, Claudia & Roberts, Leah. 2007. Processing *wh*-dependencies in a second language: A cross-modal priming study. *Second Language Research* 23(1). 9–36. DOI: https://doi.org/10.1177/0267658307071600

Fodor, Janet Dean 1978. Parsing strategies and constraints on transformations. *Linguistic Inquiry* 9(3). 427–473.

Fox, Danny & Hackl, Martin. 2006. The universal density of measurement. *Linguistics & Philosophy* 29. 537–586. DOI: https://doi.org/10.1007/s10988-006-9004-4

Frazier, Lyn. & Clifton Jr, Charles. 1989. Successive cyclicity in the grammar and the parser. *Language and cognitive processes* 4(2). 93–126. DOI: https://doi.org/10.1080/01690968908406359

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68. 1–76. DOI: https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Marantz, Alec & Miyashita, Yasushi & O'Neil, Wayne (eds.), *Image, language, brain*, 95–126. Cambridge, MA: MIT Press.

Goodall, Grant. 2021. Sentence acceptability experiments: What, how, and why. In Goodall, Grant (ed.) *The Cambridge Handbook of Experimental Syntax*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781108569620

Goodall, Grant. 2022. Theory and Experiment in Syntax. New York & London: Routledge. DOI: https://doi.org/10.4324/9781003160144

Grohe, Lydia & Schulz, Petra & Müller, Anja. 2011. How children "Copy" long-distance structures: The production of complex Wh-questions in German. In *Proceedings of the 35th annual Boston University Conference on Language Development*, 233–245.

Gutierrez, María J. 2005. The acquisition of English LD *wh*-questions by Basque/Spanish bilingual subjects in a school context. Unpublished PhD dissertation, University of the Basque Country, Leioa, Spain.

Hartshorne, Joshua K. & Tenenbaum, Joshua B. & Pinker, Steven. 2018. A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition* 177. 263–277. DOI: https://doi.org/10.1016/j.cognition.2018.04.007

Huang, Nick & Almeida, Diogo & Sprouse, Jon. 2022. How good are leading theories of bridge verbs? An experimental evaluation. *Paper presented at the West Coast Conference on Formal Linguistics*.

Ionin, Tania. 2021. Acceptability studies in L2 populations. In Goodall, Grant (ed.) *The Cambridge Handbook of Experimental Syntax*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781108569620.015

Jessen, Anna & Festman, Julia & Boxell, Oliver & Felser, Claudia. 2017. Native and non-native speakers' brain responses to filled indirect object gaps. *Journal of Psycholinguistic Research* 46(5). 1319–1338. DOI: https://doi.org/10.1007/s10936-017-9496-9

Johnson, Adriene & Fiorentino, Robert & Gabriele, Alison. 2016. Syntactic constraints and individual differences in native and non-native processing of *wh*-movement. *Frontiers in Psychology* 7. 549. DOI: https://doi.org/10.3389/fpsyg.2016.00549

Keffala, Bethany. 2011. Resumption and gaps in English relative clauses: Relative acceptability creates an illusion of 'saving'. *Berkeley Linguistics Society* 37. 140–154. DOI: https://doi.org/10.3765/bls.v37i1.846

Khomitsevich, Olga. 2008. Dependencies across phases: From sequence of tense to restrictions on movement. Ph.D. dissertation, Utrecht University.

Kim, Boyoung. 2015. Sensitivity to islands in Korean-English bilinguals. Doctoral dissertation, UC San Diego.

Kim, Boyoung & Goodall, Grant. 2011. Age-related effects on constraints on *wh*-movement. In Herschensohn, Julia & Tanner, Darren (eds.), *Proceedings of the 11th Generative Approaches to Second Language Acquisition Conference (GASLA)*. Cascadilla Proceedings Project.

Kim, Boyoung & Goodall, Grant. 2016. Islands and non-islands in native and heritage Korean. *Frontiers in Psychology* 7. 134. DOI: https://doi.org/10.3389/fpsyg.2016.00134

Kim, Boyoung & Goodall, Grant. In press. The source of the *that*-trace effect: New evidence from L2 English. *Second Language Research*. DOI: https://doi.org/10.1177/02676583221104604

Kim, Eunah & Baek, Soondo & Tremblay, Annie. 2015. The role of island constraints in second language sentence processing. *Language Acquisition* 22(4). 384–416. DOI: https://doi.org/10.1080/10489223.2015.1028630

Kim, Jeong-eun & Nam, Hosung. 2017. Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition* 39(3). 431–57. DOI: https://doi.org/10.1017/S0272263115000510

Kluender, Robert & Kutas, Marta. 1993. Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience* 5(2). 196–214. DOI: https://doi.org/10.1162/jocn.1993.5.2.196

Kuznetsova, Alexandra & Brockhoff, Per B. & Christensen, Rune H. 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

Leal, Tania. & Slabakova, Roumyana & Farmer, Thomas A. 2017. The fine-tuning of linguistic expectations over the course of L2 learning. *Studies in Second Language Acquisition* 39(3). 493–525. DOI: https://doi.org/10.1017/S0272263116000164

Love, Tracy & Maas, Edwin & Swinney, David. 2003. The influence of language exposure on lexical and syntactic language processing. *Experimental Psychology* 50(3). 204. DOI: https://doi.org/10.1026//1617-3169.50.3.204

Marinis, Theodore & Roberts, Leah & Felser, Clauda & Clahsen, Harald. 2005. Gaps in second language sentence processing. *Studies in Second Language Acquisition* 27(1). 53–78. DOI: https://doi.org/10.1017/S0272263105050035

Martohardjono, Gita 1993 *Wh*-movement in the acquisition of a second language: A cross-linguistic study of three languages with and without movement. Ithaca, New York: Cornell University dissertation.

McDaniel, Dana & Chiu, Bonnie & Maxfield, Thomas L. 1995. Parameters for *wh*-movement types: Evidence from child English. *Natural Language & Linguistic Theory* 13(4). 709–753. DOI: https://doi.org/10.1007/BF00992856

Momma, Shota. 2022. Producing filler-gap dependencies: Structural priming evidence for two distinct combinatorial processes in production. *Journal of Memory and Language,* 126. DOI: https://doi.org/10.1016/j.jml.2022.104349

Moulton, Keir. 2015. CPs: Copies and Compositionality. *Linguistic Inquiry* 46(2). 305–342. DOI: https://doi.org/10.1162/LING_a_00183

Omaki, Akira & Schulz, Barbara. 2011. Filler-gap dependencies and island constraints in second-language sentence processing. *Studies in Second Language Acquisition* 33(4). 563–588. DOI: https://doi.org/10.1017/S0272263111000313

Ortega-Santos, Iván & Reglero, Lara & Franco Elorza, Jon A. 2018. *Wh*-Islands in L2 Spanish and L2 English: A Poverty of the Stimulus and Data Assessment. *Fontes Linguae Vasconum* 126. 435–471. DOI: https://doi.org/10.35462/FLV126.7

Pearl, Lisa & Sprouse, Jon. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition* 20(1). 23–68. DOI: https://doi.org/10.1080/10489223.2012.738742

Phillips, Colin. 2013. On the nature of island constraints II: Language learning and innateness. In Sprouse, Jon & Hornstein, Norbert (eds.), *Experimental syntax and island effects*, 132–157. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139035309.007

Pliatsikas, Christos & Johnstone, Tom & Marinis, Theodore. 2017. An fMRI study on the processing of long-distance *wh*-movement in a second language. *Glossa: A Journal of General Linguistics* 2(1). 101. DOI: https://doi.org/10.5334/gjgl.95

Pliatsikas, Christos & Marinis, Theodore. 2013. Processing empty categories in a second language: When naturalistic exposure fills the (intermediate) gap. *Bilingualism: Language and Cognition* 16(1). 167–182. DOI: https://doi.org/10.1017/S136672891200017X

Plonsky, Luke & Marsden, Emma & Crowther, Dustin & Gass, Susan M. & Spinner, Patti. 2020. A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research* 36(4). 583–621. DOI: https://doi.org/10.1177/0267658319828413

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

Roeper, Thomas & de Villiers, Jill. 2011 The Acquisition Path for *Wh*-Questions. In de Villiers, Jill & Roeper, Thomas (eds.), *Handbook of Generative Approaches to Language Acquisition. Studies in Theoretical Psycholinguistics* 41. 189–246. Springer, Dordrecht. DOI: https://doi.org/10.1007/978-94-007-1688-9_6

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

Saddy, Doug. 1991. *Wh*-scope mechanisms in Bahasa Indonesia. *MIT Working Papers in Linguistics* 15, 183–218.

Schulz, Barbara. 2011. Syntactic creativity in second language English: *wh*-scope marking in Japanese-English interlanguage. *Second Language Research* 27(3). 313–341. DOI: https://doi.org/10.1177/0267658310390503

Slavkov, Nikolay. 2015. Long-distance *wh*-movement and long-distance wh-movement avoidance in L2 English: Evidence from French and Bulgarian speakers. *Second Language Research* 31(2). 179–210. DOI: https://doi.org/10.1177/0267658314554939

Sprouse, Jon & Almeida, Diogo. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: A Journal of General Linguistics*. 2(1). 236. DOI: https://doi.org/10.5334/gjgl.236

Sprouse, Jon & Hornstein, Norbert. 2013. Experimental syntax and island effects. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139035309

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

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

Thornton, Rosalind J. 1991. Adventures in long-distance moving: The acquisition of complex wh-questions. Storrs, CT: University of Connecticut dissertation.

Truswell, Robert. 2011. Events, phrases, and questions. New York: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199577774.001.0001

Wakabayashi, Shigenori & Okawara, Izumi. 2003. Japanese learners' errors on long distance *wh*-questions. In Wakabayashi, Shigenori (ed.), *Generative approaches to the acquisition of English by native speakers of Japanese*, 215–246. Berlin: Mouton.

Wanner, Eric & Maratsos, Michael. 1978. An ATN approach to comprehension. In Halle, Morris & Bresnan, Joan & Miller, George A. (eds.), *Linguistic Theory and Psychological Reality*, 119–161. Cambridge, MA: MIT Press.

White, Lydia & Juffs, Alan. 1998 Constraints on *Wh*-movement in two different contexts of non-native language acquisition: Competence and processing. In Flynn, Suzanne & Martohardjono, Gita & O'Neill, Wayne (eds.), *The generative study of second language acquisition*. 11–130. Hillsdale, NJ: Lawrence Erlbaum.

Yamane, Maki. 2003. On interaction of first-language transfer and universal grammar in adult second language acquisition: *WH*-movement in L1-Japanese/L2-English interlanguage. Storrs, CT: University of Connecticut dissertation.