Definitely Islands? Experimental investigation of definite islands

Anissa Neal & Brian Dillon*

Abstract. Experimental work on islands has used formal acceptability judgment studies to quantify the severity of different island violations. This current study uses this approach to probe the (in-)violability of definite islands, an understudied island, in offline and online measures. We conducted two acceptability judgment studies and find a modest island effect. However, rating distributions appear bimodal across definites and indefinites. We also conducted a self-paced reading experiment, but found no significant effects. Overall, offline, definite islands differ from other uniform islands, but online, the results are more complicated.

Keywords. syntactic islands; definiteness; psycholinguistics

1. Introduction. This paper aims to investigate the offline and online processing status of an understudied class of island: definite islands. By definite islands, we mean the apparent ban on extraction from inside definite determiner phrases that was first observed in Ross' dissertation (Ross 1967).

(1) a. Who did Irina see **a** picture of _?
b. *Who did Irina see **that/his** picture of _?

In this work, we will first consider offline judgments to determine how acceptable speakers find these constructions in isolation, and then move to online processing to test whether speakers are sensitive to definite islands in real time.

1.1. DEFINITE ISLANDS. One interesting feature of definite islands is their somewhat variable status (Chomsky 1973). Explanations for this gradience vary (Chomsky 1973, Keller 2000, Davies & Dubinsky 2003), but it is a shared intuition the example below is intermediate in terms of acceptability (Chomsky 1973).

(2) ?Who did Irina see **the** picture of _?

One class of accounts characterizes these as syntactic violations. The DP may be a bounding node that blocks extraction (Chomsky 1977, 1973, Davies & Dubinsky 2003, Huang 2018). However, extraction from a DP is possible if certain criteria are met.

(3) Who did Irina write the cruel article about _?

For example, Davies and Dubinsky argue that in (3) the presence of a verb of creation, *write*, and a semantically related result nominal, *article*, can override the blocking effect of the DP through abstract noun incorporation. Huang (2018) suggests an account that relies on a bound possessor with unvalued phi-features allowing *wh*-movement from the DP. Other accounts take a more semantic

*Many thanks to Maayan Keshev, Dave Kush, Ana Arregui, and UMass's psycholinguistic and semantics workshops. Authors: Anissa Neal, University of Massachusetts, Amherst (anneal@umass.edu) & Brian Dillon, University of Massachusetts, Amherst (bwdillon@umass.edu).

approach, such as the work done by Simonenko (2015) on definite DPs in Austro-Bavarian German. Simonenko shows that extraction from strong definites creates an uninformative statement by presupposing the content of the possible answers. This provides one potential explanation for their unacceptability, under the assumption that unacceptability can result from an uninformative statement.

Definite islands are also predicted by more recent-discourse based accounts. The Background Constituents are Islands by Goldberg (2013), as the name suggests, posits that backgrounded elements cannot be extracted, and are thus islands. Goldberg considers background elements to be constituents that are neither part of the focus domain nor the primary topic. Elements that are part of a presupposed clause could then be considered backgrounded. As she and others note, the position of the gap must be within the asserted content of the utterance to be a licit extraction (Erteschik-Shir 1973), and cannot be presupposed. Assuming definite DPs are presupposed content, it would be unacceptable to subextract from definite DPs under this account.

Lastly, Hofmeister & Sag (2010) also investigate referential processing inside the DP. They note that specificity and/or referentiality may consume processing resources, and that filler-gap dependencies are more easily processed when the intervening material is less complex. Definiteness may add another layer of difficulty, since a definite DP identifies and situates a specific referent in a discourse model. On this view, the low acceptability for extraction from definite DPs is thought of as a reflection of an increased processing toll for filler-gap dependencies in this environment.

1.2. PROCESSING ISLANDS. Much psycholinguistic work has investigated if the parser posits gaps inside island structures (Phillips (2006) for a more comprehensive review), and the bulk of it has used a small subset of islands (e.g., relative clause, complex NP, wh-islands).

In a study very similar to the current one, Tollan & Heller (2015) found sensitivity to definite islands in an offline task but not in an online task. They investigated two key points: (i) how sensitive is the parser to definiteness in processing filler-gap dependencies, and (ii) if, and how, the type of wh-phrase influences the parsing of the dependency. Their first experiment was an online, self-paced reading task using the filled-gap effect paradigm. The filled-gap effect refers to the finding that readers have difficulty when another word is already in a position where they expect to see a gap (Stowe 1986). This paradigm can be used to determine where readers are actively positing possible gap sites as they parse a sentence. Examples of their stimuli are below. They also include a manipulation of d-linked vs. non d-linked fillers.

(4) a. Which singer did Lizzie see [a/the movie about Elvis Presley] with _ ?
b. Who did Lizzie see [a/the movie about Elvis Presley] with _?
c. Did Lizzie see [a/the movie about Elvis Presley] with Kate Bush?

Participants were given a preceding context that made the question felicitous, and were then presented with the sentences word-by-word. Yes/no questions were included as a baseline condition to measure the size of the filled-gap effect. The authors predicted no effect of definiteness in the first gap position at "*a/the movie*," and compare only the d-linked and non d-linked items. They find a significant slowdown in this region for the d-linked phrase, suggesting participants had greater difficulty and hence a greater filled-gap effect, with *which* than *who*. In the second gap position, they find no definiteness effect in the yes/no questions, and focus on the filled-gap

sentences. They do find a main effect of wh-type with *which*-NPs having slower reading times. However, there is no effect of definiteness.

The authors also ran an offline, question completion study.

(5) a. Which singer did Lizzie see a/the movie about _____

b. Who did Lizzie see a/the movie about _____

c. Did Lizzie see a/the movie about _____

Participants were given the same short, preceding contexts as Experiment 1 and then asked to complete the sentence prompts. Ending the sentence by adding a question mark was a possible response. The authors reasoned that if the participants find a gap inside an island acceptable then they should add a question mark after *about* in all cases. However, if this is an unacceptable gap site, they may continue the sentence to place the gap outside the DP (i.e., *Who did Lizzie see the movie about Elvis Presley with _?*). Yes/no questions were once again used as a control. The offline results showed evidence for a definiteness effect: Participants placed gaps inside indefinite NPs more than definite NPs.

Thus the authors found mixed evidence of sensitivity to definiteness in processing. Furthermore, they use the same methodology used in this study, an online self-paced reading study paired with an offline task. However, one of the main differences between this study and the one done in this paper, aside from the d-linking manipulation, is different baselines. In the Tollan and Heller study, they use yes/no questions as their control condition for both the offline and online studies. While yes/no questions and wh-questions share certain similarities, there are a variety of syntactic and semantic differences that could disrupt the interpretation of the results. The controls in the present study manipulate only the presence of a filler, given in (7).

The Tollan and Heller results also raise the possibility that the parser will entertain gaps inside definite islands in real-time processing, in apparent violation of a grammatical constraint. This would be a surprising conclusion in light of the broader literature, and so bears further scrutiny. Therefore, it is crucial to see if the Tollan and Heller results replicate, and if there is actually a distinction between the offline and online results.

2. Experiment 1: Acceptability Judgment. As seen in (1), there is a range of judgments associated with definite islands. The purpose of Experiments 1a and 1b is to determine if naive informants share the intuitions reported in the literature. That is, just how acceptable do speakers consider extractions from *the*-DPs to be. To investigate this, we used the factorial paradigm for islands developed by Sprouse et al. (2016). This methodology has been successfully applied to a variety of different islands (e.g relative clause, complex NP, subject, adjunct, and *wh*-islands) in a variety of languages.

The factorial design is useful for several reasons. It allows for a quantitative measure of an "island effect." As Sprouse et al. (2016) notes in his overview of this design, there many extra-syntactic factors that could go into the decreased acceptability observed in islands. For example, long distance dependencies are considered less acceptable than short distance ones, such as in the example below, where the filler is much further in the second case than in the first.

(6) a. INDEFINITE, MATRIX: Tara knows who _ found **a** photo of Yelena.

b. INDEFINITE, EMBEDDED: Tara knows who Fiona found **a** photo of _.

c. DEFINITE, MATRIX: Tara knows who _ found **the** photo of Yelena.

d. DEFINITE, EMBEDDED: Tara knows who Fiona found **the** photo of _.

Furthermore, the complexity of the island itself could impact the acceptability. Different types of islands could introduce a certain amount of complexity related to structure, meaning, or both. Sentences containing islands may thus, by virtue of the island structure alone, have a lower acceptability. With the factorial design, extra-syntactic factors like dependency length and complexity are independently estimated and accounted for. The effect of length can be measured by subtracting (6d) from (6c). The effect of the island structure can be measured by subtracting (6a) from (6c). The total effect is measured by subtracting (6b) from (6c). The length and structure effects are then subtracted from the total effect, and the remainder is the size of the "island effect." The island effect is quantified as this difference-in-differences (DD) score, with larger values indicating more severe island penalties.

2.1. EXPERIMENT 1A. EMBEDDED JUDGMENT STUDY.

Participants. 41 native American English participants were recruited for an online acceptability judgment study run on IbexFarm (Drummond 2020) through the Prolific Academic platform and paid $4 each.

Materials. The study was a 2x2 within-subjects factorial design with factors of DISTANCE: LONG, SHORT and DEFINITENESS: INDEFINITE, DEFINITE

(7) a. The journalist guessed who _ promoted a/the ridiculous photo of Madonna.

b. The journalist guessed who Charlie promoted a/the ridiculous photo of _.

There were a total of 24 experimental items and 48 filler items, designed it to have approximately equal numbers of acceptable and unacceptable items.

Of the filler items, 33 of the 48 were ungrammatical, and about half of the ungrammatical fillers shared similar characteristics to the experimental items. Some of the fillers were taken from the Sprouse et al. (2016) experiment.

Procedure. Participants were given instructions on how to rate acceptability and tested on three practice sentences. Each sentence was presented in full, and the participant was asked to give the provided sentence a rating on a scale from 1 to 7. 1 was the most unacceptable, and 7 was the most acceptable. The experimental lists were constructed in Latin Square fashion; each participant saw only one experimental token from the above paradigm for each item.

They were encouraged to use the full range of the scale in the instructions. Participants could complete the experiment at their own pace on IbexFarm; it was expected to take 20 minutes to complete, and the average completion time was 15 minutes.

Analysis. The participants read a series of questions to ensure they were properly following instructions. If the accuracy of their response was less than or equal to 60%, the participant was excluded.

A total of 39 out of 41 participants were analyzed. The data were z-transformed before analysis to account for different scale usages across participant. The factors were sum coded (Definiteness: Definite = -0.5, Indefinite = 0.5; Distance: Short = -0.5, Long = 0.5) and a mixed-effects

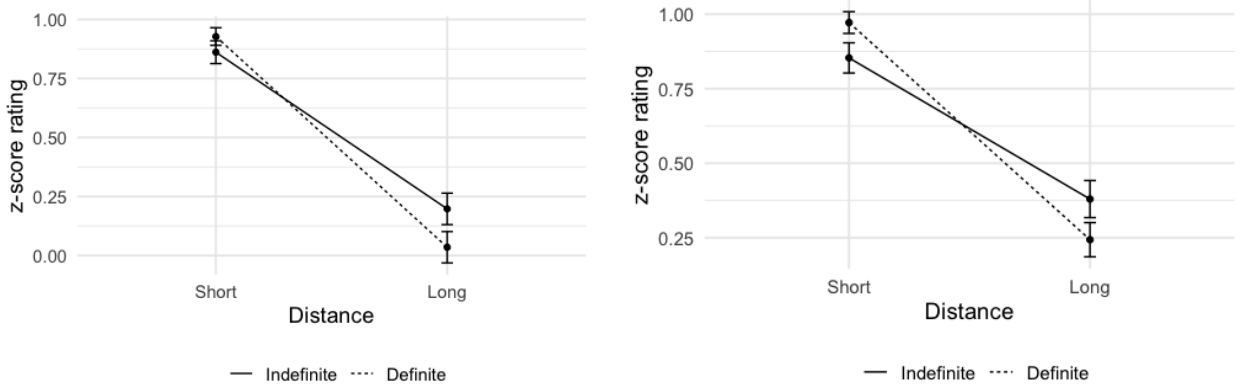
	Experiment 1a				Experiment 1b			
	Estimate	SE	*t*	*p*	Estimate	SE	*t*	*p*
DISTANCE	-0.04	0.06	-0.73	0.001	0.06	0.07	8.99	<0.001
DEFINITENESS	0.79	0.07	11.08	0.46	0.002	0.06	0.028	0.98
DEFINITENESS:DISTANCE	0.22	0.11	1.92	0.055	0.25	0.11	0.003	0.028

Table 1: Experiment 1 Linear Regression Results

linear regression model using the lmer test of the lme4 package in R (Bates et al. 2015, R Core Team 2020) was fit to the z-scored data. Following Matuschek et al. (2017), random slopes were removed until convergence

2.1.1. RESULTS. Figure (1a) presents the interaction plot for the z-scored means. The linear model found a significant main effect of distance and a marginal interaction (Table 1). Overall, the long filler-gap conditions were less acceptable, but the length penalty was larger for the definite conditions.

2.1.2. EXPERIMENT 1A DISCUSSION. We observed a large effect of distance on acceptability. Sub-extraction from inside the DP was much worse than short distance extraction across both types of DPs, definite and indefinites. The low ratings for long distance extraction are likely related to the fact that people prefer shorter filler-gap dependencies. Previous work on cross-clausal extraction (e.g., McElree et al. (2003)) indicates that the longer the distance between the filler and the gap, the more processing difficulties arise. The marginal interaction is interesting as it hints at the possibility of an island effect, but it is inconclusive. Sprouse et al. (2016) notes that the factorial paradigm produces three ways in which to observer an island effect: the presence of a significant interaction, visual absence of parallel lines on the interaction plot, and a difference-in-differences score that is greater than 0. The results of Experiment 1a satisfy all but the first. Interestingly, Sprouse et al. notes that the magnitude of a DD score is a concern for syntactic theory. While these two of these three points would appear to suggest an island effect in definite DPs, the numerically small DD leaves this question open.



(a) Interaction plot for 1a

(b) Interaction plot for 1b

Figure 1: Experiment 1 Interaction Plots

2.2. EXPERIMENT 1B: DIRECT JUDGMENT STUDY. The second experiment followed the same methodology and design as the first experiment but tested direct *wh*-questions instead of embedded indirect *wh*-questions. The reason for this was to test pragmatic-semantic accounts that tie this effect straightforwardly to the semantics of direct questions Simonenko (2015).

Participants. A total of 43 native American English participants were recruited for an online acceptability judgment study run on IbexFarm through the Prolific Academic platform. As in the first experiment, they were paid $4. Filters were put in place to ensure that participants that had completed the previous study could not participate in this study.

Materials. The stimuli was based off those of Experiment 1a, but they were adjusted to be direct questions. The matrix clause was removed, and the embedded verb, which contains the DP, became the main verb.

(8) a. Who _ published a/the horrible article about Gina?
b. Who did Olivia publish a/the horrible article about _?

As in Experiment 1a, there were a total of 24 experimental items and 48 filler items.

Procedure. The procedure was the same as in Experiment 1a.

Analysis. The same exclusion criteria from Experiment 1a were also applied. A total of 40 out of 43 participants were analyzed. One item was removed from analysis due to coding error making a total of 23 experimental items. Analysis was identical to Experiment 1a.

2.2.1. RESULTS. Table 1b presents interaction plot, and Table 1 the results of the linear model. We observed a significant effect of distance and a significant interaction of definiteness and distance, visualized in Figure 1b.

2.2.2. EXPERIMENT 1B DISCUSSION. Experiment 1b differs from Experiment 1a in that there was a significant interaction between distance and definiteness. That is, there appears to be a super-additive effect in the sense that the low acceptability of (8b) cannot be explained through the individual effects of distance or definiteness alone. Again, we observed a large effect of distance. We are, however, hesitant to conclude that these results show evidence of an island effect in direct questions, but not in embedded. First, a direct comparison between the two is difficult due to a lack of power. Second, while there is a significant interaction in direct questions, one of the key benefits of the Sprouse paradigm is a measure of an islands magnitude. The DD score for both experiments are of a lower magnitude than other observed islands (see Figure 2), and in fact, are somewhat similar to each other.

2.3. EXPERIMENT 1 DISCUSSION. The DD scores of both experiments were very close, 0.22 and 0.25, respectively. These DD scores are on the lower end of DD scores observed for island constructions when compared to other islands in the literature. While Kush et al. (2019) states that there is no true threshold for DD scores to be representative of a real island effect, the range from previous studies fall within 0.75 to 1.25. Figure 2 is a summary of DD scores across previous experiments following the same factorial design. Results of the current study are presented in red. The dashed-grey line indicates that there was a significant interaction of distance and definiteness

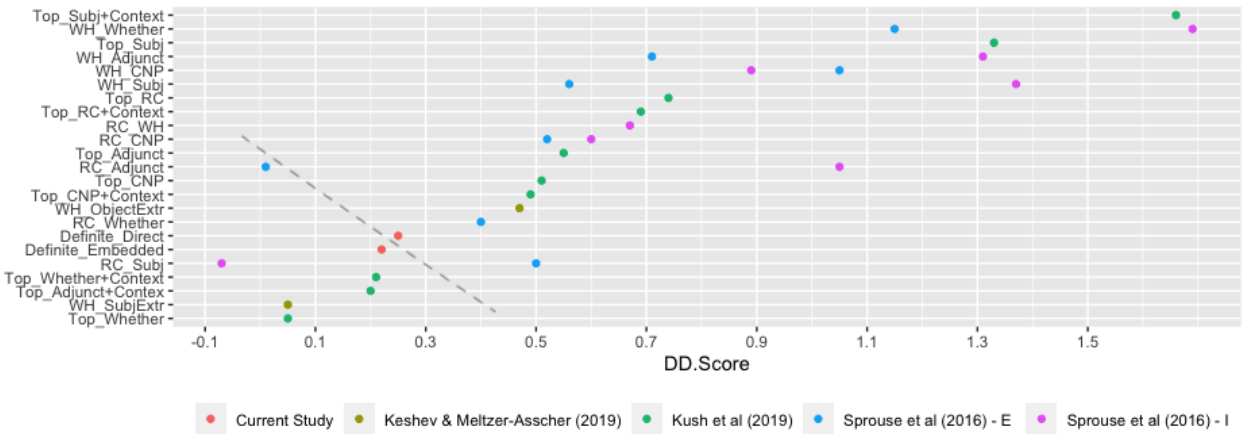


Figure 2: DD scores across studies.

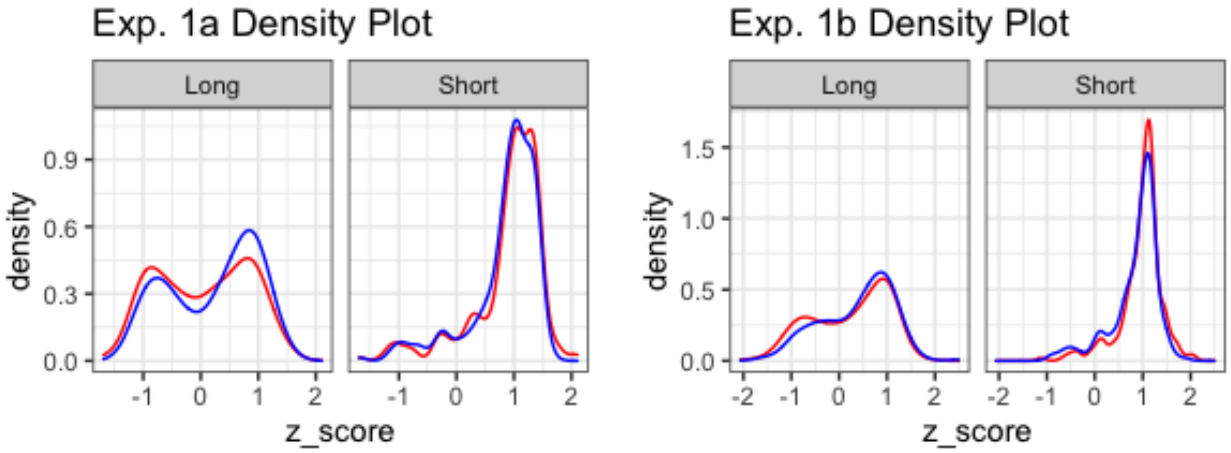


Figure 3: Participant Rating Distribution for Experiment 1. Red indicates definites, and blue indefinites.

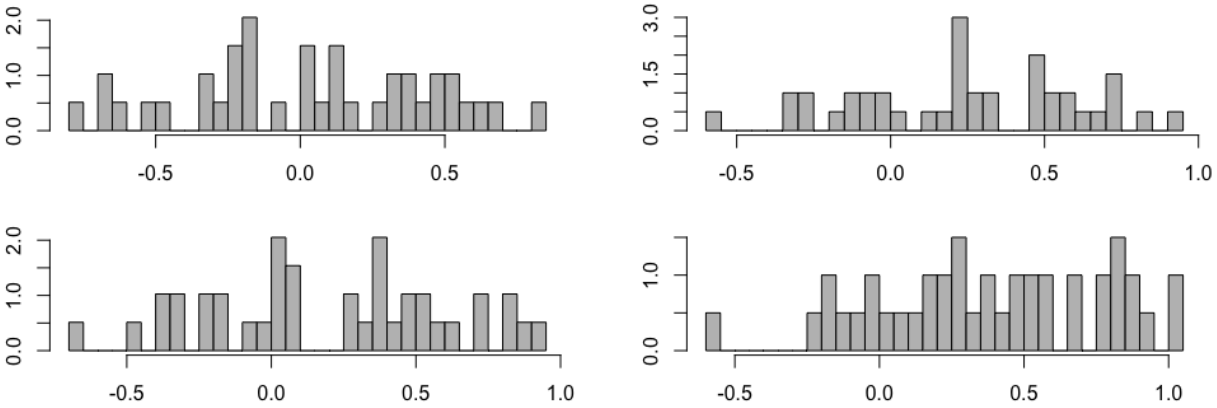
for the islands above and to the right of the line. The languages used in the studies were Hebrew, Norwegian, English, and Italian, respectively.

As seen in Figure 2, definite islands maintains a borderline placement compared to other islands. The DD scores for both conditions are on the lower side, patterning more with non-islands in English and other languages. For example, the RC Adjuncts from Sprouse et al. (2016) have a DD score of 0.01. Sprouse and colleague concluded that the apparent unacceptability of RC Adjuncts can be explained through an effect of distance, rather than a true "island effect." While the definite islands produced a larger DD than did RC Adjuncts, it is still appears quite small compared to the more robust islands with larger DD scores. This raises the question as to whether the definite islands are akin to other islands, or if the effect could be explained as compounding effects of distance and definiteness. While the results of Experiment 1 were instrumental in developing a further understanding of how acceptable these islands are offline, it does not present entirely

conclusive evidence. Experiment 2 investigates if speakers will posit gaps inside definite islands during online processing. Given the results of Experiment 1, it is unclear what behavior we might observe in definite islands in terms of active gap positing. While previous evidence suggests that gap-filling does not occur inside islands, the nebulous judgments from Experiments 1 suggest that these are not particularly strong island environments. Furthermore, the results from Tollan & Heller (2015) suggest that there may be a offline-online asymmetry for definite islands, as they found evidence of a definiteness effect offline but not online.

A look at the z-score rating distribution also suggests something interesting. Figure 3 is a density plot that shows the z-scored ratings of all the experimental items across the experiments. There appears to be a bimodal distribution for both experiments in the case of long distance extraction. Since this figure is based off the z-scored ratings, zero here indicates deviation from the mean for each group. Therefore, the bimodality suggests that there are a fair amount of ratings that are higher than average and lower than average for the cases of long distance extraction. For the cases of short distance extraction, they appear to be consistently higher than average. These two patterns are observed for both the definites and the indefinites.

This bimodality could be caused by several things. First, perhaps particular items in the experiment could be more acceptable than others. If item-wise variability was driving this bimodal pattern, then we would expect the difference-of-differences score by item in E1a to be predictive of their ratings in E1b, since they share lexical material. To check this, a correlation between the DD scores of each item for the two experiments was done. The resulting Pearson correlation was 0.22, and regressing the DD values onto each other was not significant. While this could just reflect low power, to some degree it suggests that the bimodality is not due to item-specific factors. If the item bias was present, it should still show up despite this change. The subjects themselves could also be bimodal. That is, there could be some participants who consistently rate the island conditions above the average and those that do not. To investigate this possibility, a histogram of each participant's z-scored rating for each condition was plotted. If subjects are patterning bimodally, then the histograms (Figure 4) should match the density plot. This does not appear to be the case.



(a) 1A: Top - Definite; Bottom - Indefinite

(b) 1B: Top - Definite; Bottom - Indefinite

Figure 4: Subject Means for Long Distance Extraction

Another key aspect is that bimodality is found in both the definite and indefinite DPs. Speakers are also not consistently rating the indefinites at average or higher in the long distance extraction cases. While some of this could be due to the fact that, as seen in both Experiments 1a and 1b, distance does seem to greatly impact acceptability, it is still interesting to observe this in what are considered grammatical sentences. The fact that neither item- nor subject-wise variability appears to be the source, this raises several questions about the true nature of the bimodality.

3. Experiment 2: Self-Paced Reading. Experiment 2 is an online measure attempting to address how the parser handles *the*-DP islands in real time. It is a self-paced reading experiment that uses a filled-gap effect paradigm, discussed in Section 1.2. Under this design, if participants are actively positing gaps inside the definite islands we should expect to see a slowdown in reading times if a DP-internal gap position is filled. An example stimulus is below.

(9) a. The journalist guessed **who** Charlie promoted $\__A$ a/the ridiculously scandalous photo of $\__B$ Einstein and Max Planck to _.

b. The journalist guessed **that** Charlie promoted a/the ridiculously scandalous photo of Einstein and Max Planck to the scientific magazine.

There are two possible gap sites for each item. The first, labeled A above, comes directly after the embedded verb, "promoted" (e.g., *The journalist guessed who Charlie promoted)*. The second, labeled B, is inside the DP, after "of" (e.g., *The journalist guessed who Charlie promoted a ridiculously scandalous photo of*). Gap A is a baseline to determine is participants are actively positing gaps in general; we predict no effect of definiteness here. Gap B investigates whether participants are positing gaps inside the DP; there should be an observed definiteness effect here if definite islands pattern like other islands.

Participants. 45 American English speakers all recruited through the Prolific platform. They were paid $5 for their participation.

Materials. The stimuli, (9), for the self-paced reading experiment are identical to those used in Experiment 1 with a few adjustments to create a filled-gap effect. The name inside the DP was extended with a coordination to allow for spillover, and a continuation was added to provide a grammatical gap site. This resulted in a 2x2 within-subjects design with factors of DEFINITENESS: INDEFINITE AND DEFINITE and PHRASE TYPE: WHO AND THAT. There were a total of 24 items, distributed in four Latin-Squared lists. There were a total of 40 filler items, not including 4 practice items to get participants accustomed to the task. A quarter of the fillers mirrored the experimental items, and the remaining filler items were unrelated.

Procedure. The experiment was run on IbexFarm using a word-by-word self-paced reading paradigm. Participants were given detailed instructions on how to complete the experiment and asked a series of comprehension questions to ensure they were paying attention. After four practice items, the experiment began. Participants were presented with a "+" for 1500 ms after which the screen changed to become the first word of the sentence. They then had to press the spacebar to move throughout. Once they had completed the entire sentence, they were asked a yes/no question and used the keyboard to answer. The instructions and practice items emphasized the importance

of reading at a natural pace and trying to understand the sentences in an attempt to ensure participants were paying attention. After the experiment ended, participants were asked to rate a single set of sentences similar to the experimental ones with the binary options of "good" or "bad". They were taken to a set of post-experiment screening questions, and the experiment concluded.

Analysis. The data were pre-processed to remove participants who failed to maintain an accuracy of 60% or above on the instruction questions as in Experiment 1. Participants were also removed if in a post-experiment question their responses to an open-ended question indicated they might be a bot. We also screened for compliance with the experimental instructions to ensure participants were reading the sentences in a word-by-word manner, as opposed to simply 'clicking through' at a fixed pace. To do this, two regression models were fit to each participant's log-transformed RT data. One had only an intercept (the null model), and the other model used region as a predictor. If participants are varying the speed at which they move through the sentence, the the region model should perform better. These two models were fit for each participant and a likelihood ratio test was performed. If we were not able to reject the null model at $\alpha = 0.2$, then the participant was removed from further analysis.

3.1. RESULTS. For both gap sites, A and B, there were no significant effects. The spillover region also did not find any significant results.

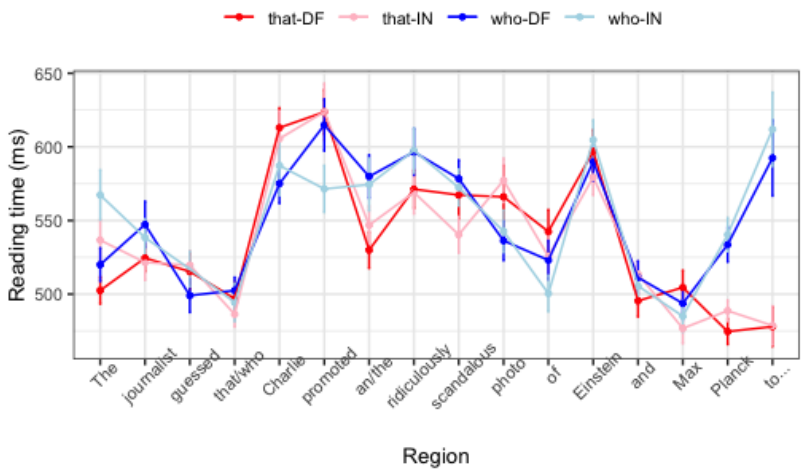


Figure 5: Experiment 2 Reading Times (ms)

	1st Gap				2nd Gap			
	Critical		Spillover		Critical		Spillover	
	β	p	β	p	β	p	β	p
PHRASE TYPE	-0.04	0.08	-0.03	0.19	-0.013	0.58	-0.003	0.86
DEFINITENESS	-0.01	0.53	-0.02	0.45	0.003	0.87	-0.02	0.33
PHRASE:DEFINITENESS	-0.05	0.25	0.03	0.47	0.014	0.73	-0.02	0.53

3.2. EXPERIMENT 2 DISCUSSION. No reliable filled-gap effect was observed at either position. In the first gap, there was a marginal main effect of PHRASE TYPE (p=0.08). In the second, we fail to see any significant effects. The initial gap was meant as a baseline to ensure that we could

measure a filled-gap effect with our materials. Figure 5 suggests there is a numerical trend towards a filled-gap effect, though it is not reliable in our analysis. This could reflect low statistical power. Turning to the gap site inside the DP, Gap B, we saw no clear differences in the mean reading times across any of the four conditions. Participants, it would appear, do not seem to be positing gaps inside definite or indefinite DPs.

This observation raises several questions. First, is whether a filled-gap effect is at all present inside a DP; the current results yield no evidence for this. Second, is whether this filled-gap effect would be modulated by definiteness. This, once again, was also not found in the current study. However, the early marginal effect in the initial gap position could indicate that our failure to find an effect is due to a lack of power. Increasing the power could further clarify both of the above questions. If the reading times remain equal across all four conditions at the DP-internal gap site, this would be evidence to support a claim that participants are not attempting to posit gaps inside DPs.

In sum, this present study finds no significant evidence of active gap filling inside DPs, definite or indefinite. This is somewhat in line with the findings of Tollan & Heller (2015); they find that there is no effect of definiteness between *who* and *which NP*. They do, however, observe a filled-gap effect inside the *which NP* condition which is not modulated by definiteness. It is also unclear if the extraction cases (i.e., *who*-phrases and *which NP*-phrases) differed significantly from the non-extraction cases (i.e., *yes-no* questions).

4. Conclusions and Future Work. Definite islands present varied behavior across offline and online experiments. The offline results suggest a weak island effect that is smaller than other observed islands in the experimental syntax literature. It also reveals a bimodal distribution of judgment where participants show varied ratings for both definite and indefinite DP extractions. This bimodality does not appear to be driven by item- or subject-wise variability. In a reading time study, we failed to find evidence for active gap filling inside the DP, making it difficult to test whether definiteness restricts gap filling inside a DP. Taken together, the picture that this work presents is that definite islands, and DPs in general, consist of greater nuance than originally thought at both an offline and online level.

References

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

Chomsky, Noam. 1973. Conditions on transformations. In S Anderson & P Kiparksy (eds.), *A Festschrift for Morris Halle*, 232–286. New York: Holt, Rinehart & Winston.

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

Davies, William D. & Stanley Dubinsky. 2003. On Extraction From NPS. *Natural Language & Linguistic Theory* 21(1). 1–37. 10.1023/A:1021891610437. https://doi.org/10.1023/A:1021891610437.

Drummond, Alex. 2020. Ibex Farm. https://spellout.net/ibexfarm/.

Erteschik-Shir, Nomi. 1973. Discourse Constraints on Dative Movement. *Syntax and Semantics* 12.

Goldberg, Adele E. 2013. Backgrounded constituents cannot be "extracted". In Jon Sprouse & Norbert Hornstein (eds.), *Experimental Syntax and Island Effects*, 221–238. Cambridge: Cambridge University Press. 10.1017/CBO9781139035309.012. https://www.cambridge.org/core/product/identifier/9781139035309%23c00870-943/type/book-part.

Hofmeister, Philip & Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language* 86(2). 366–415. 10.1353/lan.0.0223. http://muse.jhu.edu/content/crossref/journals/language/v086/86.2.hofmeister.html.

Huang, Nick. 2018. The bound possessor effect: a new argument for the phasehood of definite DPs [manuscript]. In *North East Linguistic Society 48*, .

Keller, Frank. 2000. *Experimental and Computational Aspects of Degrees of Grammaticality*: University of Edinburgh PhD.

Kush, Dave, Terje Lohndal & Jon Sprouse. 2019. On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language* 95(3). 393–420. 10.1353/lan.2019.0051. https://muse.jhu.edu/article/733277.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Douglas Bates. 2017. Balancing type i error and power in linear mixed models. *Journal of Memory and Language* 94. 305 – 315. https://doi.org/10.1016/j.jml.2017.01.001. http://www.sciencedirect.com/science/article/pii/S0749596X17300013.

McElree, Brian, Stephani Foraker & Lisbeth Dyer. 2003. Memory structures that subserve sentence comprehension. *Journal of memory and language* 48(1). 67–91. ISBN: 0749-596X Publisher: Elsevier.

Phillips, Colin. 2006. The Real-Time Status of Island Phenomena. *Language* 82(4). 795–823. 10.1353/lan.2006.0217. http://muse.jhu.edu/content/crossref/journals/language/v082/82.4phillips.pdf.

R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Ross, John Robert. 1967. *Constraints on Variables in Syntax*: MIT PhD. https://eric.ed.gov/?id=ED016965.

Simonenko, Alexandra. 2015. Semantics of DP Islands: The Case of Questions. *Journal of Semantics* 33. ffv011. 10.1093/jos/ffv011.

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

Stowe, Laurie A. 1986. Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes* 1(3). 227–245. 10.1080/01690968608407062. https://doi.org/10.1080/01690968608407062. Publisher: Routledge eprint: https://doi.org/10.1080/01690968608407062.

Tollan, Rebecca & Daphna Heller. 2015. Elvis presley on an island: wh dependency formation inside complex np objects. In *North East Linguistic Society 46*, https://www.researchgate.net/publication/303306297.