Resumptive Pronouns Can Ameliorate Illicit Island Extractions

Abstract

Syntax literature reports that resumptive pronouns (RPs) ameliorate island violations, but much psycholinguistic literature has found RPs to be no more acceptable than straightforwardly island-violating gaps, in spite of the fact that island production tasks consistently elicit RPs. However, prior psycholinguistic studies have typically compared RP and illicit gap conditions indirectly. We posit that RP island amelioration in comprehension is undetectable when subjects cannot engage in comparison of alternative sentences, and thus that the apparent production-comprehension split arises from methodological differences between perception and production experiments.

We present six experiments crossing three island types in two tasks (full sentence forced choice and forced choice fill-in-the-blank), manipulating gap location (Island vs Nonisland). We find that RPs are preferred in Islands and gaps in Nonislands (p=0.0001). This suggests that RPs do ameliorate island violations and that the production-comprehension split is a methodological artifact.

Keywords

resumptive pronouns, island violations, acceptability judgment, amelioration, forced choice

1. Introduction

Whether or not resumptive pronouns (RPs) can ameliorate island violations in English is a controversy that divides the literature into two camps: studies rooted in theoretical syntax, and studies rooted in psycholinguistics. In the syntax literature, it has been claimed that RPs can "repair" island violations (Chomsky 1986, Kroch 1981, McDaniel & Cowart 1999, Creswell 2002, Cf. Asudeh 2011). Thus a *wh*-RP dependency across a syntactic island is said to be more acceptable than a corresponding *wh*-gap dependency across the island, as illustrated in (1).

(1) a. * Which man$_i$ did Jane say that [the parent who scolded ___$_i$] forgave the babysitter's mistake?

b. ? Which man$_i$ did Jane say that [the parent who scolded him$_i$ forgave the babysitter's mistake?

On the other hand, much of the psycholinguistics literature has found no acceptability contrast between the two types of dependencies in English (Alexopoulou & Keller 2007, Omaki & Nakao 2010, Heestand et al. 2011, Han et al 2012, Polinsky et al. 2013, a.o.), possibly because RPs are "intrusive" in English, and consequently not grammatically licit as they are in languages such as Irish and Palestinian Arabic (Sells 1984, Erteschik-Shir 1992, Prince 1990, Schlonsky 1992, a.o.). Studies on RPs in English have used Likert scale and Magnitude Estimation tasks to measure relative acceptability of these constructions in a variety of different island contexts. Curiously, the psycholinguistic literature has frequently found that RPs do not repair island violations in the way predicted by the syntax literature.

Against this background, the current study explores whether these distinct and conflicting observations are due to differences in the type of decision probed in the different judgment tasks in the syntax and psycholinguistics literature. In order to address this question, we conducted a series of controlled experiments similar to the typical minimal pair comparison used informally in syntactic judgment studies; namely a binary forced-choice experiment (Myers 2009, Sprouse et al. 2013). This methodological choice allows us to gather empirical observations over a large sample size as was done for the scalar rating tasks, while at the same time using a decision format more similar to that used to make the original claims of amelioration.[1] Our aim for these experiments is to serve as a connection between the syntax and psycholinguistics literatures, demonstrating the ameliorative properties of resumptive pronouns on island violations in several syntactic contexts and briefly speculating why these results were not observed in psycholinguistic studies that did not afford subjects an opportunity to compare island-violating dependencies with RPs vs. with gaps.

2. Resumptive pronouns

Previous studies have reported that RPs can salvage island violations (Ross 1967, Kroch 1981, McKee & McDaniel 2001), ease processing difficulty (Prince 1990, Asudeh 2004), and/or increase referent accessibility (Ariel 1999). Among others, Asudeh (2004) emphasizes that RPs reduce processing cost and tend to occur in long, complex sentences and when substantial additional material intervenes between an extracted noun and the tail of its dependency (e.g., Hofmeister & Norcliffe 2013, Erteschik-Shir 1992, Dickey 1996).

Many psycholinguistic studies in English have addressed the relative acceptability of RPs in island contexts, although there is no clear consensus on the role that RPs may play in the production or comprehension process (Cf. Ferreira & Swets 2005, Han et al. 2012, Clemens et al. 2012, Asudeh 2004, Asudeh 2011). Various acceptability-rating studies, which used tasks such as Likert scale rating (Heestand, Xiang & Polinsky 2011, Polinsky et al. 2013) and magnitude estimation (Alexopoulou & Keller 2007, Omaki & Nakao 2010), have shown no difference in acceptability between gaps and RPs across islands. On the other hand, Ferreira & Swets (2005) have shown that in island violation contexts, sentences with RPs were produced more often than sentences with gaps, even when speakers were given time to plan their utterances.

It has been suggested that at least in otherwise highly taxing environments, such as in deeply embedded sentences, intrusive RPs serve as a sort of (ungrammatical but informative) signpost for the processing mechanism (Hofmeister & Norcliffe, 2013). In other words, an RP can, by overtly expressing the tail of a *wh*-dependency, relieve some of the cognitive load on the speaker's limited resources in situations where incremental production (or potentially comprehension) would benefit from an overt link to the position that triggered the dependency search (Ferreira & Swets 2005). This argument has typically been made in reference to RPs located in positions that grammatically permit a gap, but such a mechanism may also be consistent with a limited local comprehension mechanism (Beltrama & Xiang 2016; Asudeh, 2004).

3. Experiments

One previously unexplored explanation why RPs show such divergent effects is that the difference in acceptability between RP-sentences and gap-sentences is normally quite subtle in terms of what rating tasks are able to measure. When native speakers make acceptability judgments for these sentences, they normally report that the difference is unclear but that RP-island-extraction sentences feel somewhat better than gap-island-extraction sentences (Keffala, 2013). It is possible that the island-amelioration effects of RPs are sufficiently weak that gradient acceptability-rating tasks are not sensitive enough to detect them. To test this possibility, we adopted a binary forced-choice paradigm. In forced-choice (FC) tasks, minimal pairs are presented to the participants, and participants are asked to select from the minimal pair the option that is more acceptable. It has been reported that binary forced-choice tasks are more sensitive than gradient rating tasks for detecting differences among experimental conditions and more statistically powerful than other tasks (Myers 2009, Sprouse & Almeida 2014, Sprouse, Schütze & Almeida 2013). This may be due to the forced nature of the task; subjects cannot choose to select neither option (or both options) or abstain from choosing. This feature of the binary forced-choice task does not, however, preclude subjects selecting an option at random, which could produce a split between the two options at chance levels.[2] In this way, there are three logical outcomes: (1) the first option could be chosen more frequently than the second, (2) the second option could be chosen more frequently than the first, or (3) there could be no statistical difference in choices. Because of the limited number of outcomes in this paradigm, it should give us a maximal chance of detecting subtle contrasts that may have been obscured by more gradient acceptability rating analyses (Sprouse, Schütze & Almeida, 2013).

We conducted two types of FC experiments. One type is a full-sentence forced-choice task (hereafter FC), in which two sentences were presented in a pair and participants were asked to choose which of the two was more acceptable. The other task is a binary-choice fill-in-the-blank

(FiB) task, in which two minimally different phrases were presented below a sentence with an underlined blank. Participants were instructed to choose which phrase best fills in the blank to complete the sentence. The use of both FC and FiB tasks in this study allows the possibility of within-study replication, since each set of stimuli was presented to two non-overlapping groups, each participating in one of the two tasks. Thus, if we find similar results in both tasks (across participants), this result can strengthen the conclusions drawn.

3.1 Design and Stimuli

We conducted three binary FC experiments and three binary FiB experiments, each pair of which tested the following island environments: Relative Clause islands (RC islands) (Creswell 2002; Han et al. 2012; Heestand et al. 2011; Polinsky et al. 2013), adverbial clause adjunct islands (Adjunct islands) (Heestand et al. 2011; Polinsky et al. 2013), and *wh*-islands (McDaniel & Cowart 1999; Alexopoulou & Keller 2007). The experiments manipulated the factor of Island (Island versus Non-island). In the Island condition, the gap or RP was located in an object position within an island, which is an illicit extraction site. In contrast, in the Non-island condition, the gap or RP was not located inside an island, and was in an object position close to the end of the sentence (a licit extraction site). For example, in the RC island in (2), the gap or RP is embedded inside the relative clause in (2a) and in a non-island environment (such as the matrix clause in this case) in (2b):

(2) a. Island

Which woman did Carlos report that [$_{\text{island}}$ the newscaster who exposed **her/ø**] threatened the detective's case?

b. Non-island

Which woman did Carlos report that [$_{\text{island}}$ the newscaster who exposed the criminal] threatened **her/ø**?

In the FC experiment, the two sentences were presented together, as illustrated in (3), and the participants were asked to choose the more acceptable sentence of the pair.

(3) Which woman did Carlos report that the newscaster who exposed threatened the detective's case?

Which woman did Carlos report that the newscaster who exposed her threatened the detective's case?

In the FiB task, only one sentence is presented, containing a blank indicated by an underline. In this experiment, the blank always corresponds to the verb phrase. Directly below the sentence, two phrases were presented as in (4):

(4) Which woman did Carlos report that the newscaster who ___ threatened the detective's case?

exposed exposed her

Participants were asked to choose which of these two phrases best fit the incomplete sentence.

For each of the six experiments, 80 unique subjects with I.P. addresses restricted to the United States were recruited through Amazon Mechanical Turk (MTurk). Each experiment (FiB and FC formats for the three island types) consisted of a total of 108 items distributed in two counterbalanced lists, with twelve target items and 96 fillers. Each item displayed a minimal pair of sentences (e.g., Gap vs RP) from which the participant would select one, as shown in (3) and (4). In each pair of experiments testing the same type of island, the content of the items was identical. The fillers were taken from unrelated experiments, none of which involved islands or resumptive pronouns. Since forced-choice tasks explicitly reveal to the participant the crucial manipulation, some of these fillers were designed to obscure the purpose of the manipulations by using superficially similar alternations that did not interfere with the target stimuli of this study. The sample set of stimuli is summarized in (5) through (7). The counterbalanced lists ensured each participant would only see one version of each item, either (a) or (b), thus a participant saw six tokens from each of the two conditions (Island and Nonisland).

(5) RC island

a. Island condition

Which woman did Carlos report that the newscaster who exposed threatened the detective's case?

Which woman did Carlos report that the newscaster who exposed her threatened the detective's case?

b. Non-island condition

Which woman did Carlos report that the newscaster who exposed the criminal threatened?

Which woman did Carlos report that the newscaster who exposed the criminal threatened her?

(6) Adjunct island

a. Island condition

Which woman did Carlos report that, when the newscaster exposed, the criminal threatened the detective's case?

Which woman did Carlos report that, when the newscaster exposed her, the criminal threatened the detective's case?

b. Non-island condition

Which woman did Carlos report that, when the newscaster exposed the detective's case, the criminal threatened?

Which woman did Carlos report that, when the newscaster exposed the detective's case, the criminal threatened her?

(7) *Wh*-island[3]

a. Island condition

Which woman did Carlos question how the newscaster exposed?

Which woman did Carlos question how the newscaster exposed her?

b. Non-island condition

Which woman did Carlos report that the newscaster exposed?

Which woman did Carlos report that the newscaster exposed her?

What is predicted for these manipulations? First, we consider a hypothesis derived from the bulk of the literature that uses acceptability rating tasks: RPs in object positions do not ameliorate island violations because the acceptability of RP-sentences and gap-sentences in island-violating contexts do not differ. In this case, we expect that readers will not display a preference for the sentences with an RP in an island. Consequently, participants' selection of one option or another will be at chance regarding the experimental manipulation and will reflect the lack of amelioration seen in previous acceptability rating studies. This is because, under this hypothesis, the acceptability of the sentence is not affected by the choice of an RP or a gap. Second, if the RP's presence is sensitive to island contexts, i.e., RPs are employed specifically to ameliorate island violations and are not acceptable outside of this context, we expect RPs should be chosen significantly less often than gaps in the Non-island conditions, where gaps are licit. However, if RPs are sensitive to the complexity or other cognitively costly property of the intervening sentence regardless of whether an island is violated or not, then we expect a high rate of RP selection even in Non-island conditions. Since gaps are licit in the Non-island condition, we do not necessarily expect RP selection to exceed gap selection in this case, although it is a possibility. If RPs are a preferred method of recalling the tail of a long dependency (Hofmeister & Norcliffe, 2013), then they may be preferred to gaps at the end of a highly complex sentence.

3.2 Results and Discussion

Results were analyzed using logistic regression, with random slopes of item and subject (binomial general linear mixed models fit by Laplace approximation, with maximally expanded random effects structures that allowed convergence)[4] (Barr et al., 2013; Bates et al., 2015). Binomial data from the forced-choice selection were coded with a 0 for a selection containing a gap and 1 for a selection containing an RP. These data were then centered on 0 in order to prevent miscalculations due to anomalies introduced by the character 0 (Agresti, 2002). The fixed effect was the condition (Island or Non-island), with random intercepts of subject and item. It may be noted that forced-choice paradigms are shown to consistently distinguish between two options which are similarly shown to differ in scalar rating paradigms (Sprouse, Schütze & Almeida 2013). Additionally, a 1-sample proportions test with continuity correction was used to determine whether the proportion of RP or gap selections chosen within a given condition was significantly different from chance, with the test value set at .5[5]. This analysis should determine whether a preference for one option was consistently expressed over the other, or whether no preference was expressed. The rate of RP/gap selections in each experiment is summarized in Table 1, and the statistical analyses of the results are summarized in Table 2.

Table 1.

Figure 1.

Table 2.

The analysis revealed that there was a main effect of Island in all three experiments, indicating that RPs were selected at a higher rate in Island conditions than in Non-island conditions, overall. In fact, RPs were chosen significant more often than gaps in the Island condition in all experiments, and gaps were chosen significantly more often than RPs in the Non-island conditions, except in the Adjunct Island experiment. In the Adjunct Island FC experiment, RPs were chosen significantly more often than gaps across the board, although there was still a distinct preference for RPs in the Island condition compared with the Non-island condition. The same overall pattern can be seen in the Adjunct Island FiB experiment, but the difference between RP and gap selection in the Non-island condition was not significant, with RPs and gaps selected approximately in equal proportions. The result of the 1-sample proportions test shows that all other conditions displayed significant differences in RP and gap selection and thus that these differences in preference are reliable across island types and task types. That is, subjects were consistently expressing a preference for RPs as compared to gaps in islands across the three island contexts and two task types.

4. General Discussion

There are four main findings from our series of studies. First, the consistent preference for RPs in island contexts across all experiments indicates that RPs do indeed improve island-violating sentences compared to the corresponding gaps. This is consistent with the island amelioration effects reported in the syntax literature. Second, that a strong preference for RPs was seen in the island contexts in both FC and FiB experiments supports our claim that RPs can ameliorate island violations in comprehension tasks[6], contrary to the suggestions by (Han, et al., 2012; Ferreira & Swets, 2005; Heestand, et al., 2011). Third, the observation that gaps were not preferred in the Non-island context in the Adjunct Islands experiments is compatible with the claim that RPs appear at the tails of long dependencies or in deeply embedded structures (e.g., Dickey 1996, Asudeh 2011). However, that a similar dispreference for gaps was not shown in the Non-island conditions in the *wh-* and RC Islands experiments could be problematic for this explanation, unless the center-embedded structure created by the adjunct clause somehow specifically triggers the process by which non-island RPs are produced (and sometimes accepted). If this is the case, the absence of a preference for gaps in the Non-island condition of the Adjunct Island experiment could be attributed to the relatively increased complexity of the material that intervenes between the extracted materials and the (licit) gap, or possibly due to the embedded nature of the adjunct clause (Cf. Hofmeister & Norcliffe 2013, for RPs in center-embedded contexts).[7] Fourth and finally, the result of these experiments overall suggests that RPs may indeed improve island-violating sentences, contrary to the conclusions of many previous acceptability-rating studies. In the FC tasks, specifically, the participants were asked to choose the sentence that is more acceptable to make the task as similar as possible to more traditional acceptability judgments. Furthermore, as we have discussed, it is possible that the RP island-amelioration effect is subtle or weak. The fact that a clear effect of island-amelioration can be seen in the FC and FiB experiments, despite the paucity of positive results in more gradient comprehension tasks and the sensitivity and statistical power of forced-choice tasks, supports this position (Myers 2009; Sprouse, Schütze & Almeida 2013).

Although resumptive pronouns may not rescue island violations completely, the results of the reported experiments strongly indicate that they do improve island-violating sentences. In what follows, we discuss the possible interpretations of such results.

Our study finds a substantial and reliable preference for RPs over gaps in islands, which is problematic for the claim that object-extracted RPs are always unacceptable in English. How can an RP be as unacceptable as an object-extracted illicit gap, yet be preferred to the corresponding gap?[8] One possibility is that RPs do improve island violations, but RPs and gaps have indistinguishably similar acceptability ratings. This could be due to a floor effect in which these types of ungrammatical sentences are rated as the lowest possible acceptability given an otherwise well-formed sentence. Alternatively, this could be due to a ceiling effect in which sentences with island violations are never rated higher than a certain value, so even though RPs are perceived as better than gaps, their ungrammaticality prevents participants from expressing this nuanced reaction. Another way of conceptualizing this possibility is that RPs do show the rescuing (or otherwise ameliorating) effect for island violations, as suggested by some previous studies (McDaniel & Cowart 1999, Creswell 2002). However, the effects of RPs is not enough to be detected in gradient rating studies, in which the ratings are compressed in the lowest part of the scale, even in more flexible tasks such as magnitude estimation. That is, a ceiling effect may be preventing any distinction from obtaining between the gap and RP conditions in studies such as Heestand, et al. (2011) and the works cited therein, in conjunction with a floor effect preventing a more nuanced distinction between different flavors of ungrammaticality. It's not clear how fruitful a direction of inquiry into distinguishing different "types" of unacceptable ratings would be for the question at hand, since that question is no longer directed at the syntax or processing of resumptive pronouns, but rather at what acceptability rating tasks are capable of detecting on a fundamental level.

Another possibility is that the observed unacceptability of RPs is due to their ungrammaticality, but some influence of processing increases their acceptability as compared to gaps in island contexts. In other words, if one must (ungrammatically) extract material from an object position within an island, an RP is better than a gap for both the speaker and listener. On the other hand, this is potentially incompatible with theories of processing that account for the production of RPs by (local) evaluation mechanisms that allow for productions of RPs but reject RPs in (global) analysis (e.g., Dickey 1996, Asudeh 2011, Beltrama & Xiang 2016). Given these types of theories, it is not clear how RPs could elicit a consistent and reliable preference over gaps during reading-based tasks. This is because the mechanism that prevents their acceptability should not be able to distinguish between gaps and RPs in terms of their ungrammaticality during comprehension. That is, under this hypothesis, the robust preference for RPs in island contexts we have reported is unexpected.

5. Conclusion

Contra claims that RPs have no ameliorative effect in comprehension and only serve a processing purpose in production, we detect a strong preference for RPs in islands in both our FC task and FiB task. Requiring subjects to choose a sentence in each pair, regardless of "goodness" relative to some fully, incontrovertibly grammatical sentence may have magnified the relatively weak acceptability differences between gaps and RPs within island violating contexts. In sum, our observations refute the claim that RPs are as dispreferred as gaps in island contexts. In conjunction with other recent studies, however, we suggest that differences in the types of tasks used to obtain readers' judgments may account for this discrepancy in the literature.

Acknowledgements

We are grateful to Peter Baumann for his valuable comments on an earlier version of this work and the anonymous reviewers for their valuable insights and suggestions. This work has been supported in part by NSF grants BCS-1323245 and BCS-1348677 awarded to Masaya Yoshida.

References

Alexopoulou, Theodora, & Frank Keller. 2007. Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language* 83:110– 160.

Ariel, Mira. 1999. Cognitive universals and linguistic conventions: The case of resumptive pronouns. *Studies in Language* 23:217–269.

Arppe, Antti, & Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory*. Volume 3(2):131–159.

Asudeh, Arshia. 2004. *Resumption as resource management*. Doctoral dissertation, Stanford University, Palo Alto, CA.

Asudeh, Ash. 2011. Local Grammaticality in Syntactic Production. In *Language From a Cognitive Perspective: Grammar, Usage, and Processing*, ed. by Emily Bender & Jennifer Arnold, 51-79. Stanford: CSLI Publications.

Barr, Dale J. 2013. Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, 4, 328.

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth, & Harald Baayen. 2015. Parsimonious mixed models. *arXiv preprint arXiv*:1506.04967.

Beltrama, Andrea & Ming Xiang. 2016. Unacceptable but comprehensible: the facilitation effect of resumptive pronouns. *Glossa* 1(1):29, 1-24.

Chomsky, Noam. 1986. *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Clemens, Laurel E., Adam A. Morgan, Maria Polinsky, & Ming Xiang. 2012. *Listening to resumptives: An auditory experiment*. Paper presented at the 25th Annual CUNY Conference on Human Sentence Processing (CUNY 25), City University of New York, 14–16 March.

Creswell, Cassandre. 2002. Resumptive pronouns, *wh*-island violations, and sentence production. In *Proceedings of the Sixth International Workshop on Tree-Adjoining Grammar and Related Frameworks* (TAG+6), 101–109. Università di Venezia.

Dhar, Ravi, & Itamar Simonson. 2003. The effect of forced choice on choice. *Journal of Marketing Research*, 40(2), 146-160.

Dickey, Michael Walsh. 1996. Constraints on the sentence processor and the distribution of resumptive pronouns. In *University of Massachusetts Occasional Papers in Linguistics 19: Linguistics in the Laboratory*, ed by M. Dickey and S. Tunstall. Amherst, MA: GLSA Publications.

Erteschik-Shir, Nomi. 1992. Resumptive pronouns in islands. In *Island constraints: Theory, acquisition, and processing*, ed. by Helen Goodluck and Michael Rochemont, 89–108. Dordrecht: Kluwer.

Ferreira, Fernanda, & Benjamin Swets. 2005. The production and comprehension of resumptive pronouns in relative clause "island" contexts. In Anne Cutler (ed.), *Twenty-first Century psycholinguistics: Four cornerstones*, 263-278. Mahway, NJ: Lawrence Erlbaum Associates.

Han, Chung-hye, Noureddine Elouazizi, Christina Galeano, Emrah Görgülü, Nancy Hedberg, Jennifer Hinnell, Meghan Jeffrey, Kyeong-min Kim, & Susannah Kirby. 2012. Processing strategies and resumptive pronouns in English. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, 153–161.

Heestand, Dustin, Ming Xiang, & Maria Polinsky. 2011. Resumption still does not rescue islands. *Linguistic Inquiry* 42:138–152.

Hofmeister, Philip, & Elisabeth Norcliffe. 2013. Does resumption facilitate sentence comprehension. In Philip Hofmeister & Elisabeth Norcliffe (eds.), *The Core and the Periphery: Data-Driven Perspectives on Syntax Inspired by Ivan A. Sag*. Stanford, CA: CSLI Publications.

Keffala, Bethany. 2013. Resumption and gaps in English relative clauses: Relative acceptability creates an illusion of 'saving'. In *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, ed. by Chundra Cathcart, I-Hsuan Chen, Greg Finley, Shinae Kang, Clare S.

Sandy, and Elise Stickles, 140-154. Berkley: University of California Berkley, Berkley Linguistic Society.

Kroch, Anthony. 1981. On the role of resumptive pronouns in amnestying island constraint violations. In *Papers from the 17th Regional Meeting, Chicago Linguistic Society*, ed. by Robert A. Hendrick, Carrie S. Masek, and Mary Frances Miller, 125–135. Chicago: University of Chicago, Chicago Linguistic Society.

McDaniel, Dana, & Wayne Cowart. 1999. Experimental evidence for a minimalist account of English resumptive pronouns. *Cognition*, 70:B15–B24.

McKee, Cecile, & Dana McDaniel. 2001. Resumptive Pronouns in English Relative Clauses, *Language Acquisition*, 9:113–156.

Myers, James. 2009. Syntactic judgment experiments. *Language and Linguistics Compass*, 3:406–423.

Omaki, Akira, & Chizuru Nakao. 2010. Does English resumption really help to repair island violations? *Snippets* 21:11-12.

Polinsky, Maria, Laurel E. Clemens, Adam M. Morgan, Ming Xiang, & Dustin Heestand. 2013. Resumption in English. In Jon Sprouse & Norbert Hornstein (eds.), *Experimental Syntax and Island Effects*, 341–359. New York: Cambridge University Press.

Prince, Ellen. 1990. Syntax and discourse: A look at resumptive pronouns. In *Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society: Parasession on the Legacy of Grice*, ed. by Kira Hall, Jean-Pierre Koenig, Michael Meacham, Sondra Reinman, and Laurel A. Sutton, 482–497. Berkeley: University of California, Berkeley Linguistics Society.

Ross, John Robert. 1967. *Constraints on variables in syntax*. Doctoral dissertation, MIT, Cambridge, MA. [Published as Infinite syntax! Norwood, NJ: Ablex, 1986.]

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Schütze, Carson T., & Jon Sprouse. 2014. Judgment data. In Robert Podesva & Devyani Sharma (eds.), *Research Methods in Linguistics*, 27–50. New York: Cambridge University Press.

Sprouse, Jon, Carson T. Schütze, & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001– 2010. *Lingua*, 134:219–248.

Tables and Figures

Table 1: Proportion of RP selections by island type

Forced Choice	*Conditions*	*RP%*	*CI (95%)*
RC island	Island	82	77–87
	Non-Island	38	32–44
Adjunct island	Island	79	73–84
	Non-Island	57	50.4–63
Wh-island	Island	63	56–69
	Non-Island	13	9–18
FiB			
RC island	Island	80	74–85
	Non-Island	25	19–31
Adjunct island	Island	80	74–85
	Non-Island	51	44–57
Wh-island	Island	60	53–66
	Non-Island	11	7–16

Figure 1: Proportion of RP/Gap selections for each island type and task type.

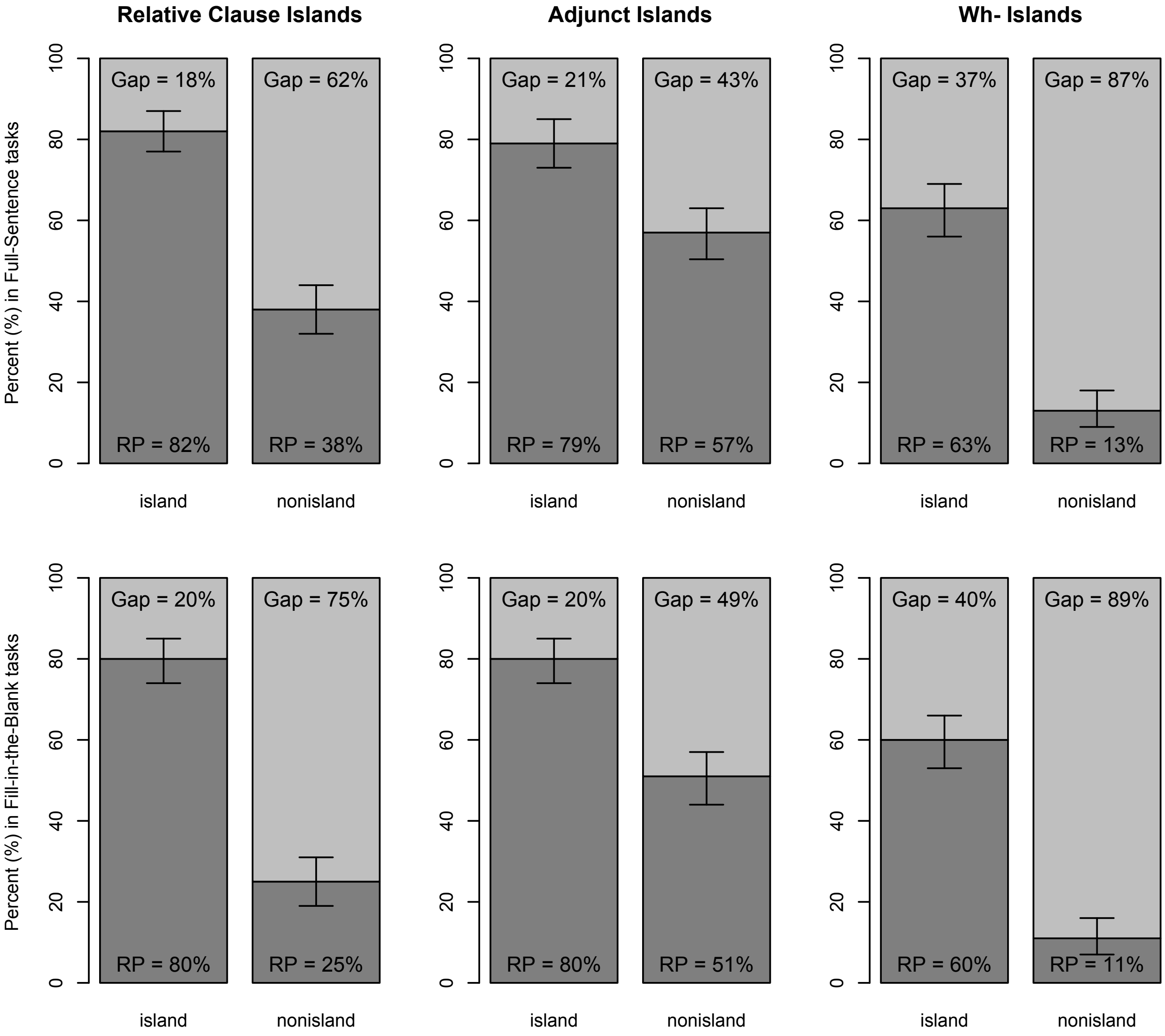

Table 2: Summary of analyses

Forced Choice	*Main Effect of Island*	*Condition*	*Chance level of performance?*
RC island	$\beta=-0.44$, SE=0.04, $\chi^2(1)=126.3$, p=0.0001	Island	No: $\chi^2(1)=97.5$, p=0.0001, 95% CI = 0.77–0.87
		Non-Island	No: $\chi^2(1)=13.54$, p=0.001, 95% CI = 0.32–0.44
Adjunct island	$\beta=0.22$, SE=0.04, $\chi^2(1)=34.1$, p=0.0001	Island	No: $\chi^2(1)=77.22$, p=0.0001, 95% CI = 0.73–0.84
		Non-Island	No: $\chi^2(1)=4.28$, p=0.038, 95% CI = 0.504–0.63
Wh-island	$\beta=-0.49$, SE=0.03, $\chi^2(1)=168.4$, p=0.0001	Island	No: $\chi^2(1)=14.5$, p=0.0001, 95% CI = 0.56–0.69
		Non-Island	No: $\chi^2(1)=128.7$, p=0.0001, 95% CI = 0.09–0.18
FiB			
RC island	$\beta=-0.55$, SE=0.04, $\chi^2(1)=197.5$, p=0.0001	Island	No: $\chi^2(1)=85.2$, p=0.0001, 95% CI = 0.74–0.85
		Non-Island	No: $\chi^2(1)=61.0$, p=0.0001, 95% CI = 0.19–0.31
Adjunct island	$\beta=0.29$, SE=0.04, $\chi^2(1)=56.1$, p=0.0001	Island	No: $\chi^2(1)=84.4$, p=0.0001, 95% CI = 0.74–0.85
		Non-Island	Yes: $\chi^2(1)=0.04$, p>.1, 95% CI = 0.44–057.
Wh-island	$\beta=0.49$, SE=0.03, $\chi^2(1)=169.0$, p=0.0001	Island	No: $\chi^2(1)=9.2$, p=0.002, 95% CI = 0.53–0.66
		Non-Island	No: $\chi^2(1)=145.7$, p=0.0001, 95% CI = 0.07–0.16

[1] Our use of *amelioration* will refer to the effect measured by acceptability judgments, which may be considered distinct from *repair*, which we take to indicate a process whereby a fully ungrammatical construction is altered to become fully grammatical.

[2] Alternatively, other pressures on the participants could result in an unintended bias that could push what "chance" is in one direction or another (Dhar & Simonson 2003).

[3] These stimuli are notably shorter than the others. Although longer (length-matched) stimuli were tested, their results are not included in this study because they did not contribute any new insights to our already consistent results. Furthermore, since longer dependencies are shown to induce RPs more readily than shorter ones (Hofmeister & Norcliffe 2013), any confirmatory

results observed in the short *wh*-island set of experiments is more precisely attributable to the nature of the island, rather than another characteristic of the sentence (i.e., dependency length).

In fact, in a post hoc analysis, we found significant differences in overall proportion of RP selections between experiments, in which the shortest stimuli (*wh*-sentences) had the fewest RP selections, the longest stimuli (adjunct islands) had the most RP selections, and mid-length stimuli (relative clause islands) in the middle (β=0.18, SE= 0.046, $\chi^2(5)$=77.21, p<0.0001). This is consistent with the findings of Hofmeister & Norcliffe (2013).

[4] Since the maximal models did not converge, random effects were removed until the model converged. Incidentally, all LMER models converged in the following template:

lmer(choice ~ condition + (1 | subj) + (1 | item) , data = data).

To determine p-value, this model was compared with an ANOVA to a reduced model with *condition* removed.

[5] In a chi-squared test for homogeneity, an observed proportion in a sample population (e.g., number of RP selections) can be compared to an expected proportion (e.g., chance level is set at 50%). In this case, we compare the observed proportion in one condition to the chance level of 50% and found that in all but one condition, the observed proportion was significantly different from chance. A binomial analysis was performed and reported to compare the RP and Gap conditions for each experiment, so a proportions test supplements our analysis to show that in the vast majority of conditions reported (all but one), we are justified in saying that the subjects were not choosing at random.

The details of how this statistical test is structured can be found here: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prop.test.html

[6] We treat the FiB task as a comprehension task, although it might also have a production component. Specifically, we consider the FiB task as a comprehension task because the participant only gets two choices with which to fill in the blank, so they must read and comprehend both in order to select one.

[7] We cannot resolve this issue with the data at hand, and direct the reader to explanations offered in Dickey's (1996) *Strategic Shunting Hypothesis* and Asudeh's (2011) LFG-based account. But this point is not central to the present study, and we leave it for future investigations.

[8] An anonymous reviewer suggests that preference between two forms and relativity acceptability of those forms are not necessarily related measures. While the authors agree with this observation, we believe treating gradient acceptability ratings and forced-choice preferences as related is a reasonable first step to take in uniting the results of this study with the literature at large. In particular, it is possible that *preference* measures frequency in colloquial language (Arppe & Järvikivi 2007). This may confound questions about grammatical constructions or variability in speaker populations or lexical choice. However, if we follow the extensive literature suggesting that RPs in English are ungrammatical, we must ask *why* a preference would be apparent in a reading task when it has only been reliably detected in language production tasks in the past. Our answer, at least for now, is that a preference detected in a forced choice task *could* be a reflection of the frequency with which it occurs in colloquial language. However, this would have to be squared with the theories of language processing that prevent the parser from coming to a globally coherent analysis of RPs and predict their uniform unacceptability in gradient rating tasks. In other words: why should readers or speakers have a preference here if there is no possibility of a processing benefit?