The amelioration effect of *which* on strong/weak islands in English: an experimental study

Sandra Villata & Jon Sprouse
New York University Abu Dhabi

The islands literature recognizes two classes: strong islands, which prevent extraction of all wh-items, and weak islands, which allow extraction of complex wh-items (*which book*) (e.g. Pesetsky 1987; Szabolcsi & Lohndal 2017). Early reports from informal judgment experiments acknowledged that the amelioration is variable, with some extractions fully acceptable and others less so (e.g. Ross 1967, Pesetsky 1987, Rizzi 1990). Recent formal experiments have shown a *reduction*, but never an elimination, of the island effect (e.g., Goodall 2015, Sprouse et al. 2016). Our first question is empirical: we want to settle the facts by precisely quantifying the effect size for the widest possible range of islands. We tested 20 islands with simple and complex wh-items using the factorial definition of islands. We collected ~200 unique participants per island/wh-type on MTurk using the CloudResearch accepted list (~8000 total). We derived Bayesian posteriors for each island effect. We find that weak islands show partial amelioration (a reduction in the island effect size), but never complete elimination. Our second question is theoretical: the interaction between island types (strong, weak) and wh-types (simple, complex) is a challenge for most existing theories. Impenetrability theories (e.g. Chomsky 2001, Müller 2010) lack a mechanism that renders the edge of the phases unavailable as a function of the wh-type. Discourse-based approaches (e.g. Erteschik-Shir 1973, Abeillé et al. 2020) have no mechanism to restrict the focusing effects of complex wh-phrase to certain islands and not others. Here we focus on featural Relativized Minimality (fRM; Rizzi 2018) as it can explain the interaction of island type and wh-item by positing that weak islands are intervention structures (strong islands are not), and can explain partial amelioration through reduced feature overlap between complex wh-phrases and interveners. Anticipating slightly, we find partial amelioration for wh-islands, which fRM is well-suited to explain. We also observe partial amelioration for 2 out of 4 types of noun complement islands, an island type that fRM is not designed to explain. We also find no amelioration for negative islands, suggesting that negative islands may fall outside of fRM. Our final question is architectural: binary grammatical approaches explain gradience in effect sizes through extra-grammatical mechanisms, like processing differences between sentences. But we can see no independently established processing differences that predict an interaction between wh-types and structures when everything is lexically matched. Though we have chosen to focus on fRM for concreteness, our general conclusion is that the partial amelioration of weak islands requires a grammar-internal explanation.

Experiments: We tested 11 weak islands and 9 strong islands (see Fig. 1). For each island, we constructed 16 lexically matched quadruplets in a 2x2 factorial design crossing STRUCTURE (island, non-island) and DEPENDENCY LENGTH (short, long), illustrated in (1):

1a. Who/Which waiter thought that the chef burned the dish? (short, non-island)
b. Who/Which waiter wondered whether the chef burned the dish? (short, island)
c. What/Which dish did the waiter think that the chef burned? (long, non-island)
d. What/Which dish did the waiter wonder whether the chef burned? (long, island)

The factorial design isolates the effect of the island violation in the interaction of STRUCTURE X LENGTH. Each combination of island type and wh-type (simple, complex) was tested in its own experiment (40 total experiments). Participants rated 2 tokens per condition on a 7-point scale. The experiment consisted of 9 practice items, 8 target items and 14 fillers.

Results. We z-transformed judgments prior to analysis to remove scale bias. We calculated linear mixed-effects models and Bayes factors to see the presence of island effects (Fig.1), and then fit Bayesian linear mixed-effects models to calculate 95% credible intervals for simple and complex wh-items (Fig. 2). Fig.1 shows a statistically significant interaction

indicative of an island effect for both simple and complex wh across all island types except two factive islands. This shows that amelioration is never complete, hence partial at best. Fig. 2 quantifies the amelioration through 95% credible intervals for the island effect size for simple and complex wh-items: we see partial amelioration for 4/5 wh-islands, 1 factive island, and 2/4 noun complement islands, but not negative islands or strong islands.

Discussion. Our most important finding is that the amelioration for complex wh-items is partial. fRM distinguishes among full, partial/reduced, and zero feature overlap, and maps these into levels of grammaticality (Rizzi 2018). This captures the partial amelioration of wh-islands under the assumption that *if, whether,* and *why* share with the extracted wh-item the +Q feature of question operators. When the extracted element is complex it also bears an additional +N feature, reducing the feature overlap and improving acceptability. The cases of *who* and *which* may involve double-name penalty confounds: for *who*, it decreases the simple short conditions obscuring the predicted amelioration; for *which* it decreases the complex short conditions causing the appearance of amelioration that is not predicted by fRM. (We are running follow-ups with *what* and *what NP* to confirm.) Negative islands are explained in fRM through the neg operator sharing a feature with wh. But our results show no amelioration for negative islands, suggesting that they are not weak islands, and perhaps not part of fRM. Among factives, only emotives show a typical island effect. This supports Kartunnen's (2016) claim that only emotives are fully factive. Emotives show partial amelioration, lending support to an fRM analysis based on a null factive operator (Melvold 1991). Lastly, 2/4 noun complement islands show evidence of amelioration. These are "make the claim" and "hear the rumor" – two that are likely to be lexicalized as a unit (e.g. Ross 1967). This warrants additional study, as it suggests that lexicalized complex NPs become similar to weak islands (e.g., a verb and operator), while non-lexicalized NPs retain their full impenetrable islandhood. Our final **architectural question** is about the nature of the grammar itself. It is unlikely that the interaction of wh-type with structure can be explained by processing difficulty. This suggests either that binary grammars must find an additional mechanism to explain the effect, or that non-binary approaches to grammar should be considered, such as generative theories with levels (e.g. Chomsky 1964, Friedmann et al. 2009, Rizzi 2018), generative theories with constraint weights (e.g. Featherston 2005), and continuous theories (e.g. Keller 2000, Villata & Tabor 2022). We will make our data set public after publication so that researchers can explore the predictions of these theories.

Table 1 *Examples of islands tested (in the simple wh- condition, the complex wh-phrase was replaced by 'what').*

Wh-	Which dish did the waiter wonder whether/if/why the chef burned?	Fact	Which dish did the waiter forget/realize/acknowledge that the chef burned?
	Which dish did the waiter wonder who/which chef burned?	NC	Which dish did the waiter make/believe the claim that the chef burned?
Neg	Which dish didn't the waiter think that the chef burned?		Which dish did the waiter hear/believe the rumor that the chef burned?
	Which dish did the waiter not think that the chef burned?	RC	Which dish did the waiter blame the chef that/who burned?
Fact	Which dish was the chef sad that the customer hated?	Adj	Which dish did the waiter sigh because/after/if the chef burned?

Fig.1 *Interaction plots: p-values and Bayes factors for the island effect (interaction term) are colored to match the legend.*

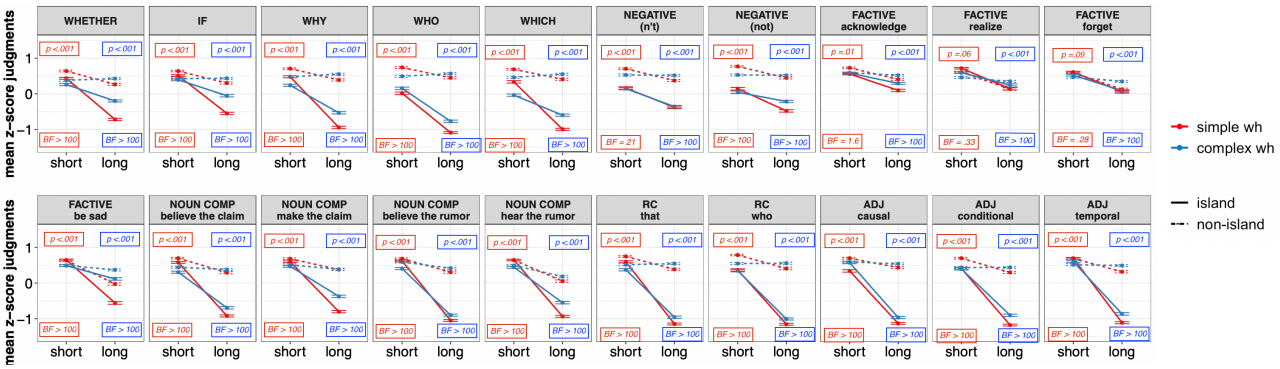

Fig.2 *95% credible intervals for the island effect (interaction term). Zero indicates no island effect.*