Does Relativized Minimality account for *wh*-island amelioration?

Abstract (100 words): This paper reports four acceptability judgment experiments that investigate the amelioration of *wh*-island violations, with a special focus on the role of D-linked *wh*-phrases. These observations have played a prominent role in theories of islands, such as a recent version of Relativized Minimality that attributes amelioration effects to the distinctness of the formal feature set on the moved constituent and intervener (Rizzi 2013). We show that the distribution of D-linking amelioration effect is not consistent with Featural Relativized Minimality's predictions, and argue that *wh*-island amelioration effects are better explained by semantic distinctness of the two *wh*-phrases.

Key words (4 to 6): Relativized Minimality, *wh*-island, D-linking, acceptability judgment, amelioration

1 Introduction

The main goal of this Remark is to explore empirical predictions of a recent version of Relativized Minimality (henceforth RM), which we will refer to as Featural Relativized Minimality (Rizzi 2013; see also Boeckx and Jeong 2003, Starke 2001). Since the original proposal in Rizzi (1990), the details of RM have undergone several revisions (e.g., Chomsky 1995, Rizzi 2004), but the general format of this locality constraint can be summarized as in (1), which is slightly modified from Rizzi (2013):

(1) In the configuration [… X … Z … Y …], X and Y cannot be connected by movement if Z c-commands Y and Z is the same structural type as X.

The locality condition as stated in (1) ensures that a movement relation cannot be established when there is a competing intervener (Z in (1)) that is structurally closer to the trace position (Y) than the ultimate landing site (X). The definition of structural relations between relevant constituents in (1) remained constant in all versions of RM, but in Featural RM, the definition of the *structural type* that constitutes a violation of RM is stated in terms of formal features of those constituents.

As will be reviewed in Section 2, Featural RM's revisions made it possible to explain a wider range of empirical data, such as amelioration of *wh*-island violations and children's non-adult-like comprehension errors (Belletti et al. 2012, Friedmann, Belletti and Rizzi 2009, Rizzi 2013). Importantly, Featural RM also makes novel empirical predictions for *wh*-island amelioration effects in sentences that involve two D-linked *wh*-phrases. We will point out that these predictions are not entirely consistent with the judgments reported in the literature, but the inconsistencies may be due to the fact that subtle changes in acceptability may be difficult to capture in informal judgments. Sections 3 to 6 present formal acceptability judgment experiments that were designed to evaluate these inconsistencies between Featural RM and previous data. Section 7 will discuss the implications of the findings, and suggest that a proper account of *wh*-island amelioration effects may need to incorporate semantic and psycholinguistic factors that influence the denotation of *wh*-phrases. Section 8 concludes the paper.

2 Featural RM: Revisions and Consequences

A critical empirical observation that led to Featural RM is the amelioration of *wh*-island violations with a D(iscourse)-linked *wh*-phrase (Pesetsky 1987). For example, it is ungrammatical to extract the bare *wh*-phrase *what* from the *wh*-island in (2a), but the extraction of the D-linked *wh*-phrase *which problem* in (2b) is considered marginal.[1] This suggests that the *wh*-island violation in (2b) is somewhat ameliorated, though its acceptability is still degraded compared to a grammatical *wh*-extraction in (2c).

(2) a. ***What** do you wonder **who** solved **__**?

b. ?**Which problem** do you wonder **who** solved **__**?

c. **Which problem** do you think that John solved **__**?

According to the original proposal in Rizzi (1990), both (2a) and (2b) violate RM due to the presence of the intervener *who* in [Spec, CP] of the embedded clause. Thus, the original RM offers no straightforward explanation for the acceptability difference between (2a) and (2b) (see Cinque 1990).

In order to account for this contrast, Rizzi and colleagues (Belletti et al. 2012, Friedmann, Belletti and Rizzi 2009, Rizzi 2013) adopted the following two changes.

First, the *structural type* that is relevant for RM is now reanalyzed as a set of morpho-syntactic features on the fronted constituent, intervener and trace position. Second, the strength of an RM violation is graded. In particular, a violation is less severe when the feature set associated with the extracted element is richer than (i.e., a superset of) the feature set associated with the intervener (Boeckx and Jeong 2003, Starke 2001). Based on the types of morpho-syntactic features assumed in works by Rizzi and colleagues, the sentences in (2) illustrate three critical feature set configurations: identity (2a), inclusion (2b), and disjunction (2c). In (2a), the extracted constituent and the intervener both contain only a [+Q] feature, and hence the feature sets are identical. This identity relation results in a severe degradation in acceptability. In (2b), the intervener only contains [+Q], whereas the feature set for the D-linked *wh*-phrase contains [+Q] as well as [+N], the latter of which represents the "referential status" of the *wh*-phrase, i.e., indicating that it is D-linked (see Cinque 1990). This configuration is called an inclusion configuration, as the extracted constituent is more richly specified, and its feature set is a superset of that of the intervener. This inclusion relation leads to a less severe degradation in acceptability, predicting that the *wh*-island effect is ameliorated relative to (2a) but that the sentence is not necessarily judged as perfect. Finally, in (2c) the embedded CP contains no [+Q] feature, and hence the feature specifications for the extracted constituent and the (potential) intervener are distinct. This is termed a disjunction configuration, which leads to no violation of Featural RM. These three feature set relations and their well-formedness statuses are summarized in Table 1.[2]

[Insert Table 1 here]

In summary, a key property of Featural RM is that it is concerned with whether the extracted constituent has a richer set of morpho-syntactic features than the intervener. This change allows Featural RM to achieve wider empirical coverage. Featural RM also leads to other novel predictions that we will not directly address here. For example, it has been argued that Featural RM also accounts for why children struggle in comprehension of object *wh*-questions with a D-linked *wh*-phrase (Belletti et al. 2012, Friedmann, Belletti, and Rizzi 2009).[3] The present paper focuses on adults' acceptability judgment data on *wh*-island violations in English and leaves aside discussions of children's data (but see Goodluck (2010) for critical comments on the relevance of Featural RM for the child comprehension data).

Despite the empirical gains described above, Featural RM also makes predictions about island amelioration that seem to be inconsistent with observations in the literature. For instance, examples such as those in (3) have been described as contrasting with each other; informal judgment data reported in the literature (Comorovski 1996, Pesetsky 1987, Shields 2008; see also Goodall 2015 for acceptability judgment data) suggests that the D-linked configuration in (3b) is more acceptable or grammatical than the non-D-linked intervener configuration in (3a).

(3) a. *Which athlete did she wonder who would recruit __? (inclusion)

b. ?Which athlete did she wonder which coach would recruit __? (D-linked identity)

In (3a) the extracted *wh*-phrase is D-linked and the intervener is a bare *wh*-phrase, whereas in (3b), both the extracted *wh*-phrase and the intervener *wh*-phrase are D-linked. Under Featural RM, the A'-chain in (3b) should be classed as an identity configuration, since both *wh*-phrases have features [+Q, +N]; we will refer to this configuration as *D-linked identity*. The A’-chain in (3a) should be an inclusion configuration, since the intervening *wh*-phrase has only the feature [+Q]. Given these featural assumptions, (3b) should be less acceptable than (3a). This is the opposite of what has been observed, and thus the predictions of Featural RM are at odds with the informal judgments in the literature. However, differences such as (3a) vs. (3b) have often been characterized as extremely subtle, and thus the reliability of data in (3) may be in question. These sentences are typically considered unacceptable or ungrammatical, and they differ only in the severity of degradation, which is not guaranteed to be readily distinguishable in informal judgments.

This paper aims to shed light on the status of this data, and therefore the treatment of Relativized Minimality, by using formal acceptability judgment experiments. The fact that the Featural RM theory makes fine-grained acceptability predictions should not be viewed as a problem; on the contrary it is a virtue, especially given that acceptability judgment experiments can provide a quantitative measure of acceptability variation that have proven useful for a variety of syntactic phenomena (e.g., Alexopoulou and Keller 2007, Featherston 2005, Hofmeister and Sag 2010, McDaniel and Cowart 1999, Sprouse and Hornstein 2013, Sprouse, Wagers, and Phillips 2012). The present paper reports four experiments that probe sentence acceptability using a 7-point scale, allowing a fine-grained measure of acceptability judgment.[4] Experiment 1 tests the validity of the claim that the D-linked identity configuration is more acceptable than inclusion with a single D-linked phrase. Experiment 2 further examines the acceptability of bare identity and inclusion. After not finding an effect of inclusion in experiment 2, experiment 3 manipulates the animacy of the extracted *wh*-phrase to more closely resemble examples in the literature. Finally, experiment 4 directly examines the effect of animacy on *wh*-island violations.

3 Experiment 1

The first experiment examines the prediction of Featural RM that *wh*-island violations with D-linked identity should be more severely degraded than *wh*-island violations with an inclusion configuration where only the extracted phrase is D-linked. We test this using a 2 x 2 design with extraction (non-extraction vs. extraction) and feature relation (non-identity vs. identity) as factors, as in (4). The extraction conditions, (4b) and (4d), contain extractions out of *wh*-islands. The non-extraction counterparts in (4a) and (4c) do not contain *wh*-island violations, and hence serve as baseline conditions.

(4) a. Non-extraction, non-identity

Which student wondered who would invite the visitor?

b. Extraction, non-identity (Inclusion)

Which visitor did you wonder who would invite ___?

c. Non-extraction, identity

Which student wondered which teacher would invite the visitor?

d. Extraction, identity (D-linked Identity)

Which visitor did you wonder which teacher would invite ___?

Featural RM predicts that the D-linked identity condition (4d) should be severely degraded because the set of features on both D-linked *wh*-phrases (*which NP*, [+Q, +N]) are identical. On the other hand, the inclusion configuration (4b) should be less degraded than (4d), because the features on the fronted phrase (*which NP*, [+Q, +N]) are a superset of the features on the intervener (*who*, [+Q]).

3.1 Method

Participants. Twenty-five self-reported native English speakers were recruited via Amazon Mechanical Turk.[5] They were paid $0.30 for their participation. The data from 3 additional participants was excluded from the analysis, as they only used the extreme ends of the scale in the pre-test phase (see below).

Materials. The stimuli for this experiment consisted of 16 sets of bi-clausal wh-questions (4). These 16 items were counter-balanced across 4 lists, so that each participant saw only one version of each target item. Forty-eight filler items of comparable length and varying acceptability were randomly interspersed with these target items for a total of 64 items. Based on our informal judgments and acceptability judgment data in the literature, we manipulated the acceptability of filler items to create three groups of fillers: those that are expected to receive high acceptability rating (good fillers), those that are expected to receive low rating (bad fillers), and sentences whose acceptability was expected to fall in between (middle fillers).[6] Fillers consisted of both declaratives and questions, which were included to ensure that the target items were not the only questions in the experiments. Having filler items with varying acceptability serves two purposes. First, this encourages the participants to use a large portion of the scale, which is critical for revealing subtle contrasts. Second, the data from fillers can serve as a baseline measure that can be used to estimate the magnitude of amelioration effects in target sentences.

Procedure. All of the acceptability judgment experiments in this paper have the same basic procedure. Participants were instructed to rate sentences on a scale from 1 (bad) to 7 (good). Before beginning the experiment, participants were provided with detailed instructions and examples to illustrate that the task is not about stylistic considerations, prescriptive norms, or the plausibility of the event described. This was followed by additional examples with varying degrees of acceptability to illustrate what type of sentences correspond to different parts of the scale. None of these example sentences used the same structure as the target sentences illustrated in (4).

Additionally, the first six experimental trials were identical for all participants and served as a pre-test phase. These six trials consisted of two highly acceptable sentences, two highly unacceptable sentences, and two marginal ones. These sentences were included to encourage participants to use the entire scale. The use of a large range of points on the scale was critical for the present study, because the target comparison involves two unacceptable sentence conditions. The acceptability contrast between such sentences may not be revealed if participants used, for example, only two extreme ends of the scale and treated the task as a binary judgment task. If participants restricted their judgments to the extreme ends of the scale (i.e., 1 and 7) on these initial items, the data from these participants were excluded from further analyses as it suggests that the participants were not reading the sentences carefully enough.

Data Analysis. All experiments in this paper use the same data analysis procedure. First, the raw judgment ratings were converted to z-scores (Schütze and Sprouse 2013). The z-score transformation converts a participant's scores to units that represent the number of standard deviations a particular rating is from that participant's mean rating. This procedure corrects for the potential that individual participants treat the scale differently, e.g., using only a subset of the available ratings, because it standardizes all participant's results to the same scale.[7]

We used linear mixed-effect models to analyze the data; these models allow the simultaneous inclusion of random participant and random item variables (Baayen, Davidson, and Bates 2008). Markov chain Monte Carlo (MCMC) sampling (n = 10,000) was used to estimate p-values for the fixed and random effects. When the results showed a significant interaction, planned pairwise comparisons were also performed to determine significance between individual conditions.

3.2 Results

Figure 2 presents the z-score transformed average ratings for each condition and for each filler type. Good filler sentences were rated as most acceptable (mean z-score = 0.80), while bad fillers were rated as least acceptable (mean z-score = -0.75). Middle fillers received ratings near participants' mean rating (i.e., near a z-score of 0, mean = -0.21). This pattern of acceptability for the fillers is common across all four experiments.

[Insert Figure 1 Here]

In the judgment data from the target items, we found that the extraction conditions, which contained an extraction out of *wh*-islands, were rated as less acceptable than the non-extraction conditions (extraction mean z-score = -0.71, non-extraction mean z-score = -0.05). Within the extraction conditions, the D-linked identity condition is rated as more acceptable than the inclusion condition (-0.58 vs. -0.84). In the non-extraction conditions, average z-scored ratings are around zero (means -0.04 and -0.07), suggesting that they were rated close to individual participants' mean ratings. This likely reflects the fact that sentences with two *wh*-phrases are generally uncommon and difficult to process.

Table 2 presents the estimated coefficients and the standard error (represented by the highest posterior density (HPD) range) for the Linear Mixed Effect model with extraction and feature relation as fixed effects and random intercepts for participants and items. Significant effects are marked by their beta estimates.

[Insert Table 2 About Here]

There is a main effect of extraction such that questions with extraction (i.e., *wh*-island violations) are significantly less acceptable than those without extraction. There is no main effect of feature relation, but there is a significant interaction of extraction and feature relation. The estimated coefficient of this interaction indicates that the feature combination had a significant effect in the extraction conditions, but not in the non-extraction conditions. This is supported by planned pairwise comparisons: the two non-extracted conditions are not significantly different from one another (t = -0.32, p > 0.1), while the D-linked identity condition is rated as significantly more acceptable than the inclusion condition (t = -3.10, p < 0.01).

3.3 Discussion

The results indicate that extraction out of a *wh*-island results in severe degradation of acceptability. More importantly, this degradation is modulated by the feature relation between the two *wh*-phrases: the D-linked identity condition shows greater acceptability than the D-linked inclusion condition. These results replicate informal acceptability judgment data in the literature (Comorovski 1996, Pesetsky 1987, Shields 2008), but are incompatible with the prediction of Featural RM that an identity configuration should be more degraded than an inclusion configuration. In fact, our results indicate that the D-linked identity configuration leads to a greater amelioration of the *wh*-island violation than an inclusion configuration.

We have so far focused only on the D-linked identity configuration. No items in this first experiment involve an identity configuration with bare *wh*-phrases, even though Rizzi's (2013) proposal critically relies on an acceptability difference between an identity configuration with bare *wh*-phrases and an inclusion configuration with a D-linked *wh*-phrase in the matrix CP. In order to confirm the presence of *wh*-island amelioration in the inclusion configuration, as predicted by Featural RM, Experiment 2 compares the inclusion condition against a bare identity condition, where both the extracted *wh*-phrase and the intervener are bare *wh*-phrases (see (2a)).

4 Experiment 2

4.1 Method

Participants. Thirty-two self-reported native English speakers participated via Amazon Mechanical Turk. They were paid $0.50 for participating.

Materials. The stimuli for this experiment consisted of 24 sets of biclausal sentences, which were constructed by using a 2x2x2 design with 3 factors: matrix *wh*-phrase (bare vs. D-linked), feature relation (non-identity vs. identity), and extraction (non-extraction vs. extraction). The experimental conditions shown in Table 3 include the same 4 conditions as Experiment 1 (those with a D-linked matrix *wh*-phrase) as well as 4 new conditions (those with a bare matrix *wh*-phrase) to test Featural RM's broader predictions for *wh*-island amelioration effects. First, the extraction conditions all involve *wh*-island violations, so their acceptability is predicted to be significantly lower than that of non-extraction conditions. Second, Featural RM predicts that the identity extraction conditions should be the most severely degraded compared to all the other conditions, including their non-extraction counterparts. It also predicts that the magnitude of degradation should not differ between the two identity extraction conditions. Third, the inclusion configuration should yield an amelioration of *wh*-island violations. Thus, the inclusion condition should yield a degradation compared to its non-extraction counterpart due to a *wh*-island violation, but the resulting acceptability should still be higher than the extracted identity conditions. Finally, the reverse inclusion configuration and its non-extraction counterpart are included in the design to test all combinations of the three factors we used in this experiment. Featural RM does not make explicit predictions for these conditions (see table 1); however, given that Rizzi and colleagues attribute the amelioration effects to the superset-subset relation of feature set between the extracted *wh*-phrase and intervener, we can infer the predictions of Featural RM to be that the acceptability of the reverse inclusion configuration should be similar to that of the two extracted identity conditions, and lower than the acceptability of the inclusion condition.

[Insert Table 3 about here]

These 24 items were counter-balanced across 8 lists, so that each participant saw only one version of a target item. Forty-eight filler items of comparable length and varying acceptability were randomly interspersed with these target items.

Procedure and data analysis. This experiment used the same procedure and data analysis steps as Experiment 1. In the statistical analysis, we added planned pairwise comparisons for the extracted bare identity, inclusion, and D-linked identity conditions, as the comparison of these three conditions is critical for establishing the amelioration of *wh*-island violations that are predicted by Featural RM.

4.2 Results

Similar to Experiment 1, all four extraction conditions were judged as less acceptable than their non-extraction counterparts (extraction mean z-score = -0.54, non-extraction mean z-score = 0.10). Among the non-extraction conditions, the non-identity bare matrix *wh*-phrase condition received the highest rating (mean = 0.25), but we will leave this aside as it bears no relevance to our goal of testing the predictions of Featural RM. The other non-extraction conditions were judged similarly with mean z-score ratings around zero (means -0.03, 0.10, and 0.09). Among the extraction conditions, the D-linked identity condition was rated as the most acceptable (mean = -0.38). The remaining three conditions received similar ratings (means -0.57, -0.58, and -0.62).

[Insert Figure 2 About Here]

The Linear Mixed Effect model analysis confirmed that the overall pattern is consistent with Experiment 1. Table 4 presents the estimated coefficients, the standard error, and the MCMC estimated *p*-value for the Linear Mixed Effect model with extraction, feature relation, and matrix *wh*-phrase as fixed effects and random intercepts for participants and items.

[Insert Table 4 About Here]

As in Experiment 1, there was a main effect of extraction and island factors, but there was no main effect of either feature relation or matrix *wh*-phrase. Importantly, there was a marginal interaction of extraction and feature relation as well as feature relation and matrix *wh*-phrase, which suggest that the feature relation factor modulates the effects of extraction or matrix *wh*-phrase type on the acceptability. Planned pairwise comparisons among extraction conditions revealed no significant difference between the bare identity condition and the inclusion condition ($t = 0.40$, $p > 0.1$), which suggests that the D-linking amelioration effect was not observed for the inclusion configuration. On the other hand, the D-linked identity condition is significantly more acceptable than the inclusion condition ($t = 2.17$, $p < 0.05$), and marginally more acceptable than the bare identity condition ($t = 1.67$, $p < 0.1$). This pattern suggests that the D-linked identity condition showed a reliable amelioration of *wh*-island violations.

4.3 Discussion

Replicating the findings from Experiment 1, *wh*-island violations with D-linked identity received a reliably higher acceptability rating than bare identity or inclusion configurations. Furthermore, there was no clear evidence for amelioration of the *wh*-

island violation in the inclusion condition. This selective *wh*-island amelioration effect is incompatible with the prediction of Featural RM that the acceptability of the inclusion configuration should be greater than bare or D-linked identity conditions.

However, the absence of an amelioration effect in the inclusion condition was somewhat surprising, given that amelioration effects in the inclusion configuration have been widely reported in the literature (Cinque 1990, Goodall 2015, Pesetsky 1987). Experiment 3 explores the role of distinctness of *wh*-phrases in amelioration of *wh*-island violations.

5 Experiment 3

Experiment 2 provided no evidence for a *wh*-island amelioration effect in the inclusion configuration. One plausible source of this unexpected finding is the number of animate nouns in the stimuli. Examples for *wh*-island amelioration in the literature typically included a single animate DP (5a), whereas the stimuli used in Experiment 2 (5b) included two animate DPs.

(5) a. Which book did you persuade which person to read __? (Pesetsky 1987)
b. Which athlete did you wonder who would recruit __? (from Table 3)

It is plausible that having two animate *wh*-phrases makes them less distinct, which may have increased confusability or processing demands in our stimuli. In fact, the psycholinguistics literature has reported a phenomenon called similarity interference where the presence of similar DPs in long-distance dependencies leads to comprehension difficulties (Gordon, Hendrick, and Johnson 2001, Gordon, Hendrick, and Levine 2002, Gordon et al. 2006, Lewis and Vasishth 2005). For example, Gordon, Hendrick, and

Johnson (2001) measured reading time for sentences with a variety of object relative clauses and cleft sentences while manipulating the types of the DPs between names, pronouns, and determiner and noun. The results showed that the reading difficulties in object relative clauses were more pronounced when the DP type (description vs. name) of the extracted DP and the subject DP matched (e.g., *It was the banker that the barber praised…*) compared to when there was a mismatch (e.g., *It was Ben that the barber praised…*). Importantly, the DP type manipulation did not affect reading difficulties in subject relative clauses or subject clefts, suggesting that the DP type match effect is restricted to cases where one of the DPs crosses the other DP via extraction, which is exactly the configuration in which Featural RM should apply. Gordon and colleagues interpreted this to follow from difficulties in either encoding the two similar DPs in memory or in retrieving the extracted DP at the verb position.

These findings from the psycholinguistics literature suggest that our inclusion configuration, which contained two animate DPs, may have unexpectedly been more difficult to process, decreasing its acceptability. In order to address this question, Experiment 3 replaces the animate extracted DP (e.g., *which athlete* in (5b)) with an inanimate DP to more closely resemble the examples from the literature.

5.1 Method

Participants. Thirty-one self-reported native English speakers participated via Amazon Mechanical Turk. They were paid $0.50 for completing the task.

Materials. The stimuli for this experiment consisted of 24 sets of biclausal sentences, following the same 2x2x2 design used in Experiment 2, with three factors:

extraction, feature relation, and matrix *wh*-phrase (see table 3). The non-extraction conditions were identical to those in Experiment 2, where the matrix *wh*-phrase was animate. In the new extraction conditions, on the other hand, the fronted *wh*-phrase was changed from an animate to an inanimate noun (e.g., *which event*). Because the animacy of the extracted NP has changed, *what* replaces *who* as the bare matrix *wh*-word in the bare identity and reverse inclusion conditions (i.e., *What did you wonder…?)*.

The 24 items were counter-balanced across 8 lists, such that each participant saw only one version of each. Forty-eight filler items of comparable length and varying acceptability were randomly interspersed with these target items for a total of 72 items.

Procedure and data analysis. The procedure and data analysis method were identical to those of Experiment 2.

5.2 Results

The acceptability judgment pattern in this experiment resembles that of Experiment 2 in that the D-linked identity condition received the highest rating among the extraction conditions (-0.06 vs. -0.62, -0.83, and -0.60). These data were submitted to Linear Mixed Effect model analyses, which used extraction, feature relation, and matrix *wh*-phrase as fixed effects and random intercepts for participants and items. The co-efficient estimates, standard error and MCMC estimated *p*-values are presented in Table 5.

[Insert Figure 3 and Table 5 about here]

The results revealed the same main effect of extraction as in the previous experiments due to the decreased acceptability of the island violating conditions (extraction mean = -0.52, non-extraction mean = 0.11). Unlike experiments 1 and 2, there is no interaction between extraction and feature relation. This is likely driven by the high acceptability of the non-extraction, non-identity condition with a bare matrix *wh*-phrase compared to the other non-extraction conditions (0.38 vs. -0.04, -0.01, and 0.10).

Next, following the data analysis procedure in Experiment 2, planned pairwise comparisons of the extraction conditions were performed to examine the precise distribution of the amelioration effect. Replicating the results of our previous experiments, the D-linked identity condition is significantly more acceptable than the inclusion condition (t = 5.62, p < 0.001) as well the bare identity condition (t = 8.06, p < 0.001). Importantly, unlike experiment 2, we found that the inclusion condition is significantly more acceptable than the bare identity condition (t = 2.38, p < 0.05).

5.3 Discussion

Once again, this experiment found that the D-linked identity condition was more acceptable than the other extraction conditions. Unlike Experiment 2, however, we found evidence for *wh*-island amelioration in the inclusion configuration, as the inclusion condition was judged as more acceptable than the bare identity condition. The fact that this effect was only found in this experiment suggests that the animacy manipulation plays a critical role in its emergence. However, in Experiment 3 extraction and animacy factors were confounded, as the extracted *wh*-phrases were always inanimate. This design does not allow a direct comparison of *wh*-island violations with animate extracted constituents to those with inanimate ones. Experiment 4 explores this issue by varying animacy within the extraction conditions.

6 Experiment 4

This experiment aims to directly compare *wh*-island violations with animate *wh*-phrases and those with inanimate *wh*-phrases. This allowed us to investigate the extent to which animacy contributed to *wh*-island amelioration effects.

6.1 Method

Participants. Thirty-four self-reported native English speakers participated via Amazon Mechanical Turk. They were paid $0.30 for completing the experiment.

Materials. The stimuli for this experiment consisted of 24 sets of biclausal sentences with a 2x2 design (6), using animacy of the matrix *wh*-phrase (animate vs. inanimate) and feature relation (inclusion vs. D-linked identity) as factors. These items were largely based on stimuli from the previous experiments.

(6) a. Animate, Inclusion

Which visitor did you wonder who would host ___?

b. Inanimate, Inclusion

Which event did you wonder who would host ___?

c. Animate, D-linked Identity

Which visitor did you wonder which family would host ___?

d. Inanimate, D-linked Identity

Which event did you wonder which family would host ___?

Given the results of experiment 3, if animacy modulates *wh*-island amelioration effects, then the inanimate conditions should be more acceptable than their animate counterparts regardless of the feature relation. Based on findings from the previous experiments, it was also predicted that the D-linked identity conditions should be reliably more acceptable than the inclusion conditions.

The 24 items were counter-balanced across 4 lists, such that each participant only rated a single item from each set. The addition of 48 length-matched filler sentences resulted in a total of 72 items.

Procedure and data analysis. The procedure and data analysis method were identical to those of previous experiments. Regardless of the presence of a significant interaction, planned pairwise comparisons of animacy within feature relation were conducted to directly test the effect of animacy on these feature relations.

6.2 Results

Figure 4 presents the mean z-score ratings in each condition. As in all three of the previous experiments, the D-linked identity conditions are rated as more acceptable that the inclusion ones (D-linked identity mean z-score = -0.22, inclusion mean z-score = -0.53). The inclusion conditions with animate and inanimate matrix *wh*-phrases do not differ in their acceptability ratings (-0.54 vs. -0.51). On the other hand, the D-linked identity condition with an inanimate *wh*-phrase was judged as more acceptable than its animate counterpart (-0.14 vs. -0.29).

[Insert Figure 4 About Here]

These data were analyzed using a Linear Mixed Effect model analysis with feature relation and animacy as fixed effects, and participants and items for random intercepts. The estimated coefficients, standard error, and MCMC estimated *p*-value for the Linear Mixed Effect model are given in table 6.

[Insert Table 6 About Here]

There is a main effect of feature relation such that the D-linked identity conditions were significantly more acceptable than inclusion conditions. This is consistent with all of our previous findings. Crucially, there was no main effect of animacy and no significant interaction between feature relation and animacy. Planned pairwise comparisons revealed that there was no reliable acceptability difference between the two inclusion conditions (t = -0.53, p > 0.1), but that the inanimate D-linked identity condition is marginally more acceptable than its animate counterpart (t = -1.96, p < 0.1). Thus, in this experiment the animacy manipulation only affected the already more acceptable D-linked identity conditions.

6.3 Discussion

Although the amelioration effect in the inclusion conditions was observed when the extracted *wh*-phrase was an inanimate noun (experiment 3), the results of experiment 4 suggest that the effect of animacy may not be as robust as one might expect. Animacy did not have a reliable effect across inclusion and D-linked identity configurations, as the D-linked identity conditions with inanimate *wh*-phrases showed a stronger amelioration effect. Taken together, the results of experiments 3 and 4 suggest that the animacy of the extracted *wh*-phrases can modulate *wh*-island amelioration effects, but that the effect can be subtle, as well as context-sensitive.

7 General Discussion

We have described four acceptability judgment studies investigating the distribution of *wh*-island amelioration effects, designed to test predictions of Featural Relativized

Minimality. In particular, we tested the acceptability of a *wh*-island violation involving two D-linked *wh*-phrases (i.e., D-linked identity) against violations with an intervening bare *wh*-phrase (i.e., inclusion) or with no D-linked *wh*-phrases (i.e., bare identity). There are two core motivations: first, the data as reported in the literature are subtle and potentially prone to gradience, and second, setting aside the subtlety, the D-linked identity configurations have been reported to have an amelioration effect on island violations (Comorovski 1996, Pesetsky 1987, Shields 2008), whereas Featural RM predicted that such violations should be ungrammatical. The combination of these two motivations led us to use formal acceptability judgment experiments, in order to have a finer-grained measure of acceptability variation.

There are two main findings from the experiments reported above. First, we found consistent evidence against the predictions of Featural RM about D-linked identity configurations: such configurations reliably led to a higher acceptability than inclusion configurations. Featural RM predicts the opposite. Moreover, a study that was conducted in parallel in French used a similar design to our experiment 2 and found the same pattern (Villata, Rizzi, and Franck, submitted). Thus, the increased acceptability of the D-linked identity configuration is robust across experiments and across English and French.

Second, we found that the D-linking amelioration effect for *wh*-island violations can be modulated by animacy, although the animacy effects were not always robust. Experiment 2 used only animate *wh*-phrases and found no evidence for *wh*-island amelioration in the inclusion configuration. Experiment 3 used inanimate nouns for extracted *wh*-phrases, and revealed evidence for amelioration in the inclusion configuration. This contrast between the experiments reveals that animacy must play a role. However, this effect did not hold robustly in experiment 4, where there was no reliable difference between the two inclusion conditions, which minimally differed in animacy of the extracted *wh*-phrase. While a complete understanding of the status of the inclusion configuration awaits further research, it is safe to conclude at this point that the *wh*-island amelioration effects for the inclusion configuration are not as robust as it has been reported in the literature (Cinque 1990, Pesetsky 1987, Rizzi 2013).

These findings are summarized in (7), which depicts the ranking of acceptability variation among the *wh*-island violations that were examined in this paper. We will now discuss the theoretical implications of these findings.

(7) Bare Identity ≤ (Reverse) inclusion with an animate *wh*-phrase extraction ≤ (Reverse) inclusion with an inanimate *wh*-phrase extraction < D-linked identity

7.1 Implications for Featural RM

Our data suggests that Featural RM does not fully account for the distribution of *wh*-island amelioration effects, especially the fact that the D-linked identity configuration led to a robust amelioration effect. We do not present this as an argument against Featural RM per se, but minimally something else must be said to account for the behavior of D-linked *wh*-items beyond the inclusion/identity featural distinction. One potential implication is that the set of morpho-syntactic features assumed in papers by Rizzi and colleagues may need to be enriched. We will explore below the addition of Topic or Animacy features below, but demonstrate that neither of these features provides a satisfactory explanation.

Luigi Rizzi (pers. comm.) suggests that the extracted D-linked *wh*-phrase has a [+Topic] feature that the intervening D-linked *wh*-phrase does not, as this feature is only licensed by the left periphery of the matrix clause (for a similar suggestion that the extracted *wh*-phrase may have a presupposition feature, see Boeckx and Jeong 2003, Grohmann 2000). If this is the case, then the sentences with two D-linked phrases are cases of inclusion rather than identity (8).

(8) **Which athlete** did you wonder **which coach** would recruit **__**?

[+Q, +N, +Topic]   [+Q, +N]   [+Q, +N, +Topic]

This amendment allows Featural RM to account for the increased acceptability of the D-linked identity configuration. However, this featural augmentation does not explain why this configuration should be reliably more acceptable than our inclusion condition with a bare *wh*-phrase in the intervener position. Given the feature sets assumed in (12), both of these configurations are inclusion configurations, which are not predicted to show a contrast in acceptability. If we were to grade acceptability based on the degree of featural overlap, the prediction would again go the wrong direction: the bare inclusion condition should have less featural overlap, and therefore be more acceptable, than the D-linked identity condition under the analysis in (8).

Another morpho-syntactic feature that may deserve to be added to the Featural RM framework is an animacy feature. It is typically assumed that animacy features do not actively participate in syntactic operations in English. However, animacy is known to play important roles in syntax of other languages (e.g., Slavic languages, see Rappaport 2003). Our observations of superior *wh*-island amelioration effects for inanimate *wh*-

phrases may be the first evidence that animacy plays an important role in English syntax as well. However, the addition of an animacy feature with the same status as e.g., [+Q] above is not fully motivated by our data either. First, it offers no explanation for the observed acceptability contrast between the D-linked identity and inclusion configuration in experiments 1 and 2. Second, using animacy features in experiment 3 would change the D-linked identity feature relation to that of a reverse inclusion (9).[8] Under this configuration, Featural RM predicts the sentence to be equally as degraded as identity configurations, which is the opposite of what was found in experiment 3. Rather, if experiment 3 is taken at face value, (9) should be ameliorated simply because the two D-linked *wh*-phrases have a different value for animacy.

(9) **Which award** did you wonder **which actress** should receive **__**?

[+Q, +N]   [+Q, +N,+animate]   [+Q, +N]

Finally, incorporating an animacy feature would predict that animacy based amelioration effects hold robustly across all *wh*-island violations, but this prediction is inconsistent with the observation in experiment 4 that the animacy manipulation selectively modulated the acceptability of the D-linked identity conditions but not the inclusion configuration. While an animacy distinction is clearly relevant, it cannot easily be captured in featural terms.

In summary, it is not obvious what featural adjustments could account for the amelioration patterns we have shown in this paper in a way that is entirely internal to the principles of Featural RM, though we are not yet ready to rule this possibility out. If this effect cannot be accounted for with featural manipulations, then (minimally) something external to the featural system must lead to the amelioration pattern.

7.2 The Role of Semantic Distinctness in Acceptability Variation

An explanation for the distribution of *wh*-island amelioration effects in our experiments must take into account the superior amelioration effects in D-linked identity configurations, as well as the fact that extraction of an inanimate *wh*-phrase sometimes led to a further increase in acceptability. Before we present such explanations, we first argue for a new descriptive generalization: the degree of semantic distinctness of the extracted *wh*-phrase and the intervener (rather than the distinctness of formal features) predicts the distribution of *wh*-island amelioration effects.

First, we will assume a broadly Hamblin semantics of *wh*-questions and assume that (i) questions denote a set of possible answers (Hamblin 1973; see also Karttunen 1977 and many others), and (ii) *wh*-phrases denote a set of potential referents (Hamblin 1973, Kratzer and Shimoyama 2002). Intuitively, the set of referents for the *wh*-item in a single-*wh* question corresponds to possible fragment DP answers to that question. Under this family of assumptions, bare *wh*-phrases like *who* denote the set of all human individuals, whereas a D-linked *wh*-phrase like *which award* would denote a presupposed set of entities satisfying the NP restrictor, in this case awards, and require the answer to the *wh*-question to be constructed from some referent in this set only. With these assumptions, let us examine the distinctness of sets of individuals or objects denoted by *wh*-phrases in table 7, which illustrates the main feature configurations that were investigated in our acceptability judgment experiments.

[Insert Table 7 About Here]

In the bare identity condition with *who* as an extracted *wh*-phrase, both the extracted *wh*-phrase and the intervener denote the set of all humans, and therefore their domains are identical and non-distinct. If the extracted *wh*-phrase is *what,* we assume that *what* denotes a set of all "things" in the world, which include human individuals.[9] Here, the set denoted by *what* is a superset of the set denoted by *who*, and these sets are thus overlapping. As for the inclusion configuration with animate *wh*-phrases, *which visitor* denotes a presupposed set of visitors, while *who* denotes a set of all human individuals. Thus, the sets of individuals denoted by these two *wh*-phrases are also overlapping. On the other hand, for the inclusion configuration with inanimate and animate *wh*-phrases, the set denoted by *which event* and the set denoted by *who* are distinct. This explains the amelioration effect that was observed in the comparison of experiments 2 and 3. Finally, in the D-linked identity conditions, the sets of individuals or objects denoted by the two *wh*-phrases (*which visitor* and *which family*, or *which event* and *which family*) are clearly distinct. Thus, these observations lead to the generalization that the *wh*-island violations that were amenable to amelioration effects were those in which the sets denoted by the extracted *wh*-phrase and the intervener are distinct. We take this as a necessary condition for *wh*-island amelioration.

The semantic distinctness of the *wh*-phrases seems to provide the beginnings of an explanation of many of the patterns in our data, one that is parallel to the featural account in Featural RM, but residing in a different part of the language faculty. However, the question remains as to why semantic features should play a role in modulating the acceptability of a syntactic phenomenon. One possible answer to this question is the psycholinguistic constraints on formation of long-distance dependencies. As noted in section 5, it has been widely observed that the processing of long-distance dependencies can be impeded when the dependencies contain two similar DPs. This similarity interference effect is considered to follow from limitations of the memory system in either encoding two similar DPs as distinct items, or in retrieving the target DPs with accurate syntactic and semantic features. It is plausible that the semantic distinctness of *wh*-phrases modulates the ease of encoding or retrieval processes, and when these processes are readily performed, participants may perceive the *wh*-island violations to be less severely degraded. In this sense, the semantic distinctness of *wh*-phrases may serve as a formal characterization of DPs that are particularly confusable for memory operations.[10]

This psycholinguistic explanation for the role of semantic distinctness would have implications for theories of islands and syntactic amelioration effects in general. One possible implication is that island constraints themselves might be reducible to cognitive constraints on memory operations, such that "island violations" merely reflect difficulties in establishing *wh*-dependencies during real-time parsing (Hofmeister and Sag 2010, Kluender and Kutas 1993; for related explanations for Superiority effects, see Hofmeister et al. 2013). While our observations of semantic distinctness and its relevance to memory interference effects are compatible with this approach, we note that this explanation has been challenged by the observation that the severity of island violations is not predicted by individual differences in working memory capacity (Sprouse, Wagers and Phillips, 2012; see also Phillips 2006). Therefore, we do not take our study to provide evidence for a purely domain-general account of island effects. This, however, means that we must suggest a new explanation for the amelioration effects we have observed.

We propose an alternative approach that situates similarity interference effects in *repair processes* that the parser initiates upon detecting a syntactic violation; we term this approach the Amelioration-as-Repair hypothesis. This explanation of amelioration effects relies on the following three assumptions. First, we assume that acceptability judgment intuitions minimally reflect the well-formedness of syntactic derivations and LF representations that the parser assigns to a given sentence. When this process fails due to linguistic or other cognitive constraints, we perceive degradation in sentence acceptability (Schütze 1996), and the severity of degradation reflects the number of constraint violations at all levels of representations (Haegeman, Jiménez-Fernández, and Radford 2014, Keller 2000, Legendre, Miyata, and Smolensky 1991, Smolensky and Legendre 2006). Second, we also assume that syntactic constraints on *wh*-islands do play an important role in accounting for the general acceptability degradation due to extraction out of *wh*-islands, and this constraint could be the original Relativized Minimality in Rizzi (1990). Finally, we also assume that in the face of sentences that violate syntactic constraints, the parser attempts to repair the structure in order to assign an interpretation to the structurally unintegrated *wh*-phrase. Such interpretive repair processes are well documented in the psycholinguistics literature on severe garden-path sentences (e.g., Christianson et al. 2001, Ferreira and Patson 2007). While this style of repair may not cancel the initial violation of syntactic constraints, it would at least ensure that the sentence receives a legitimate LF representation that can be passed onto (in Chomskyan terms) the Conceptual-Intentional system. Under these assumptions, acceptability judgment data should reflect the degree to which this repair process is able to identify a gap position inside an island and complete the *wh*-dependency for the LF representation (and potentially any cognitive burden involved in the repair).

We suggest that it is in this repair process that the similarity interference effects arise. It is well known that the parser typically respects island constraints during real-time sentence processing (e.g., Stowe 1986, Traxler and Pickering 1996); thus, initially the parser should generate an ungrammatical structure with no gap for the *wh*-phrase. This syntactic violation initiates the repair process, and searches for a gap inside an island. This interpretive process is plausibly sensitive to the semantic distinctness of *wh*-phrases, because this repair process by definition requires retrieval of constituents that were processed earlier. If the repair process fails due to similarity interference effects (e.g., in bare identity condition), the LF representation would veridically reflect the syntactic violation of *wh*-island constraint (i.e., no gap for the *wh*-phrase), and the sum of these two violations results in more severe degradation. On the other hand, if the parser identifies a gap inside an island due to the lack of similarity interference effects (e.g., in D-linked identity conditions with semantically distinct *wh*-phrases), the resulting LF representation no longer contains any violation, even though it is derived from a structure that does, and therefore the only source of acceptability degradation is the initial violation of the *wh*-island constraint (see Huang 1982 for arguments that the LF representation of islands with argument gaps does not incur any violation).

If the Amelioration-as-Repair hypothesis is on the track, it provides a new direction towards a mechanistic understanding of acceptability judgment data in general. To this day, even though acceptability judgment data has served as the primary source of data for linguists, there is very little theory of how such intuitions arise (cf. Schütze 1996), or how the process of judging sentence acceptability reflects psycholinguistic constraints.

As a final note, under this explanation there are two observations that need to be explained in future research. First, this account of amelioration does not explain why the animacy-based modulation of *wh*-island amelioration effects was not reliably observed across experiments. Second, it also does not offer a straightforward explanation for the acceptability difference observed between the two D-linked identity conditions in experiment 4. Following the psycholinguistic explanation above, we tentatively suggest that the real-time encoding and comparison of semantic distinctness information could be subject to a variety of conceptual or cognitive factors that will then impact the behavior of amelioration. For example, accessing the set of all individuals denoted by *who* may be inherently complex when it is presented out of context, as in the current experiments. This difficulty may sometimes mask the potential advantage of semantic distinctness in the inclusion configuration with an inanimate *wh*-phrase, suggesting also that it may not be generally safe to test amelioration effects out of context.

8 Conclusion

In order to systematically test predictions of Featural Relativized Minimality, we conducted four acceptability judgment experiments to investigate the distribution of *wh*-island amelioration effects, with a special focus on how it is modulated by D-linked *wh*-

phrases and animacy manipulation. We found that Featural RM in its current form failed to account for the distribution of *wh*-island amelioration effects. We suggested that a full explanation of our results requires the consideration of semantic representations, which may, in turn, be related to constraints on the sentence processing mechanisms that give rise to similarity interference effects. In particular, we introduced the Amelioration-as-Repair hypothesis: amelioration effects at least partially signal the degree to which the parser was able to repair a syntactic violation in the process of constructing an LF interpretation. This hypothesis calls for future work that re-examines amelioration effects in other syntactic environments in light of constraints on sentence processing mechanisms.

References

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

Baayen, R.H., D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59: 390–412.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72: 32–68.

Belletti, Adriana, Naama Friedmann, Dominique Brunato, and Luigi Rizzi. 2012. Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua* 122: 1053–1069.

Boeckx, Cedric and Youngmi Jeong. 2003. The fine structure of syntactic intervention. In *Proceedings of the Thirty-First Western Conference on Linguistics*, vol. 14, ed. by Brian Agbayani, Paivi Koshkinen, and Vida Samiian, 33–41. California State University, Fresno, Department of Linguistics Publications.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Christianson, Kiel, Andrew Hollingworth, John F. Halliwell, and Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology* 42: 368–407.

Cinque, Guglielmo. 1990. *Types of A'-dependencies*. (Linguistic Inquiry Monographs 17). Cambridge, MA: MIT Press.

Comorovski, Ileana. 1996. *Interrogative Phrases and the Syntax-Semantics Interface*. New York: Springer.

Dayal, Veneeta. 2002. Single-pair versus multiple-pair answers: Wh-in-situ and scope. *Linguistic Inquiry* 33: 512–520.

Featherston, Sam. 2005. That-trace in German. *Lingua* 115: 1277–1302.

Ferreira, Fernanda and Nikole D. Patson. 2007. The "good enough" approach to language comprehension. *Language and Linguistics Compass* 1: 71–83.

Friedmann, Naama, Adriana Belletti, and Luigi Rizzi. 2009. Relativized relatives: Types of intervention in the acquisition of A' dependencies. *Lingua* 119: 67–88.

Gibson, Edward, Steve Piantadosi, and Kristina Fedorenko. 2011. Using Mechanical Turk to Obtain and Analyze English Acceptability Judgments. *Language and Linguistics Compass* 5: 509–524.

Goodall, Grant. 2015. The D-linking effect on extraction from islands and non-islands. *Frontiers in Psychology: Language Sciences* 5: 1493.

Goodluck, Helen. 2010. Object extraction is not subject to Child Relativized Minimality. *Lingua* 120: 1516–1521.

Gordon, Peter C., Randall Hendrick, and Marcus Johnson. 2001. Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27: 1411–1423.

Gordon, Peter C., Randall Hendrick, Marcus Johnson, and Yoonhyoung Lee. 2006. Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32: 1304–1321.

Gordon, Peter C., Randall Hendrick, and William H. Levine. 2002. Memory-load interference in syntactic processing. *Psychological Science* 13: 425–430.

Grohmann, Kleanthes K. 2000. Prolific peripheries: A radical view from the left. Doctoral Dissertation, University of Maryland, College Park.

Haegeman, Liliane, Angel Jiménez-Fernández and Andrew Radford. 2014. Deconstructing the subject condition in terms of cumulative constraint violation. *Linguistic Review* 31: 73–150.

Hamblin, C. L. 1973. Questions in Montague English. *Foundations of Language* 10: 41–53.

Hofmeister, Philip and Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language* 86: 366–415.

Hofmeister, Philip, T. Florian Jaeger, Inbal Arnon, Ivan A. Sag and Neal Snider. 2013. The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes* 28: 48–87.

Huang, Cheng-Teh James. 1982. Logical relations in Chinese and the theory of grammar. Doctoral dissertation, MIT, Cambridge, MA.

Karttunen, Lauri. 1977. Syntax and semantics of questions. *Linguistics and Philosophy* 1: 1–44.

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

Kluender, Robert and Marta Kutas. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8: 573–633.

Kratzer, Angelika and Junko Shimoyama. 2002. Indeterminate pronouns: The view from Japanese. In *The Proceedings of the Third Tokyo Conference on Psycholinguistics*, ed. by Yukio Otsu, 1–25. Tokyo: Hituzi Syobo.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1991. Unifying syntactic and semantic approaches to unaccusativity: A connectionist approach. In *Proceedings* *of the Seventeenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on The Grammar of Even Structure*, vol. 17, ed. by Christopher Johnson, Laurel A. Sutton, and Ruth Shields, 156–167. University of California, Berkeley: Berkley Linguistics Society.

Lewis, Richard L. and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29: 375–419.

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

Pesetsky, David. 1987. Wh-in-situ: Movement and unselective binding. In *The representation of (in)definiteness*, ed. by Eric J. Reuland and Alice G. B. ter Meulen, 98–129. Cambridge, MA: MIT Press.

Phillips, Colin. 2006. The real-time status of island phenomena. *Language* 82: 795–823.

Rappaport, Gilbert C. 2003. The grammatical role of animacy in a formal model of Slavic morphology. In *American Contributions to the Thirteenth International Congress of Slavists (Ljubljana, 2003)*, vol. 1: Linguistics, ed. by Robert A. Maguire and Alan Timberlake, 149–166. Bloomington, IN: Slavica.

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

Rizzi, Luigi. 2004. Locality and left periphery. In *Structures and beyond: The cartography of syntactic structures*, vol. 3, ed. by Adriana Belletti, 223–251. Oxford: Oxford University Press.

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

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

Schütze, Carson T. and Jon Sprouse. 2013. Judgment data. In *Research methods in linguistics*, ed. by Robert J. Podesva and Devyani Sharma, 27–50. New York: Cambridge University Press.

Shields, Rebecca. 2008. What's so special about D-linking? Poster session presented at the NELS 39, Cornell University, Ithaca, New York.

Smolensky, Paul and Géraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar. Vol. 1: Cognitive architecture; vol. 2: Linguistic and philosophical implications.* Cambridge, MA: MIT Press.

Sprouse, Jon. 2011a. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87: 274–288.

Sprouse, Jon. 2011b. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavioral Research Methods* 43: 155–167.

Sprouse, Jon and Norbert Hornstein, eds. 2013. *Experimental syntax and island effects*. Cambridge: Cambridge University Press.

Sprouse, Jon, Matt Wagers, and Colin Phillips. 2012. A test of the relation between working-memory capacity and syntactic island effects. *Language* 88: 82–123.

Starke, Michal. 2001. Move reduces to merge: A theory of locality. Doctoral dissertation, University of Geneva, Switzerland.

Stowe, Laurie E. 1986. Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes* 1(3): 227–245.

Szabolsci, Anna and Frans Zwarts. 1993. Weak islands and an algebraic semantics for scope taking. *Natural Language Semantics* 1: 235–284.

Traxler, Matthew J. and Martin J. Pickering. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language* 35: 454–475.

Villata, Sandra, Luigi Rizzi, and Julie Franck. Submitted. Intervention effects in weak islands and Relativized Minimality: New experimental evidence from graded judgments.

Weskott, Thomas and Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87: 249–273.

Table 1

Taxonomy of feature set and well-formedness in Featural RM

X *Fronted phrase*	**Z** *Intervener*	**Y** *Trace position*	**Well-formedness**	**Type**
+A	+A	<+A>	Ungrammatical (*)	Identity
+A, +B	+A	<+A, +B>	Marginal (?)	Inclusion
+A	+B	<+A>	Grammatical (✓)	Disjunction

Table 2

Fixed effects summary for Experiment 1 with by-participant and by-item random intercepts.

	Estimate	*HPD95 lower*	*HPD95 upper*
Intercept	-0.04	-0.22	0.15
Extraction	-0.80 ***	-0.96	-0.62
Feature relation	-0.03	-0.20	0.14
Extraction x Feature relation	0.28 *	0.03	0.52

† *p* < 0.1, * *p* < 0.05, ** *p* < 0.01, *** *p* < 0.001

Table 3

Sample item set from Experiment 2

Bare matrix *wh*-phrase	Non-identity	Non-extraction	Who wondered which teacher would invite the visitor?
		Extraction	Who did you wonder which teacher would invite ___? *(Reverse Inclusion)*
	Identity	Non-extraction	Who wondered who would invite the visitor?
		Extraction	Who did you wonder who would invite ___? *(Bare Identity)*
D-linked matrix *wh*-phrase	Non-identity	Non-extraction	Which student wondered who would invite the visitor?
		Extraction	Which visitor did you wonder who would invite ___? *(Inclusion)*
	Identity	Non-extraction	Which student wondered which teacher would invite the teacher?
		Extraction	Which visitor did you wonder which teacher would invite ___? *(D-linked Identity)*

Table 4

Fixed effects summary for experiment 2 with by-participant and by-item random intercepts for extraction type, feature relation, and matrix *wh*-phrase type.

	Estimate	*HPD95 lower*	*HPD95 upper*
Intercept	0.10	-0.06	0.26
Extraction	-0.72 ***	-0.95	-0.53
Feature relation	-0.02	-0.23	0.18
Matrix *wh*-phrase	0.15	-0.06	0.36
Extraction x Feature relation	0.25 †	-0.03	0.57
Extraction x Matrix *wh*-phrase	-0.10	-0.39	0.20
Feature relation x Matrix *wh*-phrase	-0.26 †	-0.57	0.03
Extraction x Feature relation x Matrix *wh*-phrase	0.02	-0.38	0.46

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5

Fixed effects summary for experiment 3 with by-participant and by-item random intercepts for extraction type, feature relation, and matrix *wh*-phrase type.

	Estimate	*HPD95 lower*	*HPD95 upper*
Intercept	0.38 ***	0.22	0.52
Extraction	-0.99 ***	-1.18	-0.80
Feature relation	-0.41 ***	-0.59	-0.22
Matrix *wh*-phrase	-0.38 ***	-0.57	-0.20
Extraction x Feature relation	0.19	-0.07	0.45
Extraction x Matrix *wh*-phrase	0.40 **	0.13	0.66
Feature relation x Matrix *wh*-phrase	0.51 ***	0.25	0.78
Extraction x Feature relation x Matrix *wh*-phrase	0.24	-0.12	0.63

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6

Fixed effects summary for experiment 4 with by-participant and by-item random intercepts for feature relation and animacy of the matrix *wh*-phrase.

	Estimate	*HPD95 lower*	*HPD95 upper*
Intercept	-0.51 ***	-0.64	-0.37
Feature relation	0.36 ***	0.23	0.49
Animacy	-0.04	-0.17	0.09
Feature relation x Animacy	-0.11	-0.29	0.08

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7

Distribution of amelioration effects and semantic distinctness

Conditions	Sentence	Amelioration?	Semantic distinctness
Bare identity	Who/what did you wonder who would host___?	no	non-distinct
Inclusion	Which visitor did you wonder who would host ___?	no	non-distinct
Inclusion	Which event did you wonder who would host ___?	yes	distinct
D-linked identity	Which visitor did you wonder which family would host ___?	yes	distinct
D-linked identity	Which event did you wonder which family would host ___?	yes	distinct

Figure 1

Mean z-score acceptability rating of target questions by *wh*-phrase combination and extraction type, and mean z-score acceptability rating of filler sentences by filler type. Error bars indicate ± 1 standard error.

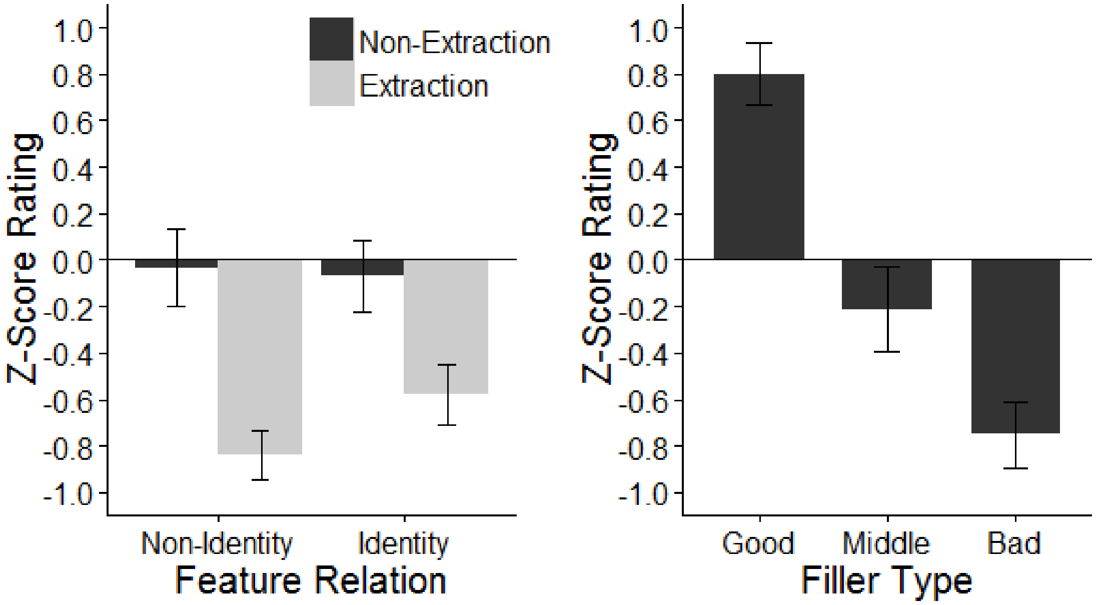


Figure 2

Mean z-score acceptability rating in experiment 2. Error bars indicate ± 1 standard error.

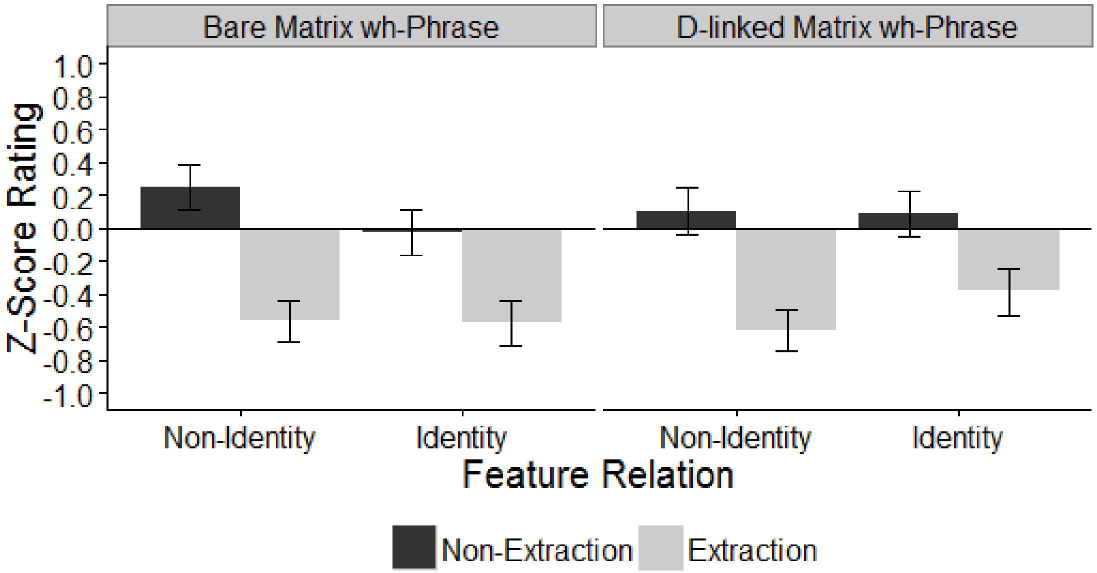


Figure 3

Mean z-score acceptability rating in experiment 3. Error bars indicate ± 1 standard error.

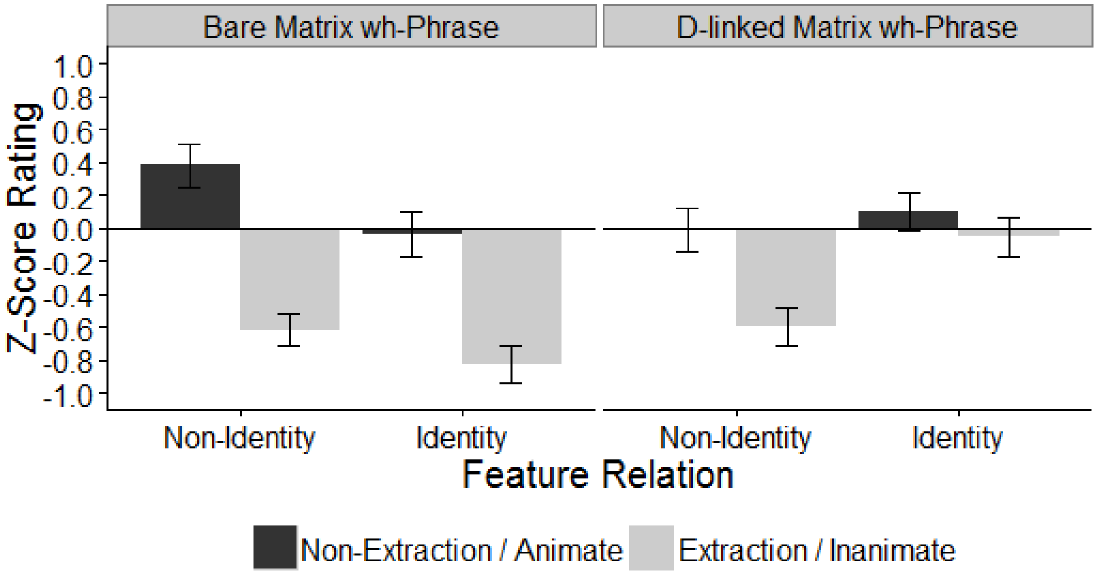
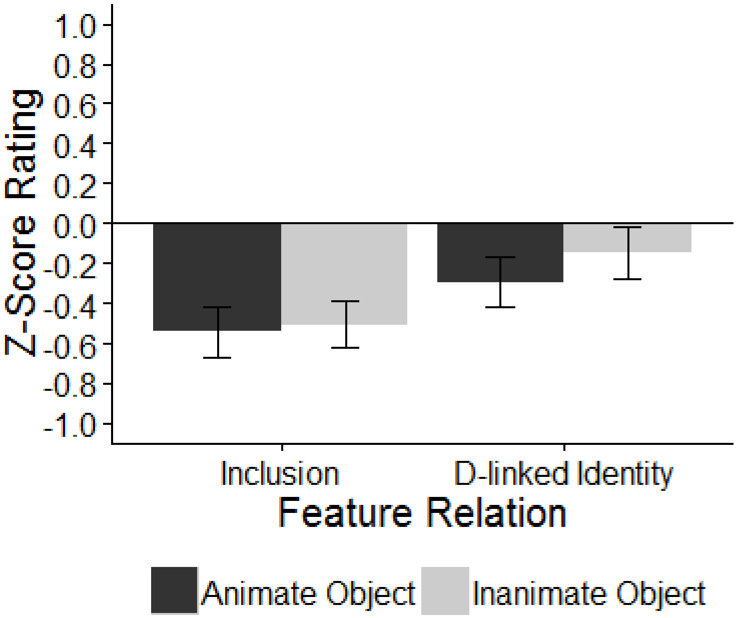


Figure 4

Mean z-score acceptability rating by *wh*-phrase combination and animacy of the object.

Error bars indicate ± 1 standard error.



Footnotes

[1] Pre-theoretically, Pesetsky characterized the difference between *what* in (2a) and *which problem* in (2b) in terms of the fact that *which* requires a salient antecedent set of problems in the discourse, whereas *what* requires no such thing; hence the term "Discourse-Linked." It remains unclear exactly what role this generalization plays in the grammar of amelioration (though we offer some discussion in section 7), and for the bulk of this paper we will use "D-linked" as a cover-term for *which*-phrases.

[2] Belletti et al. (2012) discuss another feature set relation called intersection, where one or more (but not all) features are shared by the intervener and the extracted *wh*-phrase, as seen in the Hebrew example in (i). They argue that these sentences are degraded to the same extent as inclusion configurations. We will leave aside the intersection configuration in this paper as the argument so far is solely based on children's comprehension performance, and it is impossible to construct an intersection configuration in the English *wh*-islands that are investigated here.

(i)	Tare	li	et	**ha-yalda**	**she-ha-rofa**	mecayer __.
	Show	to-me	ACC	the-girl(fem)	that-the-doctor(masc)	draws-masc
				[+R, +NP, +fem]	[+NP, +masc]	[+R, +NP, +fem]

[3] Friedmann, Belletti, and Rizzi (2009) used a picture selection task to investigate 3- and 4-year-old Hebrew-speaking children's comprehension of object *wh*-questions (ii), which either used a bare *wh*-phrase (iia) or a D-linked *wh*-phrase (iib).

(ii) a. Et mi ha-xatul noshex ?
ACC who the-cat bites
"Whom does the cat bite?"

b. Et eize kelev ha-xatul noshex ?
ACC which dog the-cat bites
"Which dog does the cat bite?"

Children's comprehension accuracy was above chance level in bare *wh*-questions like (iia) and at chance level in D-linked *wh*-questions like (iib). Friedmann and colleagues interpreted this to follow from Featural RM: In (iib), the D-linked *wh*-phrase *eize kelev* 'which dog' contains a [+Q, +N] feature set, and the subject NP *ha-xatul* 'the cat' a [+N] feature. This can be seen as an inclusion configuration, if we assume that features of constituents that are not landing sites for movement also play a role in the calculation of RM violations. On the other hand, the bare *wh*-question in (iia) has a disjunction configuration because the *wh*-phrase only contains a [+Q] feature. We note that this assumption that constituents outside of potential landing sites for movement (e.g., the subject NP in object *wh*-questions) can be treated as interveners is a major divergence from the original conceptualization of RM.

[4] Some have claimed that magnitude estimation measures are more appropriate for revealing fine-grained acceptability differences than Likert scales (Bard, Robertson, and Sorace 1996), but see Sprouse (2011a) as well as Weskott and Fanselow (2011) for empirical arguments against this claim.

[5] Recent studies have shown that acceptability judgment data collected via Mechanical Turk are indistinguishable from those collected in the laboratory (Gibson, Piantadosi, and Fedorenko 2011, Sprouse 2011b).

[6] Good fillers consisted of grammatical declaratives and questions of a comparable length to the target items, e.g., (iiia). Bad fillers were ungrammatical and consisted of structures such as filled gaps (iib) and coordinate structure constraint violations. Middle fillers consisted of minor grammatical violations such as extraction from weak islands (e.g., subject islands, complex NP islands) and use of resumptive pronouns (iiic).

(iii) a. The professor believes that the slides had been corrected. (Good)
b. It was a cigarette that I wished I had a cup of coffee. (Bad)
c. It was the war that the speech about interrupted the TV show. (Middle)

[7] We also ran the reported analyses with the raw ratings and checked that the results were unchanged in all experiments, although we will only report data and analyses based on z-scores.

[8] For the sake of the argument, we treat animacy as a privative feature in (9). There is no theoretical reason it could not be a binary feature, but the other morpho-syntactic features presented within this paper and within Featural RM (namely [+Q], [+N], and [+Topic]) are also treated as privative. If the animacy feature was binary, the case in (9) would be an example of intersection (see footnote 2) rather than inclusion.

[9] While bare *what* typically cannot refer to humans, *what* with an NP can, and so we take it to be generally unmarked.

[10] It is also possible that semantic constraints on multiple *wh*-questions directly influence the severity of island violations. For example, some analyses of the semantics and pragmatics of multiple *wh*-questions lead us to expect interactions of the domains of *wh*-phrases of the kind described in the above section (Comorovski 1996, Dayal 2002, Szabolsci and Zwarts 1993); in general, these theories do involve calculations that are sensitive to the domains of quantification and may have different semantic/pragmatic mechanisms for configurations where there is an island violation or similar effect. For example, Dayal (2002) proposes that there is an interaction between island amelioration and the interpretation of the question: focusing on superiority examples, she argues that apparent LF island amelioration involves single-pair readings, and not pair-list readings, and that these two types of readings involve different compositional mechanisms. Single-pair readings are interpreted via a choice function bound across an island boundary, and multiple-pair readings involve LF-movement and quantification over pairs constructed from the *wh*-items. The latter is blocked by islands. The interpretation of a choice function is extremely dependent on the domain that it composes with, and therefore we would expect the interpretation of ameliorated *wh*-items in general to involve calculations on this domain, and the psycholinguistic mechanisms suggested in the previous section would come into play. For multiple *wh*-items where there is no RM violation, of course, we expect either of Dayal's mechanisms to come into play, and the quantificational mechanism for pair-list readings is entirely different.