The effect of filler complexity and context on the acceptability of *wh*-island violations in Dutch

Maud Beljon, Dennis Joosen, Olaf Koeneman, Bram Ploum, Noëlle Sommer, Peter de Swart, Veerle Wilms

Radboud University Nijmegen

Abstract
Acceptability judgements of syntactic island violations are often claimed to improve by either increasing the complexity of the *wh*-filler phrase or integrating the violating sentence into a discourse. In two acceptability judgement tasks, we looked at *wh*-island violations in Dutch by varying the complexity of the filler phrase and by presenting the sentences either in isolation or with a preceding discourse. We found that neither variable had a significant effect in isolation, but that only in their combination a significant effect was observed. The same effect showed up in non-island conditions, however. This is in contrast to findings in the literature on English and French and suggests that the complexity effect in Dutch is not syntactic. We therefore conclude that *wh*-islands are strong islands in Dutch (Broekhuis & Corver 2015) and show that the contrast with English and French can be made to follow from featural Relativized Minimality (Rizzi 2017), taking into account the verb second property of Dutch.

Keywords
Island constraints, filler complexity, D-linking, context, Relativized Minimality

1. Amelioration effects in island violations

Wh-movement out of an embedded *wh*-question leads to ungrammaticality but, as is well known, the status of such sentences improves if the *wh*-filler is not a bare *wh*-word but a lexical phrase. See the contrast below, where the example in (1)b can still be degraded but is generally felt to be better than (1)a:

(1) a. *What do you wonder [who solved __]?
b. ?Which problem do you wonder [who solved __]?

This contrast, which has been confirmed in larger-scale experimental settings (cf. Hofmeister & Sag 2010; Sprouse et al 2016), has received different explanations. One prominent proposal is that lexical *wh*-phrases are D(iscourse)-linked (Pesetsky 1987) and are therefore easier to interpret than bare *wh*-words. This consequently leads to an amelioration effect in (1)b (cf. Szabolcsi and Zwarts,1993). Alternatively, the contrast could be syntactic. Rizzi (1990) has argued that (1)a violates Relativized Minimality, the ban on moving a category over another category of the same type (here, *what* over *who*). The example in (1)b is better because *which problem* contains more features than the intervener *who*, so that the intervention effect is less strong (Rizzi 2017). A third possibility is that lexical *wh*-phrases take longer to process but this makes them easier to retrieve from short-term memory at the gap site (Hofmeister & Sag

2010). Under this view, (1)b is judged better than (1)a because it causes less of a processing burden overall.

In order to distinguish between these proposals, the consequences of each should be worked out, and this proves quite difficult in practice (cf. Sprouse et al. 2013 for an extensive overview of the issues). One relatively recent issue is the potential effect of context. If D-linking improves extraction from islands, then there should be an effect of adding context, and this effect should extend to extraction of bare *wh*-words (cf. Goodall, 2015, p.7). Kush et al. (2019) looked at topicalization out of islands in Norwegian and indeed found that adding a single context sentence ameliorates extraction out of four different island types (only subject islands did not improve). They did not look at the behaviour of complex *wh*-fillers, however, so that it is unclear to what extent context eliminates the observed difference between bare and complex fillers we see in (1). In contrast, Goodall (2015) finds that filler complexity indeed ameliorates extraction out of islands but he finds the same effect in sentences without islands. He concludes that the effect of filler complexity seems general and may therefore be independent of what causes the contrast in (1). Goodall did not include context as a variable, however. The one study that did manipulate both context and complexity (Villata et al., 2016 for French *wh*-islands) only found an effect for complexity. The absence of a context effect in Villata et al. (2016) may be due to the type of context used. The authors themselves note that their context stories may not have been able to induce D-linking (p.91). A closer look at the example context provided in the article suggests another explanation: this context was unnatural for a multiple question as it left open the answer to the identity of only one of the two *wh*-elements, but uniquely specified the other. The context introduced 5 mathematical problems of which the fifth problem was singled out, as only one of the students in class managed to solve it. The context ended with the sentence "You would really like to know who the genius that has been able to do it is!", putting focus on only one of the two *wh*-elements in the target sentence *what/which problem did you wonder who solved*. Thus, the answer to *who* was not named in the context, whereas the answer to *what/which problem* was (*the fifth problem*). This unbalanced information-seeking potential of the two *wh*-elements may have affected the outcomes, in case the other contexts used were of a similar set up.

The effect of context and filler complexity on the amelioration of island violations is thus unclear at present. The goal of the current study is to add to this discussion by manipulating both variables and by using contexts that introduce alternative potential answers for both *wh*-elements in the multiple question. To do so, we ran two acceptability judgement studies. Both contrasted bare *wh*-fillers with complex ones. The first experiment presented sentences in isolation, whereas the second experiment embedded them in a supporting context. This allows us to establish the contribution of each factor in amelioration effects, if any. The language we look at is Dutch, which has been relatively understudied in the discussion on syntactic islands.

2. Experiment 1

Participants

Thirty-two native speakers of Dutch (16 female, mean age = 31.3 years, range = 18 – 59 years) volunteered in an acceptability judgement task. Participants were recruited through personal communication and social media platforms. All participants provided informed consent for their participation in this study.

Design and materials

The experiment had a 2 x 2 design (within-subjects and within-items), crossing the factors STRUCTURE (Non-island vs. Island) and FILLER COMPLEXITY (Bare *wh* vs. Complex *wh*). The materials consisted of 16 sets of sentences, each appearing in all 4 conditions (see lower half of Table 1 for examples).[1] The extracted *wh*-element in all conditions was inanimate, with *wat* 'what' as the bare *wh*-filler and *welke N* 'which N' as the complex *wh*-filler. This was done in order to minimize subject-object ambiguity, which may arise in sentences containing two *wh*-elements (Donkers et al., 2013). The subject of the embedded clause was always animate, with *hij/zij* 'he/she', each in half of the items, in the non-island structures and *wie* 'who' in the island structures. The non-island structures were embedded *dat*-clauses ('that'-clauses) introduced by the matrix verbs *denken* 'to think' or *geloven* 'to believe', each used in half of the items. The island structures involved embedded questions and were always introduced by the matrix verb *zich afvragen* 'to wonder (reflexive)'.

In addition, 36 filler items were constructed covering the full range of acceptability (12 acceptable, 12 moderately acceptable, 12 unacceptable). These items were mostly questions of varying structural types, including yes/no questions, adjunct questions, partial *wh*-movement and *wh*-copying constructions. The unacceptable fillers consisted of agreement errors and word salads, and also included 4 subject islands. All filler items, excluding the subject islands, contained an embedded clause and these were never introduced by the matrix verbs used in the experimental items. The filler items served a double purpose. The first was to counteract a potential response equalization bias. In addition, the unacceptable fillers, in particular the word salad items, served as control items to check whether participants were doing the task seriously.

The experimental items were divided over four counterbalanced lists, using a Latin square design. Each item was presented in only one condition on each list. Each participant thus saw each experimental condition 4 times. The filler items were added to each list, resulting in a total of 52 items per list. Each list was divided into 8 blocks of 6-7 items. Blocks were presented to participants in a randomized order.

Table 1: *sample experimental item. Context was only shown in Experiment 2. Note that the indices and gap positions were not shown to the participants.*

Context: (only Exp.2)	*Hans en Inge moeten een traktatie maken voor een verjaardag. Op de uitnodiging staat dat de jarige graag een appeltaart en een chocoladetaart wil. Inge heeft een duidelijke voorkeur voor één van de twee traktaties.* 'Hans and Inge have to make a treat for a birthday party. On the invitation, it says that the host would like an apple pie and a chocolate pie. Inge has a strong preference for one of the two treats.'

[1] A list of the materials including fillers can be downloaded from the OSF-repository of this study: https://osf.io/d8n7t/

No island	Bare *wh*-filler	*Wat$_i$ denk je dat zij _$_i$ gebakken heeft?* 'What do you think that she baked?'
	Complex *wh*-filler	*Welke traktatie$_i$ denk je dat zij _$_i$ gebakken heeft?* 'Which treat do you think that she baked?'
Island	Bare *wh*-filler	*Wat$_i$ vraag jij je af wie _$_i$ gebakken heeft?* 'What do you wonder who baked?'
	Complex *wh*-filler	*Welke traktatie$_i$ vraag jij je af wie _$_i$ gebakken heeft?* 'Which treat do you wonder who baked?'

Procedure

The experiment was an online acceptability judgement task that was conducted using the Qualtrics survey software (Qualtrics, Provo, UT). Items were presented in isolation in the centre of the page. Participants were instructed to imagine that the test sentences were spoken by a close friend who is a native speaker of Dutch. They were instructed to rate each sentence individually on a 7-point scale with the endpoints labelled *erg slecht* 'very bad' (1) and *erg goed* 'very good' (7) and to rely on their first reaction.

Participants were randomly assigned to one of the experimental lists. Each list started with 3 filler items of differing degrees of acceptability. These filler items served as unannounced practice items to familiarise the participants with the rating scale. It took participants on average about 10 minutes to complete the experiment.

Data analysis

Performance on the unacceptable filler items was checked before data analysis to determine whether participants had faithfully completed the task. Exclusion criteria were set on the unacceptable filler items: at most 4 out of the 12 items, or at most 2 word salad items, should be rated with 4 or higher. No participants had to be removed from further analysis based on these criteria.

The raw acceptability scores were then converted to *z*-scores per participant using all items, to correct for individual differences in scale use. We fitted a linear mixed-effects model to the standardized scores with STRUCTURE, FILLER COMPLEXITY and their interaction as fixed factors, using the *lmer* function from the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017) in R (version 3.6.1; R Core Team, 2019). The initial model included the full random structure permitted by the design of the experiment (Barr, Levy, Scheepers, & Tily, 2013). The random structure was simplified, following the recommendations in Matuschek, Kliegl, Vasishth, Baayen, & Bates (2017), until convergence was reached. Contrasts for categorical factors were effect-coded (i.e., the intercept represents the grand mean): for STRUCTURE, 'no island' was coded -0.5 and 'island' 0.5; for FILLER COMPLEXITY, 'bare' was coded -0.5 and 'complex' 0.5. Significance values for the coefficients from the selected model were calculated using the Satterthwaite approximation in *lmerTest* (Kuznetsova et al., 2017).

Results

Table 2 (left-hand side) shows the mean standardized acceptability judgement scores per condition, which are also shown in Figure 1. Table 3 (left-hand side) shows the mean unstandardized acceptability scores to help locate the results in a more familiar acceptability space.

The final model included a random intercept and random slope for STRUCTURE for items, but no random effects for participants. This model revealed a significant main effect of STRUCTURE (β = -1.92, SE_β = 0.06, 95% CI of β [-2.03, -1.81], p < .001). Sentences with an island structure (M = -.73, sd = 0.41) were consistently judged as less acceptable than their counterparts that did not contain an island structure (M = 1.19, sd = 0.51), regardless of filler complexity. There was no main effect of FILLER COMPLEXITY (β = -0.04, SE_β = 0.04, 95% CI of β [-0.12, 0.04]). Sentences with complex fillers (M = 0.21, sd = 1.08) were not judged consistently better than those containing bare fillers (M = 0.25, sd = 1.05). In fact, a complex filler led to a small numerical increase in acceptability compared to a bare filler if the sentence did not contain an island structure, but to a decrease in acceptability in sentences with an island structure. The interaction between STRUCTURE and FILLER COMPLEXITY was significant (β = -0.15, SE_β = 0.08, 95% CI of β [-0.31, 0.00], p = .05). However, this effect is marginal and should be treated with caution, as it was derived from a model specification that deviated substantially from the maximal model. In sum, we found no evidence of amelioration in Dutch *wh*-islands related to the complexity of the filler.

Table 2: *Mean standardized acceptability judgement scores (z-scores) by STRUCTURE and FILLER COMPLEXITY per experiment (associated standard deviations are given between parentheses).*

	Experiment 1 (n = 32)		**Experiment 2 (n = 32)**	
	Bare *wh*-filler	Complex *wh*-filler	Bare *wh*-filler	Complex *wh*-filler
	M (sd)	*M (sd)*	*M (sd)*	*M (sd)*
No island	1.17 (0.56)	1.21 (0.47)	1.13 (0.61)	1.30 (0.42)
Island	-0.67 (0.46)	-0.79 (0.34)	-0.77 (0.31)	-0.62 (0.47)

Table 3: *Mean unstandardized acceptability judgement scores by STRUCTURE and FILLER COMPLEXITY per experiment (associated standard deviations are given between parentheses).*

	Experiment 1 (n = 32)		**Experiment 2 (n = 32)**	
	Bare *wh*-filler	Complex *wh*-filler	Bare *wh*-filler	Complex *wh*-filler
	M (sd)	*M (sd)*	*M (sd)*	*M (sd)*
No island	6.03 (1.24)	6.12 (1.04)	5.86 (1.37)	6.27 (0.97)
Island	2.01 (1.05)	1.79 (0.85)	1.73 (0.86)	2.03 (1.16)

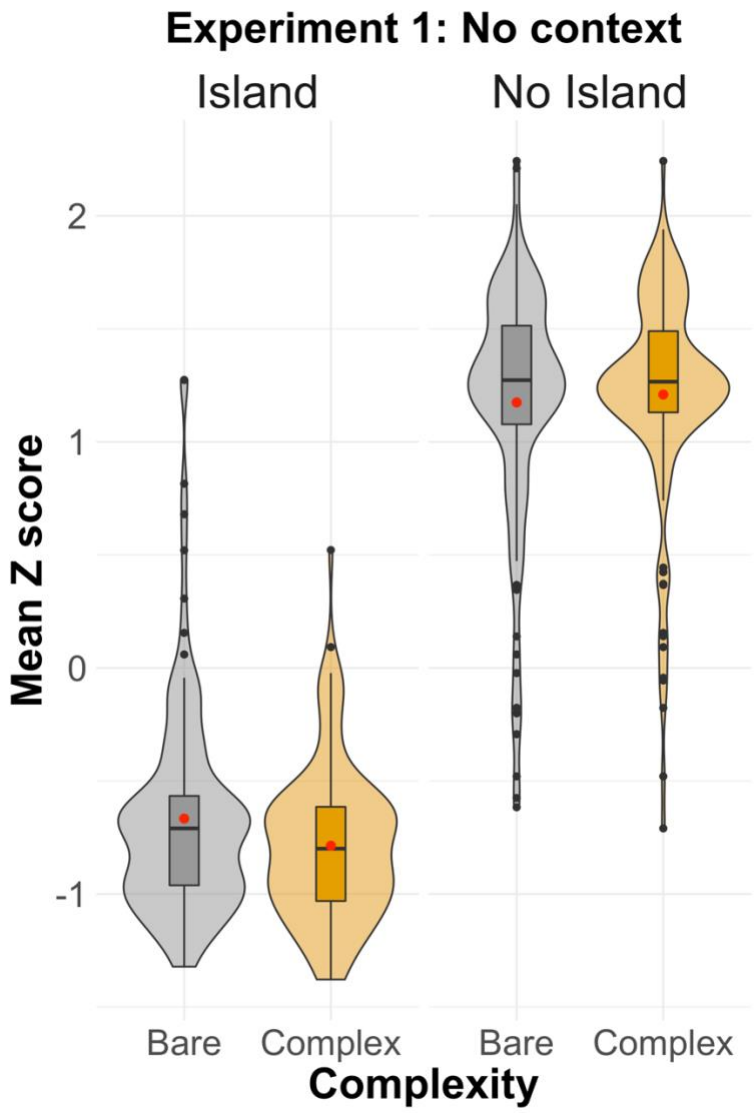

Figure 1: *Violin/boxplot of the standardized acceptability judgement scores (z-scores) by* STRUCTURE *and* FILLER COMPLEXITY *for experiment 1. The red dot encodes the mean, the black vertical line encodes the median, the top and bottom edges of the box encode the interquartile range and the transparent beans provide a density plot.*

3. Experiment 2

Participants

Thirty-two native speakers of Dutch (21 female, mean age = 34.6 years, range = 20 – 79 years) volunteered in an acceptability judgement task, none of whom participated in experiment 1. Participants were recruited through personal communication and social media platforms. All participants provided informed consent for their participation in this study.

Design and materials

Experiment 2 had the same design as experiment 1 with the addition of context. The materials from experiment 1 were embedded in a supporting context (see Table 1 for an example). Every context consisted of three sentences and was structured in the same way. The first

sentence introduced two animate protagonists, one male and one female, and an inanimate cover term (e.g. treat). The second sentence described two specifications of this cover term (e.g. cheesecake and apple pie), which served as potential antecedents to the *wh*-element in the target sentence. Crucially, this allowed for a pair-list reading for the island structures. In the final sentence, the name of one of the animate referents was repeated, and the inanimate referent was referred to using the construction *één van de twee* 'one of the two' + plural noun (e.g. *één van de twee traktaties* 'one of the two treats'). The repeated name in context sentence three was always the name mentioned first in context sentence one. Whether the first context sentence started with a male or female name was counterbalanced across items. The selected gender was matched by the embedded subject pronoun in non-island target sentences. Contexts were presented in present tense.

The 36 filler items were also embedded in contexts that were structurally similar to the ones described for the experimental items. Items were distributed over 4 lists in the same way as described for experiment 1.

Procedure

Procedure was the same as for experiment 1. The context was shown on a separate screen, followed by a new screen on which the test sentence was shown in isolation. It took participants on average about 20 minutes to complete the experiment.

Data analysis

Performance on the unacceptable filler items was again checked before data analysis using the same exclusion criteria as in experiment 1. No participants had to be removed from further analysis.

The raw acceptability scores were then converted to *z*-scores per participant using all items, to correct for individual differences in scale use. The procedure for the statistical analysis was the same as described for experiment 1.

Results

Table 2 (right-hand side) shows the mean standardized acceptability judgement scores per condition, which are also shown in Figure 2. Table 3 (right-hand side) shows the mean unstandardized acceptability scores to help locate the results in a more familiar acceptability space.

The final model included a random intercept and random slope for STRUCTURE and FILLER COMPLEXITY for items, but no random effects for participants. This model revealed a significant main effect of STRUCTURE (β = -1.91, SE_β = 0.06, 95% CI of β [-2.03, -1.79], $p < .001$). Again, sentences with an island structure (M = -0.70, sd = 0.40) were consistently judged as less acceptable than their counterparts that did not contain an island structure (M = 1.22, sd = 0.53), regardless of filler complexity. There was also a main effect of FILLER COMPLEXITY (β = 0.17, SE_β = 0.05, 95% CI of β [0.06, 0.27], p = .006). Sentences with complex fillers (M = 0.34, sd = 1.06) were judged consistently better than those containing bare fillers (M = 0.18, sd = 1.07), regardless of the structure from which they were extracted. This effect was considerably smaller than that of STRUCTURE. The interaction between STRUCTURE and FILLER COMPLEXITY was not significant (β = -0.04, SE_β = 0.07, 95% CI of β [-0.18, 0.11]). This reflects the fact that the effect of complexity was consistent across the different structures.

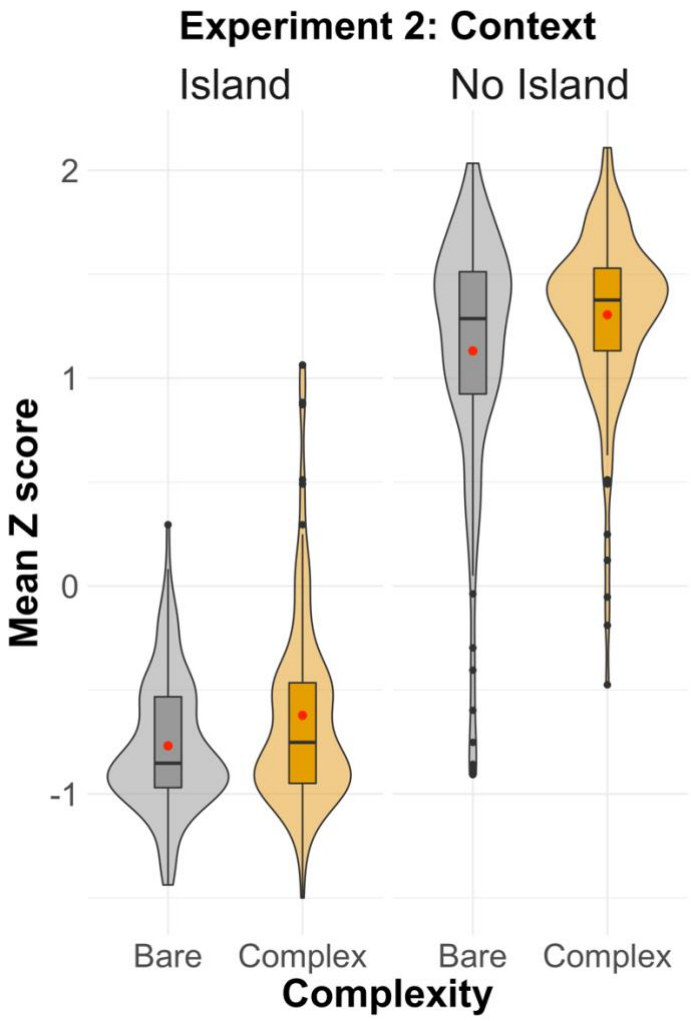

Figure 2: *Violin/boxplot of the standardized acceptability judgement scores (z-scores) by* STRUCTURE *and* FILLER COMPLEXITY *for experiment 2. The red dot encodes the mean, the black vertical line encodes the median, the top and bottom edges of the box encode the interquartile range and the transparent beans provide a density plot.*

Discussion

Experiment 2 showed an amelioration effect for sentences with complex fillers in both island and non-island structures, but the observed differences in acceptability were very small. Such an ameliorative influence of FILLER COMPLEXITY was not observed in Experiment 1. However, the conclusion that the effect of FILLER COMPLEXITY is different in and out of context, requires a statistical comparison of the results from both experiments. An effect of FILLER COMPLEXITY mediated by CONTEXT should surface as a significant interaction between the two factors. A model was fitted on the data of both experiments with the three-way interaction between STRUCTURE, FILLER COMPLEXITY and CONTEXT as fixed effects and with random intercepts for participants and items. This model indeed revealed a significant interaction between CONTEXT and FILLER COMPLEXITY (β = 0.21, SE_β = 0.06, 95% CI of β [0.10, 0.31], p < .001). The effect of FILLER COMPLEXITY was more pronounced when embedded in context. Looking at the pattern, this was not only due to the fact that the sentences with complex fillers were rated higher in experiment 2 than in experiment 1, but also to the fact that the sentences with bare fillers showed the opposite pattern: they were rated lower in experiment 2 than in experiment 1, thus widening the gap between the two types of fillers in experiment 2. In the case of island structures the

numerical patterns for bare and complex fillers were even reversed, with complex fillers scoring lower than bare *wh*-fillers in experiment 1 but higher in experiment 2. This did, however, not result in a significant three-way interaction (β = -0.12, SE_β = 0.11, 95% CI of β [-0.34, 0.09]).

4. General Discussion

The present study investigated the ameliorative potential of filler complexity and context on the acceptability of island violations in Dutch embedded *wh*-questions. Out of context, filler complexity did not have an ameliorative effect on *wh*-islands. Numerically *wh*-islands with complex fillers were in fact rated as slightly less acceptable than when they contained bare fillers but, as stated above, this result should be treated with caution. When embedded in a supporting context, the acceptability of sentences with complex fillers was higher than those with bare fillers. Although this suggests that filler complexity has an ameliorating effect at least in context, two remarks are in order.

First, the difference between bare and complex fillers was not only found because items with complex fillers received slightly higher ratings in context than out of context, but also because items with bare fillers received a slightly lower rating in context than out of context. It is unclear how the different effect that adding context had on bare and complex fillers can be accommodated in any of the approaches to the complexity effect. Even though existing working memory accounts do not explicitly address context effects, it does not seem to follow from their assumptions that bare fillers would do worse in context. A D-linking approach, by contrast, would expect the acceptability of bare fillers to increase in context, if anything. It is unclear what exactly caused our results. We speculate that speakers may have found complex fillers to be more natural continuators of the discourse than bare fillers, despite the fact that we tried to make the contexts equally suitable to both types of fillers. It may turn out to be more natural to use a *wh*-constituent that contains a lexical restriction, and thereby refer back to the lexical nominal in the discourse, than to use a bare *wh*-constituent.

Second, the ameliorating effect of filler complexity in the context condition was also observed in non-island environments, suggesting that the effect is general and not particularly suited to explain a grammatical difference between extractions of bare and complex fillers from *wh*-islands. Since grammatical accounts make no predictions for non-island constructions, the complexity effect must have an extra-grammatical source under any account. If indeed the contexts we used made complex fillers more natural continuators of the discourse than bare fillers, it is in fact expected that the same effect is found in the non-island conditions.

In addition, taking into account that the effect size of any difference in acceptability between the two types of fillers was very small, we conclude that our overall results at most provide evidence for an extra-grammatical complexity effect and that there is no reliable evidence to suggest that complexity causes a grammatical difference between the island violations due to complex or bare fillers. If so, there should be no role in the grammar of Dutch islands for a constraint referring to filler complexity. This would entail that *wh*-islands in Dutch are strong islands (in line with Broekhuis & Corver 2015: section 11.3.1.3, pp. 1397-1399) that cannot be ameliorated by context or complexity. As such, this study provides the first experimental syntactic support for this position. The situation in Dutch is in contrast to what Sprouse et al. (2016:331) found for English and Villata et al. (2016:85) for French. They observed that the complexity effect was restricted to island conditions and did not extend to non-islands. This raises two questions: (i) How can we understand the notion of a strong island

theoretically, and (ii) what causes the difference between Dutch on the one hand and English and French on the other hand with respect to absence or presence of a syntactic complexity effect in *wh*-island violations?

We offer an explanation in terms of featural Relativized Minimality (fRM). It has been proposed that movement of a complex *wh*-filler across an intervening bare *wh*-filler gives a better result than moving a bare *wh*-filler across a bare intervener because a complex *wh*-filler contains more features. The intervention effect is therefore stronger when the intervener shares all the features with the crossing *wh*-filler ('identity') than when it only shares a subset ('inclusion') (cf. Starke 2001; Rizzi 2017). This contrast is schematized in (2), using the feature system of Friedmann et al. (2009):

(2)	a.	*What do you wonder [who bought __]?	IDENTITY
		[Q] [Q]	
	b.	Which car do you wonder [who bought __]?	INCLUSION
		[Q, N] [Q]	

Theoretically, this makes sense. If the position high in the tree probes for a constituent carrying [Q], then Minimal Search (Chomsky 2013) dictates that the embedded subject must be attracted instead of the embedded object. But if there is a position in the left periphery probing for a constituent carrying [Q, N], which is what Villata et al. (2016) and Rizzi (2017) argue for, then the embedded subject can be skipped. The partial overlap then at most gives a degraded result.

In this framework, the features that trigger syntactic movement are essential in calculating relativized minimality effects. Dutch shares with English and French that a *wh*-constituent is triggered to move to clause-initial position, and this happens in main and embedded clauses alike. Now, what distinguishes Dutch from English and French is that Dutch is a verb second language. This entails that in non-*wh*-questions some phrase still has to move to initial position but only in main clauses. Since the nature of the moved constituent (*wh* vs. non-*wh*) and the environment in which it must move (all clauses vs. only main clauses) differ in these two processes, we must conclude that the features triggering *wh*-movement and phrasal movement in a V2 language must be formally distinct. This means that the C-position in a Dutch main clause question contains two rather than one feature triggering movement of a phrasal constituent. Since one of them is [Q], a *wh*-constituent will always be attracted. The nature of the other feature is hard to determine in the absence of any consensus on the trigger for phrasal movement in a V2 language (see Zwart 2011 for discussion). We will therefore informally adopt the feature [F] in order to have an input for the relativized minimality calculus. This feature explicitly indicates that we wish to remain agnostic about the precise nature of the trigger, and at the same time allows our analysis to be compatible with differing concrete triggers. The point of this paper is not to align with a particular take on the trigger for V2 but to make a claim about the absence of a complexity effect. Despite our agnosticism, our analysis still makes clear predictions, as we will note below.

We conclude that the C-position in a Dutch main clause standardly carries the feature [F], causing some phrase to move in front of the finite verb, but in main clause questions also a [Q]-feature, which limits the search for a suitable phrase to *wh*-constituents. Let us now turn to island violations in Dutch and contrast extraction of a complex and bare filler, schematized in (3):

(3) a. *Wat$_i$ vraag jij je af wie $_i$ gebakken heeft? IDENTITY
[Q, F] [Q, F]
'What do you wonder who baked?'

b. *Welke traktatie$_i$ vraag jij je af wie $_i$ gebakken heeft? IDENTITY
[Q, F] [Q, F]
'Which treat do you wonder who baked?'

The relevant point is that the embedded subject is in principle also capable of moving to the sentence-initial position and satisfying the verb second constraint. With respect to the relevant triggers, then, moving a complex or bare *wh*-filler across a bare *wh*-intervener creates a situation of identity in either case. Of course, the complex filler in (3) also carries the [N] feature, in contrast to the bare filler, but this feature does not play a role in the relevant triggers and therefore does not enter the relativized minimality calculus. After all, it is not the case that only nominal constituents can participate in satisfying the V2 constraint in Dutch.

To conclude, complexity does not lead to an amelioration effect in Dutch because it does not avoid an identity relation between the two fillers, given the V2 constraint. The fact that complexity effects have been found in English and French, both of which are not V2 languages, can be taken to underscore our analysis, but some caution is warranted.

First of all, featural Relativized Minimality does not predict that lack of verb second entails a syntactic complexity effect. For such an effect to arise, the language should not only lack the V2 property but also have a syntactic position that probes for a [Q, N] constituent, which has to be justified for each language individually (see Villata et al. 2016:79 for evidence in different languages). A more straightforward prediction of our analysis is therefore that, all things being equal, V2 languages should not display grammatical complexity effects in *wh*-islands. Future research will need to tell us if this is borne out. Another prediction, brought to our attention by an anonymous reviewer, is the following. Languages in which *wh*-islands are weak islands are known to show argument-adjunct asymmetries: extracting a *wh*-argument from a *wh*-island gives a better result than extracting a *wh*-adjunct (see for instance Sabel 2002). If our analysis is on the right track, no such asymmetry is expected for Dutch since both arguments and adjunct can satisfy the V2 constraint.

Second, it remains to be seen how robust the results are. Goodall (2015) for instance finds for English that complexity also leads to higher acceptability judgements in non-island conditions, suggesting that the effect is extra-grammatical. This is similar to what we find in the context conditions in Dutch but contrasts with what Sprouse et al (2016) find for English, and Villata et al (2016) for French, both with and without context. Future research will have to establish for each language displaying a complexity effect whether the effect is overall or only plays a role within island violations. It is only in the latter case that the grammar can be given an explanatory role.

Acknowledgments
We would like to thank two anonymous reviewers, as well as the audience at the *Grote Taaldag* for very useful comments.

References

Barr, Dale J., Levy, Roger, Scheepers, Christoph., & Harry J. Tily. (2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal." *Journal of Memory and Language* 68 (3): 255–278. doi: 10.1016/j.jml.2012.11.001

Broekhuis, Hans & Norbert Corver. 2015. *Syntax of Dutch: verbs and verb phrases. Volume 3*. Amsterdam: Amsterdam University Press.

Chomsky, Noam. 2013. "Problems of Projection." *Lingua* 130: 33–49.

Donkers, Jantien, Hoeks, John & Laurie Stowe. 2013. "D-Linking or set-restriction? Processing which-questions in Dutch." *Language and Cognitive Processes* 28 (1-2): 9–28. http://dx.doi.org/10.1080/01690965.2011.566343

Friedmann, Naama, Belletti, Adriana & Luigi Rizzi. 2009. "Relativized relatives: types of intervention in the acquisition of A-bar dependencies." *Lingua* 119 (1): 67–88.

Hofmeister, Philip & Ivan A. Sag. 2010. "Cognitive constraints and island effects." *Language* 86 (2): 366–415. DOI: https://doi.org/10.1353/lan.0.0223.

Goodall Grant. (2015). "The D-linking effect on extraction from islands and non-islands." *Frontiers in psychology* 5: 1493. https://doi.org/10.3389/fpsyg.2014.01493

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

Kuznetsova, Alexandra, Brockhoff, Per & Rune Christensen. 2017. "lmerTest Package: Tests in linear mixed effects models." *Journal of Statistical Software* 82 (13): 1–26. doi: 10.18637/jss.v082.i13

Matuschek, Hannes, Kliegl, Reinhold, Vasishth, Shravan., Baayen, Harald & Douglas Bates. 2017. "Balancing Type I error and power in linear mixed models." *Journal of Memory and Language* 94: 305–315. doi: 10.1016/j.jml.2017.01.001

R Core Team (2019). *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing*. Available at: https://www.R-project.org/.

Rizzi, Luigi. 1090. *Relativized minimality*. Cambridge: Cambridge University Press.

Rizzi, Luigi. 2017. "Comparing extractions from wh-islands and superiority effects." *Wiener Linguistische Gazette* 82: 253-261.

Sabel, Joachim. 2002. A minimalist analysis of syntactic islands. *The Linguistic Review* 19: 271–315.

Szabolcsi, Anna & Frans Zwarts. 1993."Weak islands and an algebraic semantics of scope taking. *Natural Language Semantics* 1: 235–284. doi:10.1007/BF002 63545

Sprouse, Jon, Caponigro, Ivano, Greco, Ciro & Carlo Cecchetto. 2016. "Experimental syntax and the variation of island effects in English and Italian." *Natural Language and Linguistic Theory* 34: 307–344.

Starke, Michal. 2001. *Move dissolves into merge: a theory of locality*. Doctoral dissertation. University of Geneva.

Villata, Sandra, Rizzi, Luigi, & Julie Franck. 2016. "Intervention effects and relativized minimality: new experimental evidence from graded judgements." *Lingua* 179: 76–96. DOI: http://dx.doi.org/10.1016/j.lingua.2016.03.004

Zwart, Jan-Wouter. 2011. *The syntax of Dutch*. Cambridge: Cambridge University Press.