Linguistic acceptability judgments are a critical part of the toolkit for linguistic investigation. An implicit assumption of much work using these judgments is that the data are stable, presumably reflecting the underlying nature of grammatical representation. Here, I demonstrate that for a range of constructions — including so-called island-violating sentences, those with resumptive pronouns, case errors, inflectional errors, and processing complexities — judgments systematically shift upward with repeated exposure to the offending syntactic structure. The data here show that these upward shifts cannot be solely attributed to complex sentences becoming easier to process, or ungrammatical sentences becoming grammatical: the effects are ubiquitous, although they only arise given sufficient exposure. Based on preceding work that highlights the role of adaptation and expectations in online processing and grammar, I consider these effects in terms of probabilistic expectations over syntactic structures that shift with input that diverges from prior distributions.

1 Introduction

Linguistic acceptability judgments form the backbone of much of what is postulated about the nature of grammar. In the course of many syntactic investigations, the process often begins with confident, assertive acceptability contrasts that neatly divvy up linguistic data into good vs. bad, natural vs. unnatural, or grammatical vs. ungrammatical. In the process of widening the data net and revisiting the same data points over and over, judgments frequently lose some of their sharpness and clarity. In some cases, this degradation process reaches the point that the researchers abandon the use of their own intuitions altogether, “forsaking intuitions for commonly assumed ungrammatical sentence types and relying on memorized judgments as a source of stability” (Francom 2009, p. 22). Usually, the change in judgments is upwards. What was once terrible now sounds only awkward, or what once triggered a question mark annotation now seems perfectly acceptable.

This phenomenon is not unique to linguists, despite being once labeled as ‘linguist’s disease’. Even in experimental sessions that last less than an hour, linguistically naïve participants behave similarly, shifting their judgments upward for structurally similar sentences. This was first shown formally by Snyder (2000) (but see also Nagata 1988, Hiramatsu 2000, and footnote 1), investigating the role of repetition on judgments for so-called island-violating sentences. Snyder showed that judgments for some island-violating sentences like “Who does Mary believe the claim that John likes?” increased with exposure to syntactically similar sentences. In the same study, though, sentences that violated other proposed island constraints, like the Adjunct and Left Branch constraints, failed to show such an increase or ‘satiation’ effect.

Sprouse (2007, 2009) questioned the reliability of these results on the grounds that 70% of the trials in Snyder’s experiment were arguably ungrammatical. This potentially led participants to employ a strategy where they changed their minds from ‘no’ (ungrammatical) to ‘yes’ (grammatical) for some island-violating sentences in order to produce an equal number of yes/no responses. In several replication studies with putatively equal numbers of ‘good’ and ‘bad’ sentences, Sprouse found little to no evidence of satiation for island-violating sentences and raised the possibility that “asymmetric extra-grammatical effects on acceptability may be a useful tool for identifying the grammatical status of structures that are neither clearly grammatical or clearly ungrammatical”. What improves with repetition may thus depend on grammatical status, i.e., satiation may be a litmus test for grammaticality (see Snyder 2000 for the same idea).

Since that time, a number of studies have found independent evidence that judgments for sentences with a variety of island violations rise with repeated exposure in experimental contexts without the problems of...
Snyder’s original design (Francom, 2009, Hofmeister and Sag, 2010, Hofmeister et al., 2013a, Chaves and Dery, 2013). Yet others have replicated the findings of Sprouse that judgments for island-violating sentences do not vary significantly with repeated exposure (Crawford, 2011). This naturally raises the question of why some sets of materials demonstrate such repeated exposure effects but not others.

As noted by Hofmeister et al. (2013b), the relevant studies differ on numerous dimensions, including presentation method, rating scale, the particular island constructions tested, whether the stimuli were direct vs. indirect interrogatives, and how the data were analyzed. In the current investigation, I take a wider view of this phenomenon, going beyond the domain of island effects. That is, in addition to revisiting island effect data, the goal of the current investigation is to inspect how judgments for a range of sentences vary with repeated exposure, including those with resumptive pronouns, case violations, agreement mismatches, center-embeddings, and inflectional errors. The hope is that inspecting a wider range of phenomena will provide a more incisive look at the dynamics and preconditions for this experience-driven effect.

In addition, the experiments described below add to the foregoing literature with several wrinkles. First, many of the previously described experiments examined only how repeated exposure affects judgments for island-violating sentences, but not minimally different ones, i.e., those that retained the island structure but did not violate it. Across the full range of constructions I investigate here, a marked or less acceptable construction is paired with at least one minimally different version. Second, unlike the majority of acceptability experiments reported in the literature, the judgment materials I use appear with comprehension questions which not only provide a check of reading accuracy but also ensure that participants approach the materials in a way that attempts to approximate how they process language in normal life — that is, they should try and derive meaning from the sentences.

The picture that emerges from the studies here is that repeated exposure effects are ubiquitous. With repetition of structurally related items, judgments rise for almost every single “unacceptable” structure. For the one counterexample — agreement ‘attraction’ errors — there is a fairly obvious explanation for why judgments don’t rise: judgments are already quite high.

To be clear, the current set of studies was not designed to definitively explain why judgments shift, although some possible explanations are indeed excluded. Given that judgment tasks themselves reduce a complex set of variables down to a simplified unitary response, it is highly likely that multiple factors are responsible for judgment stability and shifts. Instead of attempting to enumerate and quantify each of these variables, the primary aims here are (i) to show that repeated exposure effects are associated with multiple island-violating structures but extend well beyond the domain of islands and (ii) to exclude several candidate explanations for why these shifts occur. One possibility I exclude is the suggestion that only judgments for “grammatical” sentences shift with repeated exposure.

In Experiment 1, I re-examine a well-studied island effect tied to embedded wh-questions, alongside several unrelated constructions where acceptability contrasts are uncontroversially ascribed to either grammaticality differences or processing complexity. In Experiment 2, I test the straightforward idea that repeated exposure effects are easier to detect given more repeated exposures. The purpose here is to assess whether null result findings relate to the number of related stimuli in an experimental context. This question is evaluated in the context of an island for which previous studies have found no signs of satiation — adjunct island violations — alongside various other constructions, including case violations and agreement errors. Experiment 3 complements these results by testing for effects of repeated exposure in a context where unacceptability cannot be attributed to processing difficulty, according to the self-paced reading results of Hofmeister and Norcliffe (2014).

Much of the concluding section of this article is occupied with the question of why such effects arise.
Following ideas in Chaves (2013), one coherent explanation of the findings is that expectations borne from experience play a significant role in acceptability judgment data. That is, like online behavioral measures of language processing, offline judgments may be sensitive to the statistically-driven expectedness of syntactic structures (Fine et al., 2010, 2013, Lau et al., 2006, Levy, 2008, Levy et al., 2012, Staub and Clifton, 2006, Staub et al., 2006). For instance, experience with English unbounded dependency constructions may lead to probability distributions over possible dependency paths and how the dependency terminates, e.g., gap vs. resumptive. The acceptability of any particular realization would thus be tied to its (un)expectedness. A prediction of such a hypothesis is that judgments for such constructions are mutable (see Fine et al. 2013). That is, if new linguistic input runs counter to previous linguistic experience, then judgments will shift accordingly. Presenting rare or anomalous structures repeatedly within a linguistic context, for instance, would trigger changes in the associated probability space, making the previously unexpected structure less surprising. In contrast, presentation of a relatively expected structure would require little modification to the probability space, leading to weaker, if any, adaptation effects.

A connection to usage, which is virtually synonymous with experience, is an integral part of numerous preceding treatises on grammar (Bybee and Hopper, 2001, Bybee et al., 1994, Goldberg, 2006, Kemmer and Barlow, 2000, Langacker, 2000, MacWhinney, 1998, Tomasello, 2003). What the current research adds is a direct view of this relationship via the most widely used tool for investigating sentence grammar — acceptability judgments. Given that the bulk of the insights in the syntactic literature depend on acceptability judgments, whether gathered in a formal or informal fashion, claims about the connection between usage — and by extension, experience — and grammar have a natural starting point with judgment data.

2 Experiment 1

Unbounded dependencies in English and many other languages can, in principle, be arbitrarily long. Steep drops in acceptability often arise, however, when such dependencies cross particular syntactic boundaries or ‘barriers’, such as relative clause and interrogative clause boundaries (Ross, 1967, Chomsky, 1962, 1973, 1977, 1981, 1986), compared to sentences without any unbounded dependency or one that does not cross the relevant boundary (Sprouse, 2007, Sprouse et al., 2012). In the example in (1a), for instance, the wh-dependency crosses an interrogative clause boundary (marked with a left bracket below), putatively blocking formation of the dependency on standard transformational accounts of syntax:

(1) a. Which immigration bill was Obama unsure [whether Democrats would be able to pass through Congress]?
b. Obama was unsure [whether Democrats would be able to pass the immigration bill through Congress].

In contrast to the grammatical treatment, many island effects have been reinterpreted as a consequence of simultaneous processing costs. Many island-violating sentences pose numerous processing challenges, including long-distance dependencies spanning multiple new discourse referents, opportunities for garden-pathing or misanalysis, syntactically and semantically similar discourse referents, vague or non-specific filler phrases, and so on (Deane, 1991, Kluender, 1991, 1992, 1998, Kluender and Kutas, 1993, Hofmeister and Sag, 2010, Hofmeister et al., 2013a). In addition, while grammatical accounts predict across-the-board unacceptability of island-violating sentences, some tokens elicit relatively high ratings, and attested counterexamples to essentially every proposed island constraint have been observed in naturalistic speech or writing (see Hofmeister and Sag 2010 for a review). Such facts, as well as evidence of the online processing difficulty of such dependencies, motivate the view that a variety of island effects reflect cognitive resource limitations rather than universal syntactic constraints on dependency formation. At the heart of such accounts is the notion that multiple, simultaneous processing costs combine in such a way that severe drops in acceptability ensue, much as they do in the case of center-embedded constructions (Chomsky and Miller, 1963, Bever, 1970).
Experiment 1 tests how judgments change, if at all, for sentences that violate the Wh-island Constraint, compared to judgments for minimally different sentences. As noted above, the foregoing evidence is mixed on whether judgments for island-violating sentences shift with repeated exposure (see Sprouse 2009 for a review). The current experiment differs from the past investigations by the inclusion of a minimally different version and comprehension questions, in addition to differences in statistical analysis.

2.1 Design

The materials varied in terms of whether an unbounded dependency stayed at the level of the matrix clause or crossed an embedded clause boundary.

(2) a. The screen showed the caravan of vehicles that military officials hadn’t decided whether they should bomb yet.

b. There was a group of military officials who hadn’t decided whether they should bomb the caravan of vehicles yet.

Based on prior psycholinguistic findings and acceptability studies, it is expected that a dependency that enters into an embedded clause will lead to lower ratings than one that does not. 28 items in total were included in the materials, and each participant saw only condition of each item.

68 additional items accompanied these island materials. Among these were two sets of items that provide a means for better interpreting any changes in the judgment data relating to island effects. One set involved a classic contrast in grammaticality; while the other set contrasted processing complexity. Including these materials, therefore, allows for the assessment of how repeated exposure effects relate to conventional notions of grammaticality and complexity.

One set consisted of 20 items that appeared in two conditions, which varied in the number of inflectional errors in the sentence, viz., a finite verb occurs where a non-finite form is standardly used in English. In one condition, no errors occurred; in another, two inflectional errors occurred within the sentence, e.g., “The Arctic’s summer ice coverage could (hold/holding) its ground or even bounce back slightly from its earlier losses, even as global average temperatures (rise/rising) this year.” Again, participants saw only version. The second set was comprised of 12 sentences that varied in terms of complexity: one version contained a nested dependencies (COMPLEX) while the other had a right-branching structure (SIMPLE).

(3) a. There was an activist who a reporter that Nancy trusted wrote that Robert bribed without thinking about it.

b. There was an activist who wrote that Robert bribed a reporter that Nancy trusted without thinking about it.

The remaining 36 fillers contained a range of unrelated constructions. An analysis of the fillers (available upon request) revealed that a majority (3/4) received ratings above the experimental mean.

Judgments were collected via Amazon.com’s Mechanical Turk marketplace. Gibson et al. (2011), Sprouse (2011), and others have shown that collecting judgments via Mechanical Turk yields results that parallel those obtained with traditional laboratory methods, while simultaneously providing the advantages of large participant numbers and relatively short data gathering times. At the beginning of each list, the instructions informed participants that they should provide judgments of how good or bad the sentences sounded to them on a scale of 0-10. They were instructed to avoid rating the test items on the basis of prescriptive norms. Following each test sentence, a yes/no comprehension question acted as means for ensuring that participants comprehended the test items, e.g., Were bombs dropped on the caravan?

Datasets from any participants who indicated that English was not their native language were removed prior to analysis.

2.2 Participants

72 participants with IP addresses inside the United States received cash for taking part in the study. All

---

3This is a somewhat atypical Likert scale, but there is no reason to think the results are contingent on this choice. The experiment described here is a replication of a nearly identical experiment (more island violations per list were included in this version) using a 1-7 scale that yielded qualitatively similar results.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>embedded</th>
<th>matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISLANDS</td>
<td>5.90 (.08)</td>
<td>7.63 (.07)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ungrammatical</th>
<th>embedded</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRAMMATICALITY</td>
<td>2.62 (.11)</td>
<td>4.18 (.13)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>complex</th>
<th>simple</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPLEXITY</td>
<td>3.49 (.11)</td>
<td>6.73 (.10)</td>
</tr>
</tbody>
</table>

Table 1: Mean acceptability ratings by dataset and condition in Experiment 1. Values in parentheses indicate standard errors.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Factor</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Islands</td>
<td>DEPENDENCY[EMBEDDED=1]</td>
<td>−0.300</td>
<td>0.038</td>
<td>−7.94</td>
</tr>
<tr>
<td>Item Order</td>
<td>0.020</td>
<td>0.004</td>
<td>5.37</td>
<td></td>
</tr>
<tr>
<td>DEPENDENCY[EMBEDDED=1] × Item Order</td>
<td>0.011</td>
<td>0.034</td>
<td>3.10</td>
<td></td>
</tr>
<tr>
<td>Grammaticality</td>
<td>GRAMMATICALITY[GRAMM.=1]</td>
<td>−0.552</td>
<td>0.053</td>
<td>−10.39</td>
</tr>
<tr>
<td>Item Order</td>
<td>0.036</td>
<td>0.009</td>
<td>4.15</td>
<td></td>
</tr>
<tr>
<td>GRAMMATICALITY[GRAMM.=1] × Item Order</td>
<td>0.016</td>
<td>0.009</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>COMPLEXITY[SIMPLE=1]</td>
<td>0.286</td>
<td>0.030</td>
<td>9.44</td>
</tr>
<tr>
<td>Item Order</td>
<td>0.123</td>
<td>0.034</td>
<td>3.60</td>
<td></td>
</tr>
<tr>
<td>COMPLEXITY[SIMPLE=1] × Item Order</td>
<td>−0.031</td>
<td>0.040</td>
<td>−0.78</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Model summary for each Experiment 1 dataset's fixed effects. Each model contains a predictor for the primary manipulation, item order, and their interaction. Bracketed material indicates the treatment level of the experimental manipulation whose corresponding statistics are shown. T-values above the absolute value of 2 are considered statistically reliable.
participants self-identified as native speakers of English. 1 participant failed to complete the experiment, and 2 others answered with less than 75% accuracy. The data from these 3 participants were removed prior to analysis.

2.3 Analysis

Linear mixed effect regression models were constructed for the data associated with each set of materials using the lme4 package (version 1.1.-7) in R (version 3.1.2). Each model contained a predictor for the item order of the relevant trial, relative to other items in the same condition. This item order term was centered in each model to minimize collinearity. Secondary analyses are reported in the appendix that include models with overall trial position, which is unsurprisingly correlated with item order. Models also included a categorical factor for the given sentence manipulation. This 2-level factor was sum coded prior to analysis. For all models discussed here, I used the maximal random effect structure: random slopes for each main effect and their interaction were included for participants and items; however, any pathological terms were removed.

Acceptability z-scores were computed for each participant, and these scores act as the dependent variable. Thus, for the case of island effects, a formulation in R would be $z_{\text{score}} \sim \text{Condition} \ast \text{Order} + (\text{Condition} \ast \text{Order} \mid \text{Participant}) + (\text{Condition} \ast \text{Order} \mid \text{Item})$. Although models with nested random effect structures do not directly yield p-values, significance at the .05 level can be conservatively estimated for fixed effects coefficients with t-values which have absolute values at or above 2 (Baayen, 2008, Baayen et al., 2008, Pinheiro and Bates, 2000). All fixed effect correlations were below .2, except for those models reported in the appendix where item order strongly correlated with trial position.

2.4 Results

Model summaries for the fixed effect predictors in each dataset are provided in Table 2. Condition means are listed in Table 1.

Replicating past findings, island-violating sentences were judged to be less acceptable than sentences with island structures that a dependency did not cross, as shown by the negative coefficient in Table 2 for the island-violating condition. An overall effect of item order is also evident, but Figure 1 conveys that island-violating sentences drive this main effect. Supporting this interpretation, the main effect is qualified by an interaction of item order and dependency type (MATRX vs. EMBEDDED). A separate model fit to only the data from the island-violating condition confirms
that judgments rose significantly over the course of the experiment for such materials ($\hat{\beta} = 0.031$; SE = 0.006; $t = 4.85$). By contrast, the effect of item order does not reach conventional levels of significance, though there is an upward trend ($\hat{\beta} = 0.010$; SE = 0.006; $t = 1.83$).

For sentences with multiple grammatical errors, judgments were significantly lower than for those without errors. As with island violations, a main effect of item order is qualified by an interaction of condition and item order. This reflects the fact that judgments rose with the repetition of the ungrammatical sentences ($\hat{\beta} = 0.051$; SE = 0.011; $t = 4.35$) at a steeper rate than they did for minimally different sentences without inflection errors ($\hat{\beta} = 0.020$; SE = 0.008; $t = 2.30$), as depicted in Figure 2.

Unsurprisingly, sentences with greater complexity received lower judgments than sentences with less complexity. The model report also indicates a main effect of item order, but unlike in the other two cases, there is no reliable interaction. Inspection of Figure 3 reveals that judgments for both types of sentences increased with exposure (simple: $\hat{\beta} = 0.095$; SE = 0.047; $t = 2.03$; complex: $\hat{\beta} = 0.159$; SE = 0.048; $t = 3.34$). Arguably, this is connected to the fact that even the “simple” condition remains quite complex and difficult to parse, which is supported by the notably low ratings for this condition.

### 2.5 Discussion

For a range of constructions, repeated exposure to structurally-related items led to higher acceptability judgments. The effect of repeated exposure was most dramatic for conditions deemed relatively unacceptable to begin with. For the case of island-violating sentences, participants found them increasingly acceptable as they saw more and more such constructions, and to a lesser, judgments rose for minimally different sentences without such violations. A similar observation applies to sentences with grammatical errors. For sentences varying in complexity, judgments for both kinds of sentences shifted upward with repetition; however, this arguably reflects the fact that both types of sentences were hideously complex (Experiment 2 remedies this by using far easier “simple” sentences). Collectively, the evidence here demonstrates that judgments shift upwards for not just island-violating sentences, but for those whose unacceptability can be tied to either grammatical or processing constraints.

As nearly every construction evidenced some effect of repeated exposure, a reasonable question to ask here is whether the observed effects represent a type of task effect or whether they are attributable to repeated presentations of structurally-related items. That is, it is conceivable that judgments are rising across-the-board in response to task conditions, e.g., fatigue, loss of attention, etc.

While it is impossible to completely exclude a task-related effect, there are several reasons to think this cannot account for the pattern of data seen here. If overall trial position was the sole determinant here, we would expect to see no differences across conditions. That is, all conditions would be subject to a roughly equivalent effect of repeated exposure. Yet, the evidence strongly indicates that the steepest changes in judgments were associated with the relatively unacceptable conditions. In addition, it is possible to include overall trial position as a covariate in the models, although this introduces an unavoidable amount of collinearity, as overall trial position and item order are highly correlated. For the island dataset, by way of example, a model that includes a main effect of overall trial position yields a high correlation with item order ($r = -.93$). This multicollinearity exclusively affects the estimates for the main effects, meaning that the interaction of item order and condition remains interpretable. Across the three datasets, this interaction of item order and condition remains significant even after the inclusion of an overall trial position covariate. Finally, the additional experiments discussed here reveal a clearer picture where overall task effects seem to be significantly reduced or altogether absent.

---

4Residualization will not resolve the issue here. The issue with multicollinearity is that some amount of variance in the data is not uniquely attributable to one particular factor. Residualizing will, by stipulation, associate this variance with one factor; however, this does not address the underlying ambiguity.
Figure 2: Acceptability scores by list position with smoothed loess curve for 20 grammaticality items in Experiment 1. Shaded area shows 95% confidence interval.

Figure 3: Acceptability scores by list position with smoothed loess curve for 12 complexity items in Experiment 1. Shaded area shows 95% confidence interval.
3 Experiment 2

In all three datasets considered in Experiment 1, effects of repeated exposure were observed. This situation contrasts with some prior studies that have failed to validate such effects. One reason for these null effects may be that the number of repeated trials were insufficient or that the studies were otherwise underpowered. In particular, studies on this topic vary considerably in how many similar trials participants see. Crawford (2011), for instance, uses only 4 instances of Subject Island Constraint violations, whereas Chaves and Dery (2013) use 14; while the former reports no satiation effects, the latter does. The observation of repeated exposure effects thus may at least partly depend on the number of related trials. Here, the focus will be on the case of the Adjunct Island Constraint, which putatively blocks extraction from adjuncts. By considering another case of island-violations, we also increase the generalizability of the findings here as applied to unbounded dependency phenomena.

3.1 Design

Two groups of participants read and judged the materials for this experiment. For four different datasets, the materials were distributed across experimental lists such that one group saw twice as many as tokens as the other group, controlling for overall list length. In the island materials for this experiment, sentences with dependencies into adjunct islands were contrasted with non-island-violating sentences, as in (4).

(4) a. Just a few years ago, Mosul was a city which terrorists would have thought twice before attacking.

b. Just a few years ago, terrorists would have thought twice before attacking the city of Mosul.

List 1 contained 24 such island items (12 per condition), while List 2 had 12 (6 per condition). Three additional material types were included in each experimental list, varying in terms of (i) complexity, (ii) grammatical vs. ungrammatical case-marking, and (iii) grammatical vs. ungrammatical subject-verb agreement.

The example in (5) exemplifies materials that varied with respect to complexity: a right-branching structure is used in (5a) vs. a center-embedded structure in (5b). List 1 had 12 such items; List 2 had 24. Another set of sentences differed in the case marking of an object pronoun (nominative vs. accusative), as shown in (6). List 1 carried 24 of these items, while List 2 had 12. A third set contrasted agreement marking on a copula, as depicted in (7). List 1 had 12, and List 2 had 24.

(5) a. The rebels in the jungle captured the diplomat who pleaded with the villagers after they threatened to kill his family for not complying with their demands.

b. The diplomat who the rebels who were in the jungle captured pleaded with the villagers after they threatened to kill his family for not complying with their demands.

(6) a. The psychologist who counseled Elizabeth instructed she that negative thoughts can be chased away by positive ones.

b. The psychologist who counseled Elizabeth instructed her that negative thoughts can be chased away by positive ones.

(7) a. The advertisement on the skyscrapers was hard to read except from a long distance away.

b. The advertisement on the skyscrapers were hard to read except from a long distance away.

Participants saw only one condition of each item, and all materials were pseudorandomized for presentation order. Each group of participants read and judged an equal number of total sentences in the experiment. The net effect is that each group of participants judged 18 sentences with case or agreement problems. Moreover, of the 72 sentences each group rated, 36 were predicted to be relatively unacceptable, and 36 relatively acceptable.

3.2 Analysis

All analytical steps followed those described for Experiment 1. Separate hierarchical models were run for each group of participants and data.
3.3 Participants

96 participants (48 per version) received between $1.50 and $2.50 for taking part in the study. Participants were restricted from performing both acceptability tasks. All participants used US IP addresses and self-identified as native English speakers.

3.4 Results

Details of the effect of item order and its interaction with condition across datasets are listed in Table 4 and 5 for List 1 and List 2, respectively. Condition means are shown in Table 3.

The results reveal that judgments for adjunct island violations rose with repeated exposure; however, this effect is only reliable given twelve repetitions. In fact, item order interacts with island condition in List 1 (= 12 repetitions per condition). As revealed by separate mixed effects models per condition, this interaction reflects the fact that judgments for the island-violating condition increased significantly with exposure (\( \hat{\beta} = 0.038; \ SE = 0.012; \ t = 3.31 \)), unlike in the non-island-violating condition (\( \hat{\beta} = 0.012; \ SE = 0.008; \ t = 1.46 \)). In List 2 (= 6 repetitions per condition), however, no main effect of item order is evident, nor is there an interaction. Separate fits confirm that neither condition improves with only six repetitions (ISLAND \( \hat{\beta} = 0.049; \ SE = 0.039; \ t = 1.29 \); NON-ISLAND \( \hat{\beta} = 0.025; \ SE = 0.025; \ t = 1.01 \)). Thus, had List 2 been the only source of evidence, the conclusion would most likely be that sentences with dependencies into adjunct islands do not sound increasingly better with exposure.

As seen in Table 4, a main effect of item order is observed for the dataset with the case manipulation in List 1. Here, as in the case of the islands dataset, judgments also rose over the course of twelve presentations of case-violations (\( \hat{\beta} = 0.026; \ SE = 0.009; \ t = 3.08 \)), but not for sentences without such violations (\( \hat{\beta} = -0.003; \ SE = 0.006; \ t = -0.53 \)), leading to an interaction. The same main effect of item order recurs in List 2, although no interaction emerges in this case, as judgments in both conditions climb with item order (NOMINATIVE: \( \hat{\beta} = 0.071; \ SE = 0.036; \ t = 1.97 \); ACCUSATIVE: \( \hat{\beta} = 0.045; \ SE = 0.026; \ t = 1.77 \)).

In both lists, agreement errors fail to yield higher judgments as item order increases: there are no significant main effects of item order or interactions of item order and condition. Conversely, across both lists, there is a parallel relationship between item order and sentences that vary in processing complexity: judgments for complex, nested sentences rise with repeated exposure (List 1: \( \hat{\beta} = 0.097; \ SE = 0.024; \ t = 4.08 \); List 2: \( \hat{\beta} = 0.051; \ SE = 0.009, \ t = 5.41 \)), unlike judgments for right-branching sentences (List 1: \( \hat{\beta} = -0.007; \ SE = 0.025; \ t = -0.29 \); List 2: \( \hat{\beta} = 0.012; \ SE = 0.010; \ t = 1.20 \)).

3.5 Discussion

As in Experiment 1, judgments for a variety of constructions, including island violations, appear to be mutable. Notably, the effects were more pronounced when there were more repetitions. The methodological takeaway here is straightforward: for the best chances of observing exposure effects, the number of repetitions must be sufficiently high, although what counts as sufficient is not answered here. In addition, it is evident that the magnitude and timing of repeated exposure effects varies from construction to construction. While the island-violating cases were found to improve only given twelve exposures, judgments for sentences with case-violations and nested dependencies rose reliably even with just six. At the moment, it would be premature to argue whether such differences are idiosyncratic from construction to construction. That said, it is worth noting that both the sentences with case-violations and those with nested dependencies triggered far lower judgments than sentences with adjunct island-violations. A reasonable hypothesis to test, therefore, is that the rate that judgments shift depends on how low acceptability judgments are on the first presentation of some condition.

The only dataset that showed no effect of item order involved a manipulation of subject-verb agreement. In the agreement error examples, however, all stimuli involve number attraction, e.g., “The key to the cabinets are …” Such configurations notoriously cause people to overlook the number mismatch between the head noun and the verb, which explains the fact that these conditions received notably high ratings. Moreover, these attraction errors are not uncommon in natural speech. Based on this, I suggest that judgments for them remain stable because individuals already assign
<table>
<thead>
<tr>
<th>Dataset</th>
<th>island</th>
<th>non-island</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJUNCT ISLANDS</td>
<td>LIST 1</td>
<td>6.72 (.11)</td>
</tr>
<tr>
<td></td>
<td>LIST 2</td>
<td>6.38 (.16)</td>
</tr>
<tr>
<td>S-V AGREE</td>
<td>mismatch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIST 1</td>
<td>7.11 (.16)</td>
</tr>
<tr>
<td></td>
<td>LIST 2</td>
<td>6.22 (.12)</td>
</tr>
<tr>
<td>CASE</td>
<td>accusative</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIST 1</td>
<td>3.42 (.12)</td>
</tr>
<tr>
<td></td>
<td>LIST 2</td>
<td>3.76 (.14)</td>
</tr>
<tr>
<td>COMPLEXITY</td>
<td>center-embedded</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIST 1</td>
<td>2.39 (.13)</td>
</tr>
<tr>
<td></td>
<td>LIST 2</td>
<td>2.70 (.10)</td>
</tr>
</tbody>
</table>

Table 3: Mean acceptability ratings by dataset and condition in Experiment 1. Values in parentheses indicate standard errors.

Figure 4: Acceptability scores by list position with smoothed loess curve for island materials on List 1 in Experiment 2. Shaded area shows 95% confidence interval.
a high probability to number mismatches in this context, and the linguistic input received in the experiment does not alter this expectation.

Beyond showing the utility of testing at high levels of repetition, the results here augment those of Experiment 1. They demonstrate that at least two types of clear cut ungrammaticalities are sensitive to repeated exposure. They demonstrate that judgments rise for sentences that are uncontroversially unacceptable for reasons due to processing complexity, whereas relatively simpler sentences do not exhibit this behavior. They demonstrate that ungrammaticality per se is not sufficient to observe repeated exposure effects, even at relatively high levels of repetition. Unlike Experiment 1, where repeated exposure effects applied to nearly every condition that was tested, here we see that judgments for multiple conditions remain quite stable. This configuration of results undercuts any concern that the rise in ratings stems from a task-related effect. If general task demands or features drive the observed effects, we would predict all conditions to pattern the same, which is clearly not the case here.

I would briefly note that the data from Experiment 1 and 2 neither speak in favor nor against processing-based accounts of island effects. Plausibly, repetition may make the processing of complex dependencies easier, which accords with the fact that judgments for sentences with nested dependencies increased sharply with repeated exposure. In this sense, the evidence is certainly consistent with accounts of island effects centered on cognitive limitations. However, the evidence clearly indicates that judgments for sentences characterized by ungrammaticalities also rise with repetition. So, there is little reason to think the collected evidence distinguishes island violations from other sources of ungrammaticality.

4 Experiment 3

In both experiments, a possible explanation for the rise in acceptability judgments relates to processing facilitation, at least for island-violating conditions and those with nested dependencies. That is, if island effects arise because of cognitive constraints, then one perspective on the effects described so far is that respondents gradually improve in their ability to parse such constructions. It may even be that sentences with case and inflectional errors temporarily increase difficulty to the extent that they hinder assignment of thematic roles and understanding of temporal relations. Conceivably, encountering ungrammaticalities within
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Factor</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Islands</td>
<td>Dependency[dependency=1]</td>
<td>-0.326</td>
<td>0.051</td>
<td>-6.44</td>
</tr>
<tr>
<td></td>
<td>Item Order</td>
<td>0.024</td>
<td>0.007</td>
<td>3.71</td>
</tr>
<tr>
<td></td>
<td>Depression[dependency=1] × Item Order</td>
<td>0.013</td>
<td>0.007</td>
<td>2.05</td>
</tr>
<tr>
<td>Case</td>
<td>Grammaticity[nom.=1]</td>
<td>-0.725</td>
<td>0.048</td>
<td>-14.97</td>
</tr>
<tr>
<td></td>
<td>Item Order</td>
<td>0.011</td>
<td>0.006</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>Grammaticity[nom.=1] × Item Order</td>
<td>0.015</td>
<td>0.007</td>
<td>2.17</td>
</tr>
<tr>
<td>S-V agree</td>
<td>Grammaticity[mismatch=1]</td>
<td>-0.228</td>
<td>0.068</td>
<td>-3.36</td>
</tr>
<tr>
<td></td>
<td>Item Order</td>
<td>-0.005</td>
<td>0.023</td>
<td>-0.23</td>
</tr>
<tr>
<td></td>
<td>Grammaticity[mismatch=1] × Item Order</td>
<td>-0.005</td>
<td>0.016</td>
<td>-0.30</td>
</tr>
<tr>
<td>complexity</td>
<td>Complexity[right-branch=1]</td>
<td>0.568</td>
<td>0.048</td>
<td>11.81</td>
</tr>
<tr>
<td></td>
<td>Item Order</td>
<td>0.046</td>
<td>0.020</td>
<td>2.35</td>
</tr>
<tr>
<td></td>
<td>Complexity[right-branch=1] × Item Order</td>
<td>-0.051</td>
<td>0.015</td>
<td>-3.53</td>
</tr>
</tbody>
</table>

Table 4: Model summary for each Experiment 2 dataset’s fixed effects for List 1. Each model contains a predictor for the primary manipulation, item order, and their interaction. Bracketed material indicates the treatment level of the experimental manipulation whose corresponding statistics are shown. T-values above the absolute value of 2 are considered statistically reliable.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Factor</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Islands</td>
<td>Dependency[dependency=1]</td>
<td>-0.335</td>
<td>0.062</td>
<td>-5.39</td>
</tr>
<tr>
<td></td>
<td>Item Order</td>
<td>0.037</td>
<td>0.028</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>Depression[dependency=1] × Item Order</td>
<td>0.014</td>
<td>0.018</td>
<td>0.76</td>
</tr>
<tr>
<td>Case</td>
<td>Grammaticity[nom.=1]</td>
<td>-0.679</td>
<td>0.061</td>
<td>-11.16</td>
</tr>
<tr>
<td></td>
<td>Item Order</td>
<td>0.057</td>
<td>0.022</td>
<td>2.59</td>
</tr>
<tr>
<td></td>
<td>Grammaticity[nom.=1] × Item Order</td>
<td>0.014</td>
<td>0.024</td>
<td>0.57</td>
</tr>
<tr>
<td>S-V agree</td>
<td>Grammaticity[mismatch=1]</td>
<td>-0.370</td>
<td>0.056</td>
<td>-6.65</td>
</tr>
<tr>
<td></td>
<td>Item Order</td>
<td>0.008</td>
<td>0.007</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>Grammaticity[mismatch=1] × Item Order</td>
<td>-0.003</td>
<td>0.007</td>
<td>-0.50</td>
</tr>
<tr>
<td>complexity</td>
<td>Complexity[right-branch=1]</td>
<td>0.506</td>
<td>0.041</td>
<td>12.48</td>
</tr>
<tr>
<td></td>
<td>Item Order</td>
<td>0.031</td>
<td>0.008</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>Complexity[right-branch=1] × Item Order</td>
<td>-0.019</td>
<td>0.007</td>
<td>-2.94</td>
</tr>
</tbody>
</table>

Table 5: Model summary for each Experiment 2 dataset’s fixed effects for List 1. Each model contains a predictor for the primary manipulation, item order, and their interaction. Bracketed material indicates the treatment level of the experimental manipulation whose corresponding statistics are shown. T-values above the absolute value of 2 are considered statistically reliable.
a sentence systematically yields increased processing difficulty. The reader or listener may try and repair the input or else exert additional effort to extract a meaningful signal from the aberrant input. Upward shifts in judgments may thus be restricted to conditions for which processing effort can be reduced via experience.

Since ungrammaticality may engender processing difficulty, this would seem to create a conundrum where we cannot say for certain whether judgment shifts are tied to (changes in) processing difficulty. Luckily, there is a recently documented case where acceptability judgments conflict with processing measures. That is, what is easier to process is judged to be worse than what is harder to process. The relevant data involve resumptive pronouns in English.

A variety of sources point out that a resumptive pronoun such as *it*, in place of a gap, intuitively seems to improve the acceptability of English island-violating constructions, e.g., *Which bill was Obama unsure whether Democrats would be able to pass it through Congress?* (Ross, 1967, Kroch, 1981, Sells, 1984, Prince, 1990). One proposal for this salvaging effect of resumptives is that they act as a compensatory mechanism for comprehension or production difficulties (Givón, 1973, Hawkins, 1994, Ariel, 1999), offsetting some of the costs of processing an island-violating dependency. At the same time, resumptives appear to be extremely unacceptable in relatively simple dependency contexts, e.g., (8a), but less so with deeper embedding, (8c) (Alexopoulou and Keller, 2007):

(8)  

a. Which prisoner did the guard help him?  
b. Which prisoner did you say that the guard helped him?  
c. Which prisoner did you say that it was well-known that the guard helped him?

These observations accord with proposals that resumptive pronouns compensate for some of the costs of online dependency processing.

Recently, a variety of experimental work has challenged some of these earlier intuitions. In brief, acceptability data show that resumptive pronouns do not give rise to higher judgments than gaps in some island environments and deeply embedded contexts (Alexopoulou and Keller, 2007, Heestand et al., 2011). These results have quickly gained notoriety for good reason, but there is also a danger of mischaracterizing or oversimplifying a more nuanced landscape. First, resumptives in some island contexts are rated more highly than gaps (Keffala, 2013). This variation appears to connect to the grammatical position of the resumptive pronoun: ‘rescuing’ affects emerge in English only when the pronoun appears in subject position, according to Keffala (2013). Two, the negative effect of resumptive pronouns on acceptability diminishes with embedding (Alexopoulou and Keller, 2007, Hofmeister and Norcliffe, 2014). While this is usually reported, its significance is usually overshadowed by the lack of true rescuing. Three, self-paced reading data indicate that resumptive pronouns do indeed facilitate processing in English (Hofmeister and Norcliffe, 2014). This contradicts conclusions made purely on the basis of acceptability judgments that ‘resumption does not help the hearer’ (Heestand et al., 2011). Overall, while resumptive pronouns affect judgments in a way that calls for refinement to prior intuitions, the data consistently affirm that the processing difficulty of the context in which they appear affects overall sentence judgments.

4.1 Design

The data and materials for Experiment 3 come from Hofmeister and Norcliffe (2014), who tested the acceptability of resumptive pronouns vs. gaps in both easy (short dependencies) and hard (long dependencies) context, although we did not investigate list position in that study. 24 items appeared in four conditions, as shown below:

(9)  

a. Mary confirmed that there was a prisoner who the prison officials had acknowledged that the guard helped ___ to make a daring escape.  
b. Mary confirmed that there was a prisoner who the prison officials had acknowledged that the guard helped him to make a daring escape.  
c. The prison officials had acknowledged that there was a prisoner that the guard helped ___ to make a daring escape.  
d. The prison officials had acknowledged that there was a prisoner that the guard
helped him to make a daring escape.

Unlike in Experiments 1 and 2, sentences were presented one word at a time for a fixed period (250 ms + 16.67ms * the number of characters). In addition, instead of using a Likert scale for rating stimuli, judgments were made on an open-ended scale, using the Thermometer Judgment methodology. In this method, participants judge items relative to two reference sentences, one of which is quite good and the other quite bad.

Replicating results from Heestand et al. (2011) and Alexopoulou and Keller (2007), resumptive pronouns consistently yielded lower judgments, regardless of context. However, the difference between the resumptive and gapped conditions was smaller in the context of a long dependency (9a vs. 9b), compared to a short one (9c vs. 9d). Given that self-paced reading data established that processing times were fastest immediately after the gap/resumptive in examples like (9b), we surmised that the interaction in the acceptability data reflected the processing benefits of the resumptive in the long dependency context.

4.2 Analysis

To assess the effects of repeated exposure, the acceptability data from Hofmeister & Norcliffe (2014) were reanalyzed. For this experiment, though, the statistical analysis was carried out using Bayesian hierarchical mixed effects models via the rstan package (version 2.6.0) in R (version 3.1.2). The reason for this analytical change is that maximal random effect models failed to converge in R lme4, as well as any random effect structures that included item order. This failure to converge likely reflects the relatively small number of data points compared to the number of parameters in a random effect structure containing a three-way interaction for participants and items.

Bayesian modeling, however, appears to be more robust to smaller datasets. A hierarchical mixed effects model was constructed with the fixed effect factors of dependency length, resumption, and item order and all interaction terms. The random effect structure included random intercepts for participants and items, as well as by-participant and by-item random slopes for all fixed effect terms. The model used 5 chains, 2,000 samples per chain, a warm-up of 1,000 samples, and no thinning. Statistical inference was based on computing the posterior probabilities $P(\beta < 0)$ for acceptability decreases and $P(\beta > 0)$ for acceptability increases. These probabilities were obtained by calculating the percentage of posterior samples above or below zero. Note that, unlike frequentist methods of statistical inference, Bayesian methods yield parameter estimates with probabilities above and below zero. A probability of .96 thus entails that 96% of the posterior samples were greater than 0. For all parameter estimates, I also report the 95% Credible Interval that defines the posterior interval where we can assume that the parameter value lies with 95% certainty.

4.3 Results & Discussion

As Figure 6 illustrates, resumptive pronouns in both short and long dependency are deemed to be increasingly better with exposure. This is further confirmed by a separate model on just these two conditions, revealing a significant effect of item order ($\hat{\mu} = 0.061$, CrI Lower = 0.010, CrI Upper = 0.114, $P(\beta < 0) = 0.012$). In contrast, gapped conditions do not exhibit an overall effect of item order ($\hat{\mu} = -0.005$, CrI Lower = −0.063, CrI Upper = 0.055, $P(\beta < 0) = 0.576$). Statistically, these two generalizations yield an interaction of resumption and item order in the main analysis, shown in Table 6. However, this two-way interaction is qualified by a three-way interaction of dependency length, resumption, and item order: judgments rise with repeated exposure when the dependency ends with a resumptive pronoun, but when it ends with a gap, judgments only rise when the dependency is long, as depicted in Figure 6. In other words, the lack of an item order effect in the gapped conditions belies the fact that judgments are increasing for the complex, long condition, while slightly decreasing for the short, gapped condition (more precisely, there is a 92% probability the effect of item order on the latter condition is negative, based on a separate model fit).

What’s most notable here is that ratings appear to
Figure 6: Acceptability scores by (jittered) list position with smoothed loess curve for Experiment 3. Shaded area shows 95% confidence interval.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Mean</th>
<th>CrI lower</th>
<th>CrI upper</th>
<th>P(β &lt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length[Long=1]</td>
<td>−0.184</td>
<td>−0.251</td>
<td>−0.116</td>
<td>1</td>
</tr>
<tr>
<td>Resumption[Resumptive=1]</td>
<td>−0.293</td>
<td>−0.395</td>
<td>−0.199</td>
<td>1</td>
</tr>
<tr>
<td>Length × Resumption[Resumptive=1]</td>
<td>0.125</td>
<td>0.065</td>
<td>0.180</td>
<td>0</td>
</tr>
<tr>
<td>Item Order</td>
<td>0.025</td>
<td>−0.023</td>
<td>0.070</td>
<td>.137</td>
</tr>
<tr>
<td>Length[Long=1] × Item Order</td>
<td>0.014</td>
<td>−0.016</td>
<td>0.043</td>
<td>.168</td>
</tr>
<tr>
<td>Resumption[Resumptive=1] × Item Order</td>
<td>0.034</td>
<td>0.004</td>
<td>0.064</td>
<td>.013</td>
</tr>
<tr>
<td>Length[Long=1] × Resumption[Resumptive=1] × Item Order</td>
<td>−0.027</td>
<td>−0.058</td>
<td>0.006</td>
<td>.954</td>
</tr>
</tbody>
</table>

Table 6: Model summary for Experiment 1 for each fixed effect factor. Summary includes the posterior 95% Credible Interval (CrI), i.e. the lower CrI refers to the 2.5% bound and the upper CrI refers to the 97.5% bound. P(β < 0) indicates the probability that the parameter estimate is below zero, i.e. the factor in question lowers judgments.
improve in three out of the four conditions despite differing sentence processing profiles and differing grammatical statuses. Looking at the case of the long dependency with a gap at the tail, the prevailing presumption would be that it is less acceptable than a short dependency with a gap as a result of greater processing costs. In the conditions with resumptives, however, there is no evidence to suggest that these are harder to process than their gapped variants. For the ‘long’ dependency context, in fact, there is strong evidence that processing is facilitated by the resumptive pronoun. Thus, we have a collection of unbounded dependency constructions whose relative unacceptability reflects varied forces on judgments. Yet, all of them sound better with exposure to similar constructions. Meanwhile, the ‘best’ condition — a short dependency ending in a gap — receives lower judgments with exposure. As these effects were documented under experimental conditions that differed from those in the first two experiments (different presentation method and different scale), the observation of judgment shifts does not appear to be tied to any particular task.

5 General Discussion

Syntactic structures that appear unnatural gradually becomes less so with exposure to similar structures. According to the evidence here, this generalization applies to a range of syntactic formulations deemed to be unacceptable, compared to minimally different variants. To the best of my knowledge, this is the first data showing that systematic judgment shifts are observable beyond the domain of island-violating sentences, while simultaneously adding to the literature which has found the so-called satiation effects for islands. Most strikingly, these effects apply to constructions that have starkly different analyses in the theoretical literature: the unacceptability of key sentence types examined here has been previously attributed to ungrammaticality, processing complexity, or both. Yet in all such cases, judgment ratings tick upward with repeated exposure to the offending structures. For the relatively acceptable conditions, judgments either remained stable across the experiment or else enjoyed far weaker upward shifts. Anecdotally, examining my data from previously published acceptability studies revealed that repeated exposure effects occur across an even wider range of constructions, even with different presentation methods.

So why do such repeated exposure effects occur in some cases, but not others, and what accounts for the differences in the rate of change? What is certain is that this phenomenon does not apply to only ‘grammatical’ sentences — at least in the categorical sense that dominates the syntactic literature. This much is apparent simply via the judgments for sentences with case violations and inflectional errors, if not the resumptive examples. Even jumbled English sentences, as in (10) & (11), improve with exposure to similarly jumbled sentences, as documented by Hofmeister et al. (2013b):

(10) Iran has gun-control strict laws that bar private citizens carrying from firearms.

(11) Some of the were protesters telling the police that they have the right legal to protest in peace.

These results complement those of Luka and Barsalou (2005) who observed “structural persistence” effects for sentences like *Sam recites poems as well as playing the piano*. Grammaticality in the standard sense thus does not seem to be a precondition for the observation of repeated exposure effects, contra the suggestion of Sprouse (2009).

The data also effectively rule out the possibility that repeated exposure effects arise only for complex or costly constructions. Island-violating structures and resumptive pronouns have juxtaposed ties to processing difficulty — for islands, there is evidence that unacceptability relates to processing difficulty, whereas in the other case, the unacceptability arises in spite of processing facilitation. Yet judgments for both rise with exposure. To be clear, repeated exposure effects in judgment tasks may often reflect a kind of facilitation process (e.g., via priming), but the evidence does not show that upward judgment shifts only happen with hard-to-process or syntactically complex sentences. So, processing difficulty may be a factor in the magnitude of exposure-related effects, but it is not the sole determinant.
Moreover, the issue of equalization raised by Sprouse (2009) does not seem a likely factor here. First, Sprouse was responding to an imbalance of responses in a yes/no judgment task, whereas the current studies all utilize Likert scales. In Likert scale tasks, the distribution of responses almost always tilts to the right — participants use the high end of the scale more than the low end. Second, in none of the experiments described here were there more ‘bad’ sentences than ‘good’ ones.⁶

Some further generalities emerge from the data. First, the rate of upward shift appears to depend on how the offending structure is originally rated. This means that judgments for sentences that are relatively acceptable, compared to some minimally different variant, will usually not shift upwards simply because it is likely that the target sentences receive generally high ratings on whatever scale is used. However, in some cases, two variants may both receive low average ratings on some scale, leaving room for even the preferred variant to shift. This situation arose with the complexity dataset from Experiment 1. The second generality is that relative unacceptability appears to be insufficient to observe repeated exposure effects: in the case of agreement errors, judgments do not shift for such sentences at either 6 or 12 repetitions. The lack of any effects here was tied to the first observation — the relevant structures were already judged to be relatively acceptable.

One viable perspective on these effects is that respondents alter their judgments to meet the statistical properties of the input relative to past experience, despite no such instructions being given. That is, the mere act of encountering sentences with abnormalities — induced by grammatical or processing considerations — may cause these sentence types to be less unnatural. This view aligns with the overwhelming evidence that expectations affect various levels of language processing, from the lexical items (Kutas and Hillyard, 1980, 1984, Kutas and Federmeier, 2011, Altman and Kamide, 1999, DeLong et al., 2005, Arai and Keller, 2013) to lexical categories (Tabor et al., 1997, Gibson, 2006, Levy and Keller, 2013) to syntactic structures (Fine et al., 2010, 2013, Lau et al., 2006, Levy, 2008, Levy et al., 2012, Staub and Clifton, 2006, Staub et al., 2006). In short, there is little doubt that expectations impact the online sentence processing profile. If this profile partly determines judgment patterns, as is commonly assumed, it should come as no surprise that sentences judgments are at least somewhat sensitive to the likelihood of syntactic structures.

Notably, comprehenders alter their online processing profile as a function of recent input, according to recent evidence (Fine et al., 2010, Kleinschmidt et al., 2012, Fine et al., 2013, Wells et al., 2009). As one instance, Fine et al. (2013) report that garden-path sentences like (12b) become progressively easier to process with exposure to such reduced relative clause sentences. More significantly, the reading time difference at the key disambiguating region (in bold) between locally ambiguous sentences like (12b) and unambiguous ones like (12d) decreases:

(12) a. The experienced soldiers warned about the dangers before the midnight raid.
   b. The experienced soldiers warned about the dangers conducted the midnight raid.
   c. The experienced soldiers spoke about the dangers before the midnight raid.
   d. The experienced soldiers who were told about the dangers conducted the midnight raid.

In fact, with enough exposure to sentences like (12b), reduced relative clause sentences can be processed faster than main verb sentences like (12a). To explain such effects, Fine et al. (2013) contend that comprehenders habitually “converge towards, or even on, the statistics of the current linguistic environment . . . By generating expectations that reflect the distribution of actual events in the environment, comprehenders should, in principle, be able to reduce the average prediction error experienced during online processing, and thus process language efficiently.” That is, tuning one’s expectations to the distribution of syntactic forms in a given environment paves the way for more efficient and successful parsing procedures with the objective of extracting coherent meaning from the

⁶Note, though, that to categorize sentences as good or bad before doing an acceptability study begs the question of what the point of the study is.
linguistic signal (Fine and Jaeger, 2013, Jaeger and Snider, 2013).

Assuming analogous effects in judgment data, the repetition of an unlikely sentence structure should shift the relevant probability distribution over possible structural realizations and hence any judgments for said structure. This would mean that respondents unconsciously learn how language is used in a particular context and update their judgments to reflect the learned distributions. This perspective should not be surprising to anyone who has encountered a new dialect of their language and learned to accept or even use a previously unconscionable syntactic formulation. Conceivably, respondents in judgment tasks are doing something similar: even though a particular formulation strikes them as unnatural, its occurrence changes their perspective on what is likely or expected.

The notion that judgments at least partly reflect experience and, derivatively, expectations is not a novel idea, either. In a study that directly informs the discussion here, Chaves (2013) argues that English users develop expectations over unbounded dependency paths. In particular, the surface distribution of such paths in English leads to expectations for gaps to be located in a subject’s sister VP, i.e., after the main verb. By extension, the subject NP itself is expected to be gapless. Thus, an unbounded dependency as in (13a) would satisfy both expectations, a parasitic gap construction like (13c) would meet the first expectation but violate the second, whereas an example like (13b) violates two expectations:

(13) a. Who did the rivals of Castro shoot?
    b. Who did the rivals of shoot Castro?
    c. Who did the rivals of shoot?

In light of the cognitive costs of processing complex subjects and pragmatic constraints on questioning, as well as the attestation of fairly natural examples, Chaves reasoned that dependencies into complex subjects were unexpected in English, but not impossible. Because they reflect gradient expectations rather than categorical rules, perceptions of such dependencies should be alterable. This argument was corroborated by data from Chaves and Dery (2013), showing that repeated exposure to sentences with dependencies into subjects progressively led to higher judgments of linguistic acceptability for such structures.

What’s changing with repeated exposure may thus reflect changes in processing costs as mediated by expectations. It would be entirely compatible with the data seen here, however, to suppose that the grammar itself is changing (independently of processing costs), especially if some probabilistic grammar is entertained. After all, grammars in general can be construed as expectations over phrase structural possibilities; in standard theories, these expectations are simply categorical. However, in general, syntactic theories must acknowledge the fact that languages change at various levels of linguistic granularity, e.g., phonological, lexical, syntactic, etc. So, one view on the longitudinal changes observed here are that they mirror natural linguistic processes of adaptation to changing input and thus constitute a small-scale replica of diachronic grammatical shifts. The point is that these judgment shifts can be apprehended through the lens of a changing grammar, changing cognitive costs, or, in light of their interdependence, a combination. On any scenario, expectations driven by experience assume a much larger role in the understanding of linguistic acceptability.

All this requires the assumption that what participants are doing in judgment tasks captures some significant aspect of how language is perceived and used in everyday communication. This non-trivial issue is not unique to judgment tasks — all methods face scrutiny over whether they capture aspects of human cognition and behavior outside the laboratory. To the extent that one finds judgment data generally informative about grammatical representations and the mental states associated with them, then the data here have significant repercussions for how one views the stability of grammar and the relationship of grammar to the input. While more evidence is needed to clarify the role of expectations in these judgments shifts, an enticing aspect of this experience-driven view of judgments is that it brings the science of linguistic acceptability judgments closer to successful research programs in the language sciences that center around the role of experience in learning and language use (MacDonald et al., 1994, MacDonald and Christiansen, 2002, Misyak et al., 2010, Safran et al.,

19
1996, Stanovich and West, 1989, Wells et al., 2009, Jaeger and Snider, 2013). If experience is a primary conduit through which individuals learn and use language, then how people perceive and judge sentences of their language likely shares a similar path.

References


