The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments

Running header: The source ambiguity problem

Philip Hofmeister

Center for Research in Language

University of California-San Diego

La Jolla, CA

Telephone: (650)-387-6641

E-mail: phofmeister@ucsd.edu

T. Florian Jaeger

University of Rochester

Inbal Arnon

University of Manchester, UK

Ivan A. Sag

Stanford University

Neal Snider

Nuance Communications, Inc.

Abstract

Judgments of linguistic unacceptability may theoretically arise from either grammatical deviance or significant processing difficulty. Acceptability data are thus naturally ambiguous in theories that explicitly distinguish formal and functional constraints. Here, we consider this source ambiguity problem in the context of Superiority effects: the dispreference for ordering a *wh*-phrase in front of a syntactically "superior" *wh*-phrase in multiple *w*h-questions, e.g. *What did who buy?* More specifically, we consider the acceptability contrast between such examples and so-called D-linked examples, e.g. *Which toys did which parents buy?* Evidence from acceptability and self-paced reading experiments demonstrates that (i) judgments and processing times for Superiority violations vary in parallel, as determined by the kind of *wh*-phrases they contain, (ii) judgments increase with exposure while processing times decrease, (iii) reading times are highly predictive of acceptability judgments for the same items, and (iv) the effects of the complexity of the *wh*-phrases combine in both acceptability judgments and reading times. This evidence supports the conclusion that D-linking effects are likely reducible to independently motivated cognitive mechanisms whose effects emerge in a wide range of sentence contexts. This in turn suggests that Superiority effects, in general, may owe their character to differential processing difficulty.*

*For helpful discussion and input on this research, we thank Bruno Estigarribia, Evelina Fedorenko, Ted Gibson, Jeanette Pettibone, Laura Staum Casasanto, and Tom Wasow. Very special thanks go to Ted Gibson and Ev Fedorenko for running the participants from Experiment III at MIT's Tedlab, and to David Kettler for his help in conducting Experiment IV. For help in editing the manuscript, we thank Camber Hansen-Karr. This first author gratefully acknowledges research support from NIH Training Grant T32-DC000041 via the Center for Research in Language at UC-San Diego.

1 Acceptability Judgments

Linguists rely on the ability to distinguish grammatical sentences from their ungrammatical counterparts. The problem for the language theorist, however, is that judgments of unacceptability may derive from different sources. An example may be judged unacceptable because it violates grammatical constraints or because it is simply hard to process, interpret, or contextualize. For instance, Miller and Chomsky (1963) point out that sentences with multiple center-embeddings, as in (1a), are nearly incomprehensible and sound highly deviant to most speakers of English:

(1) a. The boy the girl the host knew brought left.

b. The boy the girl brought left.

The structure in (1a) can be derived by repeated application of the grammatical rule that licenses the relatively acceptable-sounding variant in (1b), e.g. $NP \rightarrow NP\ S$. Hence, they suggest that the unacceptability of (1a) should not be attributed to a grammatical constraint, but rather to independently motivated cognitive difficulty that arises when processing such structures. In other cases of linguistic unacceptability, however, the source of the deviance is less clear. We refer to this as the "source ambiguity problem," and it is this ambiguity that makes it challenging to understand what effects grammar imposes, and what effects should be explained in terms of independently motivated cognitive processes.

An example of this problem emerges in the discussion of multiple *wh*-questions, as in (2):

(2) Who finished what?

Such questions commonly elicit pair-list readings (Dayal, 2002; Grohmann, 2003), where an appropriate answer consists of multiple propositions or argument pairs, e.g. *John finished the stats*, *Hank finished the formatting*, etc., or single-pair readings where a single such proposition or argument pair suffices (see Ginzburg and Sag (2000) for a discussion of reprise uses).

As first noted by Kuno & Robinson (1972), examples like (3) are a considerably less natural way to ask the question posed by (2):

(3) What did who finish?

Chomsky (1973) states that this condition on *wh*-ordering actually represents an instance of a broader limitation on constituent movement, termed *Superiority*:

(4) No rule can involve X, Y in the structure:

$$\ldots X \ldots [_{\alpha} \ldots Z \ldots - WYV \ldots]$$

where the rule applies ambiguously to Z and Y and Z is superior to [m-commands] Y.

Subsequent renditions of this constraint, e.g. Attract Closest or the Minimal Link Condition (Chomsky, 1995, 2000), retain much the same spirit. In all their renditions, grammatical accounts of Superiority invoke comparisons that depend on assessments of relative structural positions. Our conclusions apply equally to any such formulation; however, we will talk uniformly of the "Superiority constraint," and will refer to examples like (3) as "Superiority violations" (SUVs).

Even within the category of SUVs, acceptability contrasts have been ascribed to differences in grammaticality. Since Karttunen (1977), it has been reported that using *which*-$\overline{\text{N}}$ phrases, as in (5), rather than bare *wh*-words like *who* and *what*, produces relatively acceptable structures (see also Wilkins (1977, 1980), Maling and Zaenen (1982), Culicover and Wilkins (1984) for similar observations):

(5) Which medicine does which patient need?

Even sentences with two bare *wh*-elements are acceptable in certain circumstances, as in (6) (see Bolinger (1978)):

(6) You look awful! What did who DO to you?

Superiority-violating structures with only bare *wh*-words also appear on the internet (Arnon, Hofmeister, Jaeger, Sag, & Snider, 2005; Clifton, Fanselow, & Frazier, 2006). Although ungrammatical examples are sometimes attested in speech or text (see Labov (1973, 1996) for discussion), the fact that these examples appear in presumably edited text speaks against their status as errors:

(7) a. What did who know and when did they know it?
[http://$www.antigonishreview.com$/bi-113/113-curb.html]

b. What did who say and who did the asserting?
[http://$www.thenation.com/doc/20030512/cockburn$]

c. What, do you think this is a game? What rules should who follow? This shitsandwich is a reality – a competition for survival between all souls...

[http://$www.cruel.com/discuss/viewTopic.php/83308$]

Therefore, it appears that *wh*-orders like those shown in (7) are acceptable under at least some circumstances.

To preserve the generalization behind the Superiority constraint, the relative acceptability of sentences like (5) calls for some refinement or addendum. Such an addendum appears with the concept of D(iscourse)-linking, presented in Pesetsky (1987, 2000) (see also Comorovski (1989)). Although a precise explanation of D-linking does not appear in Pesetsky (1987, 2000) or in the secondary literature, the following serves as a rough approximation of its intended meaning:

> "Context sets previously mentioned in the discourse qualify a phrase as D-linked, but so do sets that are merely salient (e.g. *which book*, in a context where speaker and hearer both know that reference is being made to a reading list for a course) and sets whose salience is culturally determined (e.g. *what day of the week, which sign of the zodiac*)" [p. 16].

This provides an escape hatch for examples like (5)-(7): Superiority now only applies to *wh*-phrases that are non-D-linked. Consequently, the grammar retains the relevant constraints on *wh*-phrase order and examples like (3) retain their grammatically ill-formed status. Additionally, individual lexical items and phrases are not inherently specified for D-linking. Instead, this property emerges from the context or conventionalized implicatures. Hence, nothing theoretically prevents bare *wh*-items like *who* and *what* from being D-linked.

However, within the theory of grammar assumed by Pesetsky, it is hard to see why contextualization invalidates an otherwise universal constraint such as Superiority. Pesetsky (2000) acknowledges the absence of any rationale for why D-linking voids an exception to the constraints on *wh*-phrase order: "The reason for this link between semantics and syntax is obscure, and will remain obscure even at the end of this book" (p. 16). Since there is no explanation for why contextualization voids this one syntactic constraint, D-linking provides only a descriptive label for examples which run counter to the predictions of Superiority. Moreover, to the extent that contextualization or the saliency of context sets accounts for acceptability differences, this seems to belong to the class of processing-related factors, rather than potential grammatical factors. That is, the accessibility of discourse referents and the role of contextualization have long been assumed to affect processing ease (Ariel, 1990, 2001; Prince, 1978, inter alia) but do not have a historical precedent for affecting grammaticality.

This leaves us with the following questions: Are Superiority-violating examples perceived to be worse than their Superiority-conforming counterparts because English (or some universal) grammar states restric-

tions blocking structures like the one in (3), or do processing differences give rise to the acceptability contrast between items like (2) and (3), or some combination of the two? Are non-D-linked SUVs worse than D-linked examples due to some grammatical restrictions, or is the difference reducible to processing? As with SUVs more generally, decades of syntactic research reflect the sustained belief that some grammatical constraint separates D-linked and non-D-linked SUVs.

A competing line of thought suggests that differences in processing difficulty play an important role in the differences between SUVs vs. non-SUVs and D-linked vs. non-D-linked SUVs (Arnon et al., 2005; Culicover, 2008; Hofmeister, Jaeger, Sag, Arnon, & Snider, 2007; Jaeger, 2004). In particular, this hypothesis argues that Superiority-violating multiple *wh*-questions like (3) pose significant processing difficulties that are absent in cases like (5), leading to differences in acceptability judgments. This builds on previous work arguing that processing difficulty accumulates during sentence processing in such a way that several sources of difficulty may combine to produce additive or even super-additive effects (Gibson, 1990; Kluender, 1991, 1998; Kluender & Kutas, 1993b).

In this paper, we ask whether the acceptability difference between SUVs with and without *which*-$\overline{\text{N}}$ phrases can be reduced to differences in processing difficulty. That is, we aim to determine whether observed differences *among various kinds of SUVs* can be accounted for in terms of processing principles alone. We provide evidence that the acceptability contrasts previously attributed to D-linking can be derived from differential processing difficulty due to independently motivated cognitive mechanisms affecting comprehension. In short, D-linking effects are predictable from the processing perspective, such that no special exception principle is necessary. This leads us to entertain the possibility that the perceived unacceptability of SUVs in general relates to the differential processing difficulty associated with some of the same processing mechanisms. While our results suggest that online sentence processing constraints can account for a considerable amount of observed variation in acceptability judgments, they do not exclude the possibility that grammar also contributes to Superiority effects. Particularly, while we focus on grammatical accounts that make categorical predictions, the work presented here raises intriguing possibilities about the interplay between processing and grammar, as understood by grammatical accounts with gradient constraints. We discuss some of these implications in the final discussion section.

2 Processing *wh*-dependencies

In this section, we review psycholinguistic accounts that make predictions about the processing difficulty of different multiple *wh*-phrase orderings based on properties of the *wh*-phrases. This includes predictions about cases previously attributed to D-linking, e.g. the preference for (8b) over (8a).

(8) a. What does who need?

b. Which medicine does which patient need?

The accounts we discuss also make predictions about the differential processing difficulty of SUVs vs. non-SUVs. However, we postpone discussion of these until the final section of this paper. Here, we focus on the processing pressures whose effects on SUV acceptability ratings are associated with D-linking (for a recent, more comprehensive summary of work on syntactic processing and its relationship to typology, see Jaeger and Tily (in press)).

A key feature of all SUVs is that they contain two *wh*-dependencies. For sentences like (8a) and (8b), this means that representations corresponding to the subject and object *wh*-phrases must be integrated with the verb *need* for proper interpretation. Hence, representations of the dependent arguments (the syntactic constituents required by the verbal head) must be restored from memory at the verb and assigned the appropriate thematic roles (Gibson, 1998, 2000; Hawkins, 1999; Lewis & Vasishth, 2005; McElree, 2000; McElree, Foraker, & Dyer, 2003; Van Dyke & Lewis, 2003; Vasishth & Lewis, 2006). Empirical support that representations of dependent arguments are restored from memory at the appropriate subcategorizing head comes from various measures, including reading times, reaction times, speed-accuracy trade-off tasks, and electrophysiological measures (Kluender & Kutas, 1993a; Nicol & Swinney, 1989; Osterhout & Swinney, 1993; Pickering, 1993; Pickering & Barry, 1991; Stowe, 1986; Swinney, Ford, Bresnan, & Frauenfelder, 1988; Tanenhaus, Carlson, & Seidenberg, 1985).

An increasing amount of evidence from psycholinguistics suggests that the degree to which linguistic representations overlap at both the syntactic and semantic level plays a highly important role in this retrieval process during sentence processing (Anderson, Budiu, & Reder, 2001; Gordon, Hendrick, & Johnson, 2001; Gordon, Hendrick, & Levine, 2002; Gordon, Hendrick, Johnson, & Lee, 2006; Lewis, 1996, 1999; Lewis & Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006; Van Dyke, 2007; Van Dyke & McElree, 2006). Specifically, a retrieval target may share feature values with other locally available representations that are cued at the retrieval site, thereby inducing so-called similarity-based interference. For instance, Gordon

et al. (2001) show that processing clefted constructions like (9) is easier when the two NPs preceding the subcategorizing verb are of different types (proper name vs. definite description):

(9) It was John/the barber that the lawyer/Bill saw in the parking lot.

At the verb *saw*, two unintegrated NPs compete for argument slots. Given two NPs of the same type, the phrases overlap with one another in terms of syntactic and semantic features (e.g. definiteness and syntactic form). This overlap is reduced when the two phrases differ in nominal type, and is accompanied by facilitated processing at the verb. Additional work shows that similar effects occur when syntactic differences are controlled for, but semantic overlap remains (Gordon et al., 2002; Van Dyke, 2007).

With English SUVs, the comprehender must similarly process back-to-back *wh*-elements, and then retrieve them and determine thematic relations at the head verb. By virtue of being *wh*-phrases, these phrases always overlap in terms of their phrasal type, although the extent of the overlap can vary significantly. Thus, SUVs like (8) differ from uncontroversial unary *wh*-questions like *What did Sandy need?* in terms of the similarity of the verbal arguments. At the extreme end of overlap are SUVs like (10a) & (11), where the identical *wh*-phrases substantially increase the processing difficulty of the sentence:

(10) a. Who did who see?

b. What did who see?

c. Which movie did who see?

d. Which movie did which child see?

(11) I realized which student which student likes.

In (10b), the *wh*-elements no longer overlap in terms of animacy, intuitively making this example more acceptable than (10a). The *which*-$\overline{\text{N}}$ phrase in (10c) dissociates the two arguments further. The phrases differ not only in terms of animacy, but also in terms of syntactic realization and semantic features introduced by the noun. Two *which*-$\overline{\text{N}}$ phrases with different nouns, as in (10d), presents the maximally distinctive situation. While the *wh*-elements may have the same syntactic form, they can differ on a number of different semantic dimensions, including animacy, number, humanness, abstractness, imagability, associated features, etc. In contrast, while a *which*-$\overline{\text{N}}$ phrase and a bare *wh*-item have differing syntactic forms, the bare *wh*-item may have few if any distinguishing or unique features, especially if the *wh*-phrases agree in animacy.

The SUV situation is thus quite similar to the relative clause and cleft contexts investigated by Gordon and colleagues. Even though the second of the two critical NPs appears immediately before the subcategorizing verb in cleft constructions like (9), the effects of feature overlap are significant and replicable (see also Warren and Gibson (2005)). In addition, many documented interference contrasts occur despite the fact that the lexical head triggering retrieval is unlikely to directly target the shared features that presumably account for the interference. For instance, there is little reason to suspect that the verb in (9) deploys a set of retrieval cues referencing definiteness or proper name status.

One way of modeling such interference effects is to assume that retrieval cues include fuzzy records of the original encoding to be matched against memory candidates (see Criss and McClelland (2006); Ericsson and Kintsch (1995); Nairne (1990, 2001, 2006); Oberauer and Lewandowsky (2008)). To illustrate, Nairne (1990, 2001, 2006) suggests that the probability of retrieving a memory representation increases with the similarity or feature-overlap of the retrieval cues and target, and decreases with the similarity of the cues to other memory candidates. Because the retrieval cues include versions of the original encoding, the chances for successful retrieval improve as the similarity between targets and non-targets decreases. Against this backdrop, we assume that *which*-$\overline{\text{N}}$ phrases encode more features than bare *wh*-words. In addition to the features shared with bare *wh*-words (e.g. being *wh*-words), *which*-$\overline{\text{N}}$ phrases carry further semantic and syntactic features that would make retrieval easier. For the interested reader, a detailed illustration of how Nairne's model predicts differences in SUV processing on these assumptions can be found in Appendix A.

Multiple models of sentence comprehension also argue that boosts in trace activation follow from modifying or adding to some representation (Anderson et al., 2001; Lewis & Vasishth, 2005; Vasishth & Lewis, 2006). These modifying events trigger memory access which increases the activation of the relevant memory chunk. In turn, elevated activation levels cause representations to be more accessible at future points. The process of building a complex syntactic and semantic representation thus leads to elevated activation and interference-resistant memory representations. Building on this idea, we argued in Hofmeister et al. (2007) that processing a *which*-$\overline{\text{N}}$ phrase boosts the activation level compared to processing a bare *wh*-item, as the former requires additional syntactic and semantic processing.

The evidence therefore suggests that D-linking effects can be reconsidered in terms of general processing mechanisms, particularly similarity-based interference. This leads to the prediction, then, that *which*-$\overline{\text{N}}$ phrases should elicit more efficient processing compared to bare *wh*-items in SUVs. Assuming that processing effort does influence acceptability judgments, we expect these ameliorative processing effects asso-

ciated with complex *wh*-phrases to correspond with higher judgments of acceptability—that is, sentences with complex *wh*-phrases should be deemed to be more acceptable that minimally different ones with bare *wh*-words. Moreover, we anticipate that the ameliorative effects of complex *wh*-phrases should combine, given the discussion regarding asimilarity-based interference (i.e. two maximally distinct phrases should result in the easiest processing context and the highest judgments). At the moment, we do not claim to have a precise formula for how processing difficulty within a sentence corresponds to acceptability judgments. The question here, however, is whether grammatical constraints are necessary if the effects of *wh*-form on acceptability judgments can be linked to independently motivated processing constraints.

The following experiments, therefore, test multiple hypotheses: (i) *which*-$\overline{\text{N}}$ phrases lead to higher acceptability judgments of SUVs; (ii) these effects are not specific to *which*-$\overline{\text{N}}$ phrases, but to *wh*-phrases that are syntactically and semantically richer than bare *wh*-items; (iii) *which*-$\overline{\text{N}}$ phrases produce faster response times at the retrieval region in SUVs; (iv) how individual SUVs are processed is a significant predictor of how the same tokens are judged.

In these experiments, we take note of two diagnostics relevant to the understanding of acceptability judgments: repeated exposure effects and how multiple constraints combine to influence acceptability judgments. Prior research illustrates that repeated exposure to a given type of structure may raise judgments of subsequent tokens of a similar or identical structure (Braze, 2002; Francom, 2009; Kaschak & Glenberg, 2004). Luka and Barsalou (2005), for example, demonstrate that moderately grammatical sentences receive higher acceptability ratings after participants read structurally similar tokens or identical repetitions in a preceding task. Of concern, however, is whether all types of structure show effects of repeated exposure. By some accounts, only grammatical strings should show evidence of repeated exposure, as illicit strings lack a coherent representation (see Sprouse (2009) for discussion). Relatedly, several studies suggest that certain types of island violations do not get better with repetition (Sprouse, 2009), although evidence to the contrary also appears in the literature (Braze, 2002; Hofmeister & Sag, 2010; Snyder, 2000). Consequently, we evaluate here how judgments for SUVs change with exposure, leaving interpretation of the results until the discussion section.

Secondly, we consider how multiple sources of unacceptability combine. A *which*-$\overline{\text{N}}$ phrase in either syntactic position of an SUV arguably improves acceptability ratings, but the prior literature has recognized that two *which*-$\overline{\text{N}}$ phrases are better than one (Comorovski, 1989, 1996; Pesetsky, 2000). If interference effects have a significant impact on processing and judgments of SUVS, as we suggest, then it follows

that using maximally distinctive *wh*-phrases (i.e. two *which*-$\overline{\text{N}}$ phrases) should be the optimal condition for facilitating processing and raising judgments. Note, we are not suggesting that grammatical sources of unacceptability cannot combine with processing ones or that they cannot "stack up." Rather, we consider the question of how constraints combine from the perspective of processing constraints, viz. the form and content of each *wh*-phrase should bear on the overall acceptability and processing of the construction.

In the final section of this paper, we take stock of the results and consider the source of unacceptability in SUVs generally. We discuss some additional factors that likely make SUVs harder to process than non-SUVs and how the judgment differences emerge from existing modeling work on sentence processing.

3 Experiment I: *Wh*-Phrase Type

3.1 Method

In Experiment I, we seek to evaluate intuitions from the theoretical literature that *which*-$\overline{\text{N}}$ phrases, as opposed to bare *wh*-words, improve the acceptability of SUVs. This study utilizes the Magnitude Estimation (ME) method for eliciting judgments, where a given stimulus is evaluated with respect to a chosen reference (see Bard, Robertson, and Sorace (1996) and Keller (2000)). Scores for experimental items are divided by the reference score and then logged to normalize the data. The ME analyses we present are based on the z-score of these log-transformed standardized judgments. Participants for the ME studies were recruited via online discussion forums and e-mail lists.

We analyzed all data using linear mixed-effects (LME) models, which allow for a principled way of incorporating longitudinal effects and covariates into the analysis, and are also free from the assumptions of homogenous variance and sphericity that are inherent to ANOVA (Baayen, 2004; Pinheiro & Bates, 2000). Markov chain Monte Carlo (MCMC) sampling (n=10,000) is used to reliably estimate p-values for the fixed and random factors (see Baayen, Davidson, and Bates (2008)).

We centered all fixed effects predictors before computing higher order variables (interactions). In addition to the fixed factors of *wh*-complexity, the LME models for acceptability experiments also included a fixed effect factor for logged list position. We validated the inclusion of this factor in all experiments by comparing models with and without this factor. Following Baayen et al. (2008), we always used the maximum random effect structure justified by the data as assessed by model comparison (using the R function anova) (see also http://hlplab.wordpress.com/2009/05/14/random-effect-structure/ for implementation

details). Sometimes, this led to removing the random variable for items, as it accounted for too little variation in the results. In Experiment I, the maximum likelihood-fitted model included a random intercept for participants, as well as random slopes for each participant for the effects of subject *wh*-complexity and log list position with correlation parameters; however, p-values cannot currently be estimated for models with random correlation parameters, so we report the results from the simpler model with the same random slopes but without the correlation parameters, which does not alter the results' significance. We follow this practice for all remaining studies. Items were not included as a random factor, based on model comparisons that showed it to be a non-significant source of variation.

3.2 Materials & Participants

In 20 experimental items consisting of embedded *wh*-questions, the *wh*-object and the subject *wh*-phrases appeared either as bare *wh*-words (*who* or *what*) or *which*-$\overline{\text{N}}$ phrases (e.g. *which book*):

(12) a. Mary wondered what who read.

b. Mary wondered which book who read.

c. Mary wondered what which boy read.

d. Mary wondered which book which boy read.

If *which*-$\overline{\text{N}}$ phrases facilitate dependency processing in comparison to bare *wh*-words, and processing effort predicts acceptability judgments, then two *which*-$\overline{\text{N}}$ phrases should yield the highest judgments, while those with a single *which*-$\overline{\text{N}}$ phrase should result in intermediate judgments.

41 individuals participated in this experiment over the web and did not receive any compensation. The 20 experimental items were accompanied by 32 fillers which ranged from simple, well-formed sentences, e.g. *Who loves Jane?*, to highly aberrant, ill-formed sentences, e.g. *Built what which yesterday?*

3.3 Results

Both object *wh*-phrase and subject *wh*-phrase complexity emerge as significant predictors of acceptability (see Table 1). *Which*-$\overline{\text{N}}$ phrases in both positions produce higher judgments of acceptability ($ps < .001$) compared to bare *wh*-words, as shown in Figure 1. These two main effects are qualified by a significant interaction. Figure 1 shows that this interaction happens because the WHICH-BARE condition is not significantly different from the BARE-BARE condition. In other words, a *which*-$\overline{\text{N}}$ phrase does not increase acceptability

[TABLE 1 ABOUT HERE]

[FIGURE 1 ABOUT HERE]

ratings when the subject *wh*-phrase is a bare *wh*-word in this experiment. But the higher ratings for the WHICH-WHICH condition compared to the BARE-WHICH condition evidences that the complexity of the *wh*-object phrase does influence judgments.

The results also indicate that where the critical items appear in the experiment also significantly affects judgments. As shown in Table 1 and Figure 2, acceptability judgments increase significantly over the course of the experiment. Furthermore, according to log likelihood tests, the model with list position outperforms a model without such a factor ($\chi^2 = 80.83$, $df = 4$, $p < .0001$). In contrast, judgments for filler items were not significantly affected by list position ($t = 0.76$, $p = .45$).

3.4 Discussion

These results confirm intuitions regarding the role of *wh*-phrase complexity in judgments of SUVs: *which*-$\overline{\text{N}}$ phrases correspond with higher acceptability judgments than bare *wh*-words. Perhaps surprisingly, the data seem to indicate that the complexity of the intervening *wh*-subject matters more for perceptions of well-formedness than the complexity of the object *wh*-phrase: Complex *wh*-objects did not raise judgments significantly when followed by bare *wh*-words, despite a numerical trend in that direction (see Featherston (2005) for similar findings for German SUVs).

In addition, a clear preference for two *which*-$\overline{\text{N}}$ phrases over just one appears in the data. If *which*-$\overline{\text{N}}$ phrases lead to easier processing at the syntactic head than bare *wh*-items, the results are explicable on the analysis that the easier processing leads to higher judgments for each token. As suggested earlier, making each argument phrase syntactically or semantically richer (and thus more likely to have unique features) can offset some of the retrieval difficulties posed by the feature-values that the arguments share (i.e. both are *wh*-phrases). Moreover, the longitudinal effects on judgments parallel those observed elsewhere for sentences with moderate grammaticality (Luka & Barsalou, 2005). This effect of increased ratings after repeated exposure to the syntactic structure follows naturally from the perspective that these are hard-to-

[FIGURE 2 ABOUT HERE]

process constructions which get easier with experience. Any account of these effects must thus deal with the facts that (1) exposure raises acceptability judgments and (2) the effects of *wh*-phrase type combine.

4 Experiment II: *Wh*-Phrase Complexity

4.1 Materials & Participants

The previous study yielded the unexpected finding that *which*-$\overline{\text{N}}$ phrases, compared to bare *wh*-words, in the object position of SUVs did not raise judgments in the presence of a bare *wh*-subject. In Experiment II, we address this result by exploring whether the absence of an acceptability difference constitutes a spurious null effect. The second purpose of Experiment II is to develop a more accurate understanding of why *which*-$\overline{\text{N}}$ phrases lead to higher judgments ratings than bare *wh*-words. Based on previous accounts, we might expect that the advantage associated with *which*-$\overline{\text{N}}$ phrases is lexically conditioned and does not extend to other syntactically complex *wh*-phrases. If the syntactic and semantic complexity of the *wh*-phrase affects processing and thereby judgments of acceptability, we would expect that other kinds of complex *wh*-phrases should also improve acceptability.

The materials for Experiment II consequently varied in terms of the content of the object *wh*-phrase. We included this manipulation in both SUV *wh*-orders and non-SUV orders, as shown below:

(13) a. Ted revealed what who invented.

b. Ted revealed what device who invented.

c. Ted revealed which device who invented.

d. Ted revealed who invented what.

e. Ted revealed who invented what device.

f. Ted revealed who invented which device.

SUV orders should receive significantly lower judgments than non-SUV orders, and complex *wh*-expressions should lead to higher judgments than bare *wh*-items. However, we did not investigate semantic properties distinguishing*what*-$\overline{\text{N}}$ phrases from *which*-$\overline{\text{N}}$ phrases, treating both as roughly equally more informative and syntactically more complex than the bare *wh*-word. The effect of the different kinds of *wh*-elements and *wh*-orders were compared by using contrast sum coding or deviation coding. In this coding schema, the mean

[TABLE 2 ABOUT HERE]

[TABLE 3 ABOUT HERE]

of one factor level is compared with the sum of the means. Moreover, it allows us to test for differences among the factor levels while minimizing the effects of collinearity.

18 experimental items were mixed with 52 fillers. 41 participants, none of whom participated in any of the other experiments, took this experiment via the internet. Procedure and analysis paralleled the methods described for Experiment 1 However, model comparisons justified only the inclusion of participants as random effects and by-participant random slopes for log list position in the model for SUV orders; for the model with all *wh*-orders, participants were included as a random effect factor without any random slope adjustments.

4.2 Results

As expected, non-SUV orders produced significantly higher ratings than SUV orders (see Table 2). Considering both types of orders, no effect of *wh*-phrase complexity emerges, although later list positions do result in higher judgments. The absence of main effects can be attributed to the fact that complex *wh*-phrases do not yield higher acceptability judgments in non-SUVs, as seen in Figure 3. According to additional modeling, the difference between the bare *wh*-item and the *which*-$\overline{\text{N}}$ conditions does not reach statistical significance in the non-SUV conditions ($p \geq .2$), so we do not comment on this further.

Focusing only on the SUV word orders, however, we find significant effects of *wh*-complexity. As Table 3 makes clear, bare *wh*-objects in SUVs lead to lower ratings than the overall mean. Indeed, both complex *what*-$\overline{\text{N}}$ phrases and *which*-$\overline{\text{N}}$ phrases produce higher ratings of acceptability than bare *wh*-objects in SUV word orders. Finally, for the SUV orders, the effect of list position is marginally significant: Judgments trend higher throughout the course of the experiment. List position had no significant effect on judgments for filler items ($t = 0.15$, $p = .88$).

4.3 Discussion

The data reveal that complex object *wh*-phrases followed by bare *wh*-words do raise judgments in SUV constructions. Hence, we tentatively conclude that the lack of such on effect in Experiment I represents a

[FIGURE 3 ABOUT HERE]

spurious null effect (the evidence from Experiment V also backs up this interpretation). The data in Figure 3, in combination with further findings about the ameliorative effects of *wh*-complexity, therefore point to a general relationship between the complexity of dislocated *wh*-filler phrases and acceptability that is not limited to *wh*-phrases containing the lexical item *which*. Therefore, any account of these effects must be able to relate the ameliorative effects associated with *which*-$\overline{\text{N}}$ phrases to complex *wh*-phrases more generally. And while we failed to see the same trends in the non-SUV conditions, robust evidence from Hofmeister (2007) and Hofmeister & Sag (2010) demonstrates the general ameliorative effects of *wh*-phrase complexity on resolving *wh*-dependencies. As in Experiment I, repeated exposure improves the acceptability of SUVs, although the effect is only marginal here (perhaps because the number of SUV items is half that in Experiment I).

Combined, Experiments I & II provide us with an empirical backdrop of acceptability differences that can be compared to how the same constructions are processed. It is clear that many of the basic intuitions regarding SUVs align with the experimental data, although these intuitions do not accompany a theory of why two *which*-$\overline{\text{N}}$ phrases are better than one. Moreover, these intuitions do not tell us whether the acceptability differences correspond to differences in processing, which is what we consider next.

5 Experiment III: Comprehension

5.1 Methods

While the acceptability results are consistent with the predictions of prior psycholinguistic research on dependency processing, we now seek to provide direct confirmation that manipulations of *wh*-complexity affect processing in a manner that parallels their effects on acceptability judgments.

Experiment III is a self-paced, moving window reading experiment, where participants read sentences at their own pace on a computer screen (Just, Carpenter, & Woolley, 1982). Initially, they see a screen of dashes separated by spaces, representing the words for that trial. By pressing a predefined key, a new word appears on the screen and the previous word disappears. Following standard practice, higher reading times are construed as indicators of greater processing difficulty. For the reading time analyses, we utilize length-residualized reading times (Ferreira & Clifton, 1986). Estimates of the effect of

word length are based on all experimental stimuli, including fillers (but not practice items). The LME model also included a fixed effect factor (SPILLOVER) that models the relationship between reading times at region n and $n-1$. This factor controls for the possibility that reading time differences prior to the critical region may account for a significant amount of the observed differences across conditions (see http://hlplab.wordpress.com/2008/01/23/modeling-self-paced-reading-data-effects-of-word-length-word-position-spill-over-etc/). In Experiment III, participants and items were included as random factors in the LME model, by-participant random slopes for log list position without a correlation parameter, and by-item random slopes for spillover without a correlation parameter.

In both reading time experiments we describe, comprehension questions follow all experimental items. Thus, we also report how the experimental manipulations impact comprehension question accuracy. Question-answer accuracy was analyzed using mixed logit models (Jaeger, 2008). In this case, the mixed logit model included random effects for subjects and items (no random slopes). Data from participants with a question-answer accuracy less than 67% and/or an average reading time more than 2.5 standard deviations from all other participants were discarded prior to statistical analysis. Furthermore, all reading time data come from correctly answered items. Finally, we eliminated data points more than 2.5 standard deviations from the condition mean at each word region to reduce the impact of outliers. In Experiment III, this affected 1.91% of the original data.

5.2 Materials & Participants

The materials for this study were adaptations of the 20 items from Experiment I (post-verbal PPs were added to allow for spillover effects), varying in terms of the complexity of the *wh*-object (*what* vs. *which*-$\overline{\text{N}}$ phrase) and the complexity of the *wh*-subject (*who* vs. *which*-$\overline{\text{N}}$ phrase).

(14) Ashley disclosed (what/which agreement) (who/which diplomat) **signed after receiving permission** from the president.

According to a processing account of the acceptability contrasts, *which*-$\overline{\text{N}}$ phrases should lead to faster processing at the verb than bare *wh*-words. Two *which*-$\overline{\text{N}}$ phrases should thus result in the fastest processing times, the condition with two bare *wh*-words should have the slowest times, and the conditions with one *which*-$\overline{\text{N}}$ phrase should be intermediate in processing time. To control for the possibility that processing effects may extend beyond the subcategorizing verb, we take the verb and the following three words (the

[TABLE 4 ABOUT HERE]

[FIGURE 4 ABOUT HERE]

spillover region for the verb) as our critical area of interest (bolded in (14)).

64 filler items appeared along with the critical items, 40 of which belonged to an unrelated experiment. These fillers did not contain any multiple *wh*-questions and were of variable length and complexity. 41 individuals participated in this experiment at MIT's Tedlab for $10/hr.

5.3 Results

Reading times As expected, *wh*-phrase complexity significantly impacts reading times at the verb and the adjacent spillover area (see Figure 4). As Table 4 shows, this holds true for both fronted *wh*-objects and in-situ *wh*-subjects: Participants process the retrieval region faster in conditions with one or more *which*-$\overline{\text{N}}$ phrases. Hence, these findings replicate the *wh*-complexity effects of Experiments I and II. Unlike Experiment I, however, these two main effects are unqualified by an interaction. A single *which*-$\overline{\text{N}}$ phrase in either clause-initial or subject position results in faster processing times than analogous sentences with two bare *wh*-words. These findings are not contingent upon reading time differences prior to the critical region, as the spillover variable controls for this source of variation. Word-by-word reading times are shown in Table 5.

The reading time results parallel the acceptability results in another way. The later the item appears in the experiment, the faster the participants read both the verb and the words in the spillover region. Model comparisons with and without the list position factor also suggest that its inclusion benefits the model (χ^2 = 105.45, $df = 2$, $p < .0001$).

Question Answering Question-answer accuracies also reveal effects of *wh*-phrase complexity (BARE-BARE = 84.4%, *SE* = 2.5; BARE-WHICH = 89.8%, *SE* = 2.1; WHICH-BARE = 81.5%, *SE* = 2.7; WHICH-WHICH = 95.1%, *SE* = 1.5). As is the case in the reading time results, complex *wh*-phrases in subject position facilitate comprehension performance ($\beta = 1.12$, $z = 4.40$, $p < .001$). However, in contrast with the reading data, the question answering accuracies do not yield a significant effect of object *wh*-complexity ($p = .22$).

[TABLE 5 ABOUT HERE]

Instead, the accuracy results display an interaction of object and subject complexity (β = 1.10, z = 2.18, $p < .05$), which stems from the fact that complex *wh*-objects preceding a bare *wh*-word do not improve comprehension performance. This interaction echoes the Experiment I findings.

5.4 Discussion

These reading time data present novel evidence that manipulations that increase acceptability judgments for SUVs also facilitate processing. Complex *wh*-phrases, compared to simple *wh*-words, aid in the processing of the verb (and its spillover regions) that selects both of these *wh*-phrases as dependents. The processing advantage for *which*-$\overline{\text{N}}$ phrases occurs not only in the reading times, but also in the question-answer accuracies. Moreover, as in the acceptability study, multiple *which*-$\overline{\text{N}}$ phrases yield larger effects than a single *which*-$\overline{\text{N}}$ phrase.

The combined reading time and acceptability data, therefore, suggest a relationship between processing difficulty and acceptability judgments. The parallels between the reading time and acceptability data can be viewed as evidence that processing difficulty feeds into perceptions of well-formedness. This interpretation is supported by the observation that reading times in the critical retrieval region decrease with later list positions in the experiment, while acceptability judgments *increase* with list position for the same items. Assuming these longitudinal effects are related, the upward trend in acceptability judgments is consistent with the idea that Superiority effects reflect online processing costs that can be attenuated with repeated exposure.

6 Experiment IV: Comprehension and Judgments

6.1 Materials & Participants

The conclusion that processing times influence acceptability judgments in SUVs is based on comparing the results of different experiments with different participants. Hence, we lack direct evidence that the lower judgments of acceptability accompany increased processing times. Accordingly, in our final experiment, we combine the methods of collecting comprehension data and judgment data. Here, participants read word-by-word, as in the previous experiment. After answering a comprehension question, participants rated the sentence on a scale of 1 to 7 (1 being 'extremely unnatural' and 7 being 'extremely natural'). Combining

fixed scale judgments with self-paced reading reduced the overall length and complexity of the experiment, making this combination preferable to one with ME.

The materials for this experiment were identical to those in Experiment III, except for the 72 fillers. 24 fillers belonged to a separate experiment on resumptives, 32 were grammatical sentences which were relatively easy to process, and 16 were grammatical but hard-to-process sentences, e.g. *We overheard which student an administrator that Eric criticized documented that Laura defended after the ceremony*. Based on the previous results, we expected that bare *wh*-words would yield lower judgments of acceptability and increased processing times. Moreover, we anticipated that judgments would exhibit a significant relationship with processing times at the critical verb and spillover regions in the same sentence.

For the analysis of the acceptability results, the average reading time of the words in the critical region on the corresponding trial was included as a fixed effect factor in the LME model. This model also included fixed effect factors for the two complexity manipulations, their interaction, log list position, and random effects of participants and items. Model comparisons showed that the optimal random effect structure includes by-participant random slopes for *wh*-object complexity, *wh*-subject complexity, their interaction, log list position, and residual reading time at the critical region for the corresponding trial (and correlation parameters for all of them).

For the reading time results, outlier exclusion affected 2.8% of the data. Along with the complexity factors, the reading time LME model included fixed effect factors for spillover and logged list position, and random effect factors for participants, but not items. The model also contained by-participant random slopes for spillover (without a correlation parameter), *wh*-subject complexity (with a correlation parameter) and log list position (with a correlation parameter). The question-answer accuracy model contained random effect factors for subjects and items, but no random slopes. For both the acceptability and the reading time data, we report the results from the model without correlation parameters.

32 Stanford University students were paid $14/hr for their participation in this experiment and an unrelated set of experiments.

6.2 Comprehension Results

Reading times Reading times at the verb and its three-word spillover region display main effects of both object and subject *wh*-phrase complexity (see Figure 5 & Table 6). In particular, reading times were faster in the critical region in conditions with one or more *which*-$\overline{\text{N}}$ phrase, and the condition with two *which*-$\overline{\text{N}}$

[FIGURE 5 ABOUT HERE]

[FIGURE 6 ABOUT HERE]

phrases led to the overall fastest reading times (word-by-word reading times are shown in Table 7). These main effects did not accompany an interaction. Furthermore, as in Experiment III, list position significantly predicts reading times: Reading times decrease as list position increases. In short, the reading time results replicate our findings from Experiment III.

Question answering The question-answer accuracies evidence a marginal effect of object complexity (β = .50, z = 1.86, p = .06): Complex *wh*-elements lead to improved comprehension accuracy (BARE-BARE = 85%, *SE* = 2.8; BARE-WHICH = 87.5%, *SE* = 2.6; WHICH-BARE = 89.4%, *SE* = 2.4; WHICH-WHICH = 91.9%, *SE* = 2.2). A numerical trend for a similar effect of subject *wh*-complexity exists, but does not reach significance. Hence, while these results do not replicate the significant effects of question-answer accuracy found in Experiment III, they pattern similarly.

6.3 Acceptability Results

Following the pattern of results seen so far, complex *wh*-phrases elicit higher judgments than bare *wh*-words in SUVs. This finding holds for both object and subject *wh*-phrases ($ps < .001$). As in some of our previous studies, these two factors create a significant interaction, as Table 8 specifies. In this case, the interaction is not because complex object *wh*-phrases fail to raise judgments in the presence of a bare *wh*-subject (see Figure 6). Instead, the interaction appears to occur because two *which*-$\overline{\text{N}}$ phrases raise judgments more than one would expect on the basis of individual *which*-$\overline{\text{N}}$ phrases in either position. Furthermore, judgments significantly increase with later list positions in the experiment, which aligns with the decrease in processing times.

In parallel with Experiments 1 & 2, we did not find a significant relationship between log list position and judgments for the resumptive filler items ($t = -0.17$, p = .86) or the grammatical easy sentences ($t = -0.71$, p = .48). We did, however, observe a significant effect of log list position for the challenging filler items, (t = 2.08, p = .04).

[TABLE 6 ABOUT HERE]

[TABLE 7 ABOUT HERE]

[TABLE 8 ABOUT HERE]

Mixed effects modeling also demonstrates that mean residual reading times are a highly significant predictor of acceptability judgments. As residual reading times increase in the critical region (the verb and the subsequent three spillover words), judgments of acceptability decrease. Figure 7 illustrates that this relationship appears in all four conditions. Comparisons of LME models with and without the fixed effect predictor of reading times verify the significance of this factor in the model of acceptability judgments (χ^2 = 23.06, *df* = 7, *p* = .002). Further analysis verifies that this relationship does not depend on the influence of data points at the left or right periphery. Even after trimming off the data below the 25% quantile and above the 75% quantile of the residual reading times, reading time remains a significant predictor in the model of acceptability judgment as shown by model comparisons with and without this factor (χ^2= 5.95, *df* = 1, *p* = .015).

6.4 Discussion

This dual-task experiment replicates the findings from the preceding experiments, but it also provides direct evidence for the link between processing difficulty and judgments of acceptability. Of primary interest here is that complexity manipulations generated processing differences that mirrored differences in acceptability for the same items and with the same participants. These two sets of data also allow us to confirm directly that as processing times decrease throughout the experiment, judgments rise. The hypothesis that active processing constraints factor into judgments of Superiority violations presents an explanation for these longitudinal effects. Similarly, a processing account accords with the finding that two *which*-$\overline{\text{N}}$ phrases elicit judgments higher than expected based on how sentences with just one such phrase were judged. Specifically, several factors influencing processing may combine to produce effects that rise to a level beyond what is expected on the basis of each factor in isolation. This may occur because of the existence of a processing bottleneck that limits cognitive control to one task at a time (Pashler, 1994), or to the temporary exhaustion of a finite set of cognitive resources (Cowan, 2001; Fedorenko, Patel, Casasanto, Winawer, & Gibson, 2009;

[FIGURE 7 ABOUT HERE]

Just & Carpenter, 1992; Kluender, 1998).

7 General discussion

We began this investigation by noting the difficulty of distinguishing effects of grammar and processing effort on acceptability judgments. Superiority and D-linking represent cases whose analyses are complicated by this source ambiguity problem. After highlighting the difficulties that past attempts at syntactic analyses have encountered, we explored experimental effects due to the representational richness of the *wh*-phrases in SUVs, and investigated to what extent the acceptability contrasts could be predicted on the basis of processing principles. Our experimental investigations produced several key findings: (1) judgments and reading times for SUVs vary significantly with the complexity of the *wh*-elements in parallel ways; (2) judgments increase, where processing difficulty decreases; (3) effects of the complexity of the *wh*-phrases appear to combine. These findings follow naturally from the perspective that processing pressures play a major role in judgments of SUVs, and that differences in similarity-based interference separate D-linked from non-D-linked cases. These differences are not only replicable, but the processing contrasts work in ways that subsume effects previously attributed to grammar, and go beyond grammatical accounts in explaining variance previously unaccounted for by such accounts (i.e. how and why effects of *wh*-complexity combine).

Evidence outside the domain of SUVs further suggests that the syntactic and semantic complexity of filler-phrases affects processing and judgments in a variety of contexts. Judgments for a variety of sentences with syntactic island violations, including *wh*-islands and adjunct islands (Ross, 1967), improve with the use of a *which*-$\overline{\text{N}}$ phrase, as opposed to a bare *wh*-word (Hofmeister, 2007; Hofmeister & Sag, 2010):

(15) (Who / Which employee) did Albert learn whether they dismissed after the annual performance review?

(16) I knew (who / which staff members) my boss said she was calm before meeting ___ in the White House yesterday.

Importantly, these effects also occur in contexts without proposed grammatical constraint violations or even *wh*-phrases. Hofmeister (2007, in press) presents reading time evidence that processing at retrieval sites in filler-gap dependencies improves significantly with greater syntactic and semantic complexity of the target

representation. For instance, in cleft constructions like (17) below, the complexity of the clefted indefinite has a significant linear effect on reading times immediately after the subcategorizing verb. Specifically, participants read the word regions following the verb *banned* faster after a more complex retrieval target:

(17) It was a(n) (alleged (Venezuelan)) communist who the members of the club banned ___ from ever entering the premises.

Crucially, the reading time advantage for complex phrases emerges only at or immediately after the retrieval site (i.e. they are not contingent upon any preceding reading time differences). Thus, the relationship between the complexity of linguistic representations and how they are reaccessed looks to be a general phenomenon, and not observable only in cases where grammaticality is in question.

Other research on memory for list items also points to the mnenomic benefits of increased semantic processing or elaboration (Anderson & Reder, 1979; Anderson et al., 2001; Bradshaw & Anderson, 1982; Gallo, Meadow, Johnson, & Foster, 2008; Jacoby & Craik, 1979; McDaniel, 1981; McDaniel, Dunay, Lyman, & Kerwin, 1988; Reder, 1979; Reder, Charney, & Morgan, 1986; Stein, Morris, & Bransford, 1978). This research converges on the idea that processing meaning-related features has numerous potential advantages for memory retrieval when the retrieval context centers on meaning-related properties (as in language comprehension). For instance, adding semantic features may create a network of information that links together memory traces. Bradshaw & Anderson (1982) also suggest that this elaboration provides a means for a comprehender to redundantly encode information, creating "alternative retrieval paths." McDaniel et al. (1988, p. 358) go on to suggest that creating mental representations that encode unique relationships "would favor memory performance because it would reduce interference from related encodings."

There are further reasons to think that SUVs with complex *wh*-phrases should be easier to process than those with bare *wh*-words. For instance, it may be easier for comprehenders to imagine an appropriate context for the use of an SUV with lexically restricted *wh*-elements. By a similar reasoning, the result of building a mental representation from complex *wh*-phrases may be more imageable, which also conceivably has positive effects on processing. Although it is not immediately evident how such possibilities should be formally integrated with models of sentence processing, such performance-related factors may have a hand to play in the complexity effects investigated here.

Consequently, effects previously ascribed to a grammatical principle like D-linking are likely reducible to general processing principles whose consequences appear in a variety of sentence contexts (and even

outside language use). In essence, the acceptability contrasts between SUVs with complex *wh*-phrases and those without is just as we would expect, given our knowledge of how phrasal complexity interacts with the retrieval process in sentence comprehension. An account of these effects that invokes processing constraints not only describes the relevant contrasts—it provides an independently motivated explanation. A grammatical account, no matter how observationally successful, fails to provide such an explanation. Thus, processing-based approaches to syntactic phenomena have an inherent explanatory advantage. By accepting the explanatory value of the processing account, we can ultimately lighten the obligations of grammar and unify our account of human behavior in terms of independently motivated cognitive constraints. In this sense, a processing-based perspective on these acceptability contrasts has considerable merits that make it preferable to an analysis dependent on grammatical principles.

Given that contrasts within the class of SUVs are consistent with a processing account, do we need grammatical constraints to account for Superiority effects more generally? Multiple *wh*-questions with SUV orders clearly differ from those without SUV orders, insofar as two *wh*-elements must be processed before the verbal head. That is, there is an *inherent* difference in the degree of interference expected for each type of *wh*-order. Even in relatively acceptable SUVs with two *which*-$\overline{\text{N}}$ phrases, the *wh*-elements have identical determiner heads. In non-SUV multiple *wh*-questions, only a single *wh*-phrase appears before the critical verb, making retrieval and thematic interpretation substantially easier.

Along these lines, Gordon et al. (2001) showed that when the NPs in object and subject relative clauses are of different types, e.g. pronoun vs. definite NP, the processing difference between the relative clause types is diminished. Correspondingly, utterances such as *I know which book the students mentioned* are unquestionably grammatical, but as we increase the similarity between the two embedded nominals, the sentence becomes increasingly harder to process (*I know which book who mentioned* $\leq$ *I know which book what mentioned* $\leq$ *I know which book which book mentioned*). Thus, there are reasonable grounds for suspecting that SUV *wh*-orders impose significant processing challenges that are absent in non-SUV *wh*-orders.

It may accordingly seem economical to conclude that grammatical constraints are unnecessary to explain Superiority effects. However, the contrast between multiple *wh*-orders with SUVs and without is quite large (see Table 2). The data do not rule out a more complex account of the reduced acceptability of SUVs. What the evidence shows, however, is that any account of SUVs should acknowledge the effects of processing differences. Our strategy, therefore, in tackling the source ambiguity problem is not to remove

any possibility of grammatical influence (which is probably impossible), but to highlight evidence that processing factors play a fundamental role in the acceptability contrasts.

Ultimately, discriminating between possible sources of acceptability decrements depends on assumptions about the state of the grammar. These include what grammatical constraints look like, how sensitive they are to processing difficulty and other environmental factors, etc. Some theories of grammar draw much stronger distinctions between grammar and processing difficulty than others. On rule-based, non-gradient approaches to syntax—what we have focused on here—these two sources are largely independent of each other. But under some conceptions of grammar, these sources are more tightly linked. For instance, grammaticality constraints are sometimes defined in terms of graded preferences, weights, or rankings, rather than categorical or discrete levels of grammaticality. Such views are represented in theories of usage and frequency-based grammar (Bybee & Hopper, 2001; Bybee, Perkins, & Pagliuca, 1994; Goldberg, 2006; Kemmer & Barlow, 2000; Lanckager, 2000; MacWhinney, 1998; Tomasello, 2003), stochastic OT (Bresnan, 2000; Bresnan, Dingare, & Manning, 2001; Keller, 2000; Sorace & Keller, 2005), and exemplar-based accounts (Bod, 2006, 2009; Johnson, 1997; Pierrehumbert, 2001). With a less categorical view of grammar, these theories can describe syntactic phenomena in terms of grammaticality that emerges from preferences that develop over lexicalized phrases or constructions (Arnon & Snider, 2010; Bannard & Matthews, 2008; Bresnan & Hay, 2008; Fox & Thompson, 2007; Thompson & Mulac, 1991). In turn, such preferences can be linked to factors that affect processing difficulty, e.g. frequency of use, prototypicality, etc. (Bates & MacWhinney, 1982; Bybee, 2007; Bybee & Hopper, 2001; Jaeger, 2006). Perceptions of well-formedness would thus depend not only on characteristics of an individual token, but also on previous experiences with similar structures.

Such theories make radically different predictions about how syntactic constraints should align with processing difficulty and how constraints change with experience. For example, they are perfectly compatible with the possibility that a particular phrase structure might receive higher ratings of acceptability following repeated exposure to items with the same phrase structure. In other words, grammar (or the mechanisms by which judgments are made) constantly changes and adapts to the language use environment (for evidence in favor of this view, see Bradlow and Bent (2008); Diessel (2007); Fine, Qian, Jaeger, and Jacobs (2010); Jaeger (2010); Kaschak and Glenberg (2004); Kraljic and Samuel (2005, 2007); Snider and Jaeger (submitted); Wells, Christiansen, Race, Acheson, and MacDonald (2009)). Changes to these preferences that generalize over phrases and structures potentially come from many sources, including active online

processing differences. Thus, a construction that imposes serious processing difficulty may cause it to be a dispreferred syntactic alternative, leading to infrequent usage and even near categorical judgments of acceptability. Hence, unfamiliarity with the *wh*-ordering of SUVs may well be another factor compounding the difficulty of these items. In sum, the experimental findings presented here are consistent with the possibility that partially or completely grammaticized constraints contribute to the observed behavior. Such constraints may themselves be determined by processing difficulty, or may be acting alongside active processing pressures (Hawkins, 1994, 1999, 2004). It is a matter for future research to determine whether these options can be disentangled.

This investigation of the source ambiguity problem has demonstrated how functional factors may be identified as a contributor to acceptability contrasts. But a number of outstanding issues remain. Among these is the question of how repeated exposure effects relate to the source of unacceptability. Ratings for SUVs, regardless of the *wh*-elements involved, rose throughout the experiments, but we observed earlier that the findings regarding repeated exposure are equivocal. Some suggest that only grammatical strings improve with exposure, while others hint that any kind of sentence can. Consequently, our results with respect to SUVs can lead to the following conclusions: (1) ungrammatical strings are, in fact, sensitive to repeated exposure; or (2) SUVs are not ungrammatical.

A potential resolution of apparently conflicting findings regarding repeated exposure and grammaticality appears with Francom (2009). Francom argues that interpretability rather than grammaticality is crucial in understanding repeated exposure effects. To support this claim, Francom cites Maclay and Sleator's (1960) experimental evidence that participants rated anomalous items similarly (e.g. *Get me from the kitchen a big spoon*) whether they were instructed to rate items for meaningfulness or grammaticality. This aligns with some of our own laboratory investigations, where we pseudo-randomly moved several words in a sentence to an unlicensed position (e.g *Iran has gun-control strict laws that bar private citizens carrying from firearms*). Participants rated these sentences increasingly better with exposure in three different experiments with different fillers and participants.

On this view, SUVs are judged increasingly better because they become easier to interpret with exposure. Problematically for this analysis, some errors that do not become better with repetition nevertheless leave the sentence quite interpretable (e.g. *I saw he at the store*). The extent to which a structure improves with exposure may thus reflect several interacting factors, including not only interpretability but identifiability of "what's wrong" with the sentence or how to fix it. In short, assessments of repeated exposure effects by

themselves appear to have limited value in determining the source of acceptability in any given case.

Clearly, this research highlights the difficulties involved in separating effects of grammar from effects of processing. No single type of evidence excludes the possibility that processing or grammar contributes to variation in judgments. But, following the pioneering work of Kluender (1991, 1998, 2005) and Kluender and Kutas (1993b), the work presented here provides further evidence that at least some acceptability contrasts previously attributed to arbitrary grammatical constraints can be accounted for by independently motivated processing preferences. Finally, this work challenges any subsequent grammatical accounts of Superiority effects to go above and beyond the capabilities of a processing account to predict variation and acceptability contrasts.

8 Appendix A: Modeling Effects of *Wh*-Phrase Interference

Nairne (1990, 2001, 2006) suggests that the probability of retrieving an event (or mental representation) E_1, given a retrieval cue set X_1, depends upon the similarity or feature-overlap of X_1 and E_1, as well as the similarity of X_1 to other memory candidates.

$$P_r(E_1|X_1) = \frac{s(X_1, E_1)}{\sum s(X_1, E_n)} \tag{1}$$

The similarity between a memory item and a retrieval cue is determined by the proportion of mismatching features to all retrieval cue features (*d*):

$$s(X_1, E_1) = e^{-d(X_1, E_1)} \tag{2}$$

The makeup of X_1 depends in large part on the features in E_1. For instance, if a memory chunk consists of a vector of features such as [+C +D +E −G], the operative retrieval cues contain a perfectly maintained set of feature-values [+C +D +E −G] or a degraded one with missing features ([? ? +E −G]). Loss of features can happen for numerous reasons, such as feature overwriting, whose discussion goes beyond this paper's scope. Ultimately, the greater number of contextually unique features in E_1, the greater the probability for unique features in X_1, and thus better chances for successful retrieval.

To briefly illustrate, suppose that a *which*-$\overline{\text{N}}$ phrase like *which patient* has a feature representation like [+A +D +W $f_4 \ldots f_{12}$] (the names and values are arbitrary), *who* contains a subset of these features [+A +W

[TABLE 9 ABOUT HERE]

], *which medicine* has the feature representation [−A +D +W g_4 ... g_{12}], and *what* has a subset of these [−A +W], and that no *f*-features are *g*-features and vice versa. To calculate retrieval probabilities for each *wh*-element in an SUV context where the trace has been perfectly preserved ($X_1 = E_1$), we simply plug in the appropriate numbers to get the results in Table 9. For instance, the two *which*-$\overline{N}$ phrases, according to the vector representations above, mismatch on 10 out of 12 feature-values. Hence, if E_1 corresponds to the representation for *which medicine*, the similarity of X_1 to E_1 is 1, the similarity of *which patient* to X_1 will be .43, leading to the sampling probability (1/1.43) of .70 for the correct target. As Table 9 illustrates, the highest average sampling probability occurs in the context with two *which*-$\overline{N}$ phrases and the lowest with two bare *wh*-words, while contexts with one bare *wh*-word are predicted to have an intermediate status. The equations above also guarantee that as the proportion of feature mismatches between memory candidates increases, so does the sampling probability. This model-theoretic illustration should not be interpreted as a commitment to Nairne's model or to specific probability values, but rather to the idea that the general class of retrieval models to which it belongs—those that involve retrieval cues based on the original encoding—makes predictions about the relative ease of memory retrieval for specific pairings of *wh*-elements based on the degree of feature overlap.

9 Appendix B: Experimental Materials

9.1 Experiment I, III, & IV Stimuli

Only the condition with two *which*-$\overline{N}$ phrases is shown here.

1. Anna noticed which picture which collector desired when she attended the big art sale in the museum.
2. Danny observed which box which mover lifted while he continued packing up his things.
3. John determined which vase which child broke after talking with the children in the class.
4. David guessed which continent which explorer discovered and then won the big trivia match.
5. Lisa remembered which play which author wrote but she couldn't remember the year.
6. Ellen recalled which pie which customer ate when she was at the local diner last week.
7. Ashley discovered which agreement which diplomat signed after receving permission from the president.

8. Tom knew which key which janitor took after questioning the students in the school.

9. Michael researched which castle which king built because he was writing a report on medieval architecture.

10. Chris reported which island which millionaire bought and Dan passed that information to the board of trustees.

11. Sophia revealed which device which engineer invented and then she was sued for violating privacy laws.

12. Nicole announced which prize which actor received but no one could hear her over the noisy crowd.

13. Dan forgot which reform which politician promised before the crucial elections last June.

14. Laura discovered which treasure which pirate hid but she didn't discover the hiding place.

15. Peter asked which lecture which professor presented so Mary gave him the conference schedule.

16. Sarah predicted which race which jockey won when she went to the races last Saturday.

17. Ted investigated which necklace which suspect stole and then he reported it to the police.

18. Jeanette stated which medicine which patient needed and John took note of it.

19. Mary wondered which book which student read but later the teacher told her.

20. Andrew decided which weapon which fighter used before allowing the match to begin.

9.2 Experiment II Stimuli

Only the SUV order with *which*-$\overline{\text{N}}$ phrase condition appears here. All stimuli also appeared in non-SUV orders.

1. Mary wondered which book who read.

2. Peter asked which lecture who presented.

3. John inquired which bill who paid.

4. Michael questioned which song who wrote.

5. Ted investigated which war who started.

6. David guessed which continent who discovered.

7. Nicole announced which prize who received.

8. Tom knew which key who took.

9. Sarah predicted which race who won.

10. Anna said which picture who painted.
11. Sophia revealed which device who invented.
12. Chris discussed which collection who bought.
13. Jeanette indicated which law who broke.
14. Ellen saw which fruit who ate.
15. Danny observed which box who lifted.
16. Lisa remembered which play who directed.
17. Andrew published which property who sold.
18. Laura discovered which jewelry who stole.

References

Anderson, J. R., Budiu, R., & Reder, L. (2001). A theory of sentence memory as part of a general theory of memory. *Journal of Memory and Language*, *45*, 337-367.

Anderson, J. R., & Reder, L. (1979). An elaborative processing explanation of depth of processing. In L. Cermak & F. Craik (Eds.), *Levels of processing in human memory* (pp. 385–404). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routledge.

Ariel, M. (2001). Accessibility theory: an overview. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 29–87). Amsterdam: John Benjamins.

Arnon, I., Hofmeister, P., Jaeger, T. F., Sag, I. A., & Snider, N. (2005). *Rethinking Superiority effects: A processing model.* Poster presented at the CUNY Sentence Processing Conference. Tucson, AZ.

Arnon, I., & Snider, N. (2010). More than words: frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67-82.

Baayen, R. (2004). Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers*, *1*, 1–45.

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

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, *19*(3), 241.

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*, 32–68.

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. *Language acquisition: The state of the art*, 173–218.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, *23*, 291–320.

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, *33*, 752–793.

Bolinger, D. (1978). Asking more than one thing at a time. In H. Hiz (Ed.), *Questions* (pp. 107–150). Dordrecht: D. Reidel.

Bradlow, A., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Bradshaw, G., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, *21*, 165–174.

Braze, F. D. (2002). *Grammaticality, acceptability, and sentence processing: A psycholinguistic study*. Unpublished doctoral dissertation, University of Connecticut.

Bresnan, J. (2000). Optimality syntax. In J. Dekkers, F. van der Leeuw, & J. van de Weijer (Eds.), *Optimality Theory: Phonology, syntax and acquisition* (pp. 334–385). Oxford: Oxford University Press.

Bresnan, J., Dingare, S., & Manning, C. (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG conference* (Vol. 1, pp. 13–32).

Bresnan, J., & Hay, J. (2008). Gradient grammar: an effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua*, *18*, 245–259.

Bybee, J. (2007). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Bybee, J., & Hopper, P. (2001). Introduction to frequency and the emergence of linguistic structure. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 1–24). Amsterdam: John Benjamins Publishing Company.

Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago, IL: University of Chicago Press.

Chomsky, N. (1995). *The Minimalist program*. Cambridge, MA: MIT press.

Chomsky, N. (2000). Minimalist inquiries. In D. M. Roger Martin & J. Uriagereka (Eds.), *Step by step: Essays on Minimalism in honor of Howard Lasnik* (pp. 89–155). Cambridge: MIT Press.

Clifton, C., Fanselow, G., & Frazier, L. (2006). Amnestying superiority violations: Processing multiple questions. *Linguistic Inquiry*, *37*, 51–68.

Comorovski, I. (1989). *Discourse and the syntax of multiple constituent questions.* PhD Thesis. Cornell University.

Comorovski, I. (1996). *Interrogative phrases and the syntax-semantics interface*. Norwell, MA: Kluwer Academic Publishers.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Criss, A., & McClelland, J. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language*, *55*, 447–460.

Culicover, P. (2008). Linear order effects in computing grammatical dependencies. In *SFB Workshop: Syntactic constructions, focus and meaning*. University of Tuebingen.

Culicover, P., & Wilkins, W. (1984). *Locality in linguistic theory*. San Diego, CA: Academic Press.

Dayal, V. (2002). Single-pair versus multiple-pair answers: Wh-in-situ and scope. *Linguistic Inquiry*, *33*(3), 512–520.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*(2), 108–127.

Ericsson, K., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–244.

Featherston, S. (2005). Magnitude estimation and what it can do for your syntax: Some *wh*-constraints in German. *Lingua*, *115*, 1525–1550.

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition*, *37*(1), 1.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Fine, A., Qian, T., Jaeger, T. F., & Jacobs, R. (2010). Syntactic adaptation in language comprehension. In *Proceedings of ACL: Workshop on cognitive modeling and computational linguistics* (p. 18-26). Uppsala, Sweden.

Fox, B., & Thompson, S. (2007). Relative clauses in English conversation: Relativizers, frequency, and the notion of construction. *Studies in Language*, *31*(2), 293–326.

Francom, J. (2009). *Experimental syntax: Exploring the effect of repeated exposure to anomalous syntactic structure–evidence from rating and reading tasks*. Unpublished doctoral dissertation, University of Arizona.

Gallo, D., Meadow, N., Johnson, E., & Foster, K. (2008). Deep levels of processing elicit a distinctiveness heuristic: Evidence from the criterial recollection task. *Journal of Memory and Language*, *58*, 1095–1111.

Gibson, E. (1990). A computational theory of processing overload and garden-path effects. In *Proceedings of the 13th conference on computational linguistics* (pp. 114–119). Morristown, NJ: Association for Computational Linguistics.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge: MIT Press.

Ginzburg, J., & Sag, I. A. (2000). *Interrogative investigations*. Stanford: CSLI publications.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gordon, P., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1411–23.

Gordon, P., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimantal Psychology: Learning, Memory, Cognition*, *32*, 1304–1321.

Gordon, P., Hendrick, R., & Levine, W. (2002). Memory-load interference in syntactic processing. *Psychological Science*, *13*, 425-430.

Grohmann, K. (2003). German is a multiple *wh*-fronting language. In C. Boeckx & K. Grohmann (Eds.), *Multiple wh-fronting.* Philadelphia: John Benjamins.

Hawkins, J. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hawkins, J. (1999). Processing complexity and filler-gap dependencies across grammars. *Language*, *75*,

244-285.

Hawkins, J. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hofmeister, P. (2007). *Representational complexity and memory retrieval in language comprehension.* PhD Thesis, Stanford University.

Hofmeister, P. (in press). Representational complexity and memory retrieval in language comprehension. *Language and Cognitive Processes*.

Hofmeister, P., Jaeger, T. F., Sag, I. A., Arnon, I., & Snider, N. (2007). Locality and accessibility in *wh*-questions. In S. Featherston & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base* (pp. 185–206). Berlin: Mouton de Gruyter.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *82*, 366-415.

Jacoby, L., & Craik, F. (1979). Effects of elaboration of processing at encoding and retrieval: trace distinctiveness and recovery of initial context. In L. Cermak & F. Craik (Eds.), *Levels of processing in human memory* (pp. 1–21). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jaeger, T. F. (2004). Topicality and Superiority in Bulgarian wh-questions. In *Proceedings of formal approaches to Slavic linguistics: The Ottawa meeting 2003* (Vol. 12, pp. 207–228). Michigan Slavic Publications.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished doctoral dissertation, Stanford University.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62.

Jaeger, T. F., & Tily, H. (in press). On language 'utility': Processing complexity and communicative efficiency'. *WIRE: Cognitive Science*.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. Mullenix (Eds.), *Talker variability in speech processing.* San Diego, CA: Academic Press.

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Just, M., Carpenter, P., & Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228–238.

Kaschak, M., & Glenberg, A. (2004). This construction needs learned. *Journal of Experimental Psychology: General*, *133*, 450–467.

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Unpublished doctoral dissertation. University of Edinburgh, College of Science and Engineering, School of Informatics.

Kemmer, S., & Barlow, M. (2000). Introduction: A usage-based conception of language. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language.* Stanford, CA: CSLI.

Kluender, R. (1991). *Cognitive constraints on variables in syntax*. Unpublished doctoral dissertation, University of California, San Diego.

Kluender, R. (1998). On the distinction between strong and weak islands: A processing perspective. In P. Culicover & L. McNally (Eds.), *Syntax and semantics 29: The limits of syntax* (pp. 241–279). San Diego, CA: Academic Press.

Kluender, R. (2005). Are subject islands subject to a processing account? *Proceedings of the 23rd west coast conference on formal linguistics*, 475–499.

Kluender, R., & Kutas, M. (1993a). Bridging the gap: Evidence from erps on the processing of unboudned dependencies. *Journal of Cognitive Neuroscience*, *5*, 196-214.

Kluender, R., & Kutas, M. (1993b). Subjacency as a processing phenomenon. *Language and Cognitive Processes*, *8*, 573-633.

Kraljic, T., & Samuel, A. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.

Kraljic, T., & Samuel, A. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Labov, W. (1973). *Sociolinguistic patterns*. Philadelphia, PA: Univ of Pennsylvania.

Labov, W. (1996). When intuitions fail. In L. D. L. McNair K. Singer & M. Aucon (Eds.), *Papers from the parasession on theory and data in linguistics: Chicago Linguistic Society* (Vol. 32, pp. 77–105). Chicago.

Lanckager, R. (2000). A dynamic usage-based model. In S. Kemmer & M. Barlow (Eds.), *Usage-based models of language.* Stanford, CA: CSLI.

Lewis, R. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*, 93–115.

Lewis, R. (1999). Accounting for the fine structure of syntactic working memory: Similarity-based inter-

ference as a unifying principle. *Behavioral and Brain Sciences*, *22*, 105–106.

Lewis, R., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1-45.

Lewis, R., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*, 447–454.

Luka, B., & Barsalou, L. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, *52*, 436–459.

Maclay, H., & Sleator, M. (1960). Responses to language: Judgments of grammaticalness. *Journal of American Linguistics*, *26*, 275–282.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, *49*, 199–227.

Maling, J., & Zaenen, A. (1982). A phrase structure account of Scandinavian extraction phenomena. In P. Jacobson & G. Pullum (Eds.), *The nature of syntactic representation* (pp. 229–282). Dordrecht: Reidel.

McDaniel, M. (1981). Syntactic complexity and elaborative processing. *Memory and Cognition*, *31*, 35–43.

McDaniel, M., Dunay, P., Lyman, B., & Kerwin, M. L. (1988). Effects of elaboration and relational distinctiveness on sentence memory. *The American Journal of Psychology*, *101*, 357-369.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*, 111–123.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*, 67–91.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 419–492). New York: Wiley.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*, 251–269.

Nairne, J. S. (2001). A functional analysis of primary memory. In H. Roediger, J. S. Nairne, & A. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 282–296). Washington, D.C.: American Psychological Association.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 27–46). Oxford: Oxford University Press.

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research*, *18*, 5–19.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review*, *115*, 544–576.

Osterhout, L., & Swinney, D. (1993). On the temporal course of gap-filling during comprehension of verbal passives. *Journal of Psycholinguistic Research*, *22*, 273–286.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*, 220-244.

Pesetsky, D. (1987). *Wh*-in-situ: movement and unselective binding. In E. Reuland & A. ter Meulen (Eds.), *The representation of (in)definiteness* (pp. 98–129). Cambridge: MIT Press.

Pesetsky, D. (2000). *Phrasal movement and its kin*. Cambridge: MIT Press.

Pickering, M. (1993). Direct association and sentence processing: A reply to Gorrell and to Gibson and Hickok. *Language and Cognitive Processes*, *8*, 163–196.

Pickering, M., & Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes*, *6*, 229–259.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS statistics and computing*. New York: Springer.

Prince, E. (1978). A comparison of *wh*-clefts and it-clefts in discourse. *Language*, *54*, 883–906.

Reder, L. (1979). The role of elaborations in memory for prose. *Cognitive Psychology*, *11*, 221–234.

Reder, L., Charney, D., & Morgan, K. (1986). The role of elaborations in learning a skill from an instructional text. *Memory and Cognition*, *14*, 64–78.

Ross, J. R. (1967). *Constraints on variables in syntax*. Unpublished doctoral dissertation, MIT. (Published in 1986 as *Infinite Syntax!* by Ablex: Norwood, N. J.)

Snider, N., & Jaeger, T. F. (submitted). *Syntax in flux: Structural priming maintains probabilistic representations.*

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, *31*, 575–582.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*, 1497–1524.

Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*,

40, 329–341.

Stein, B., Morris, D., & Bransford, J. (1978). Constraints on effective elaboration. *Journal of Verbal Learning and Verbal Behavior*, *17*, 707–714.

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

Swinney, D., Ford, M., Bresnan, J., & Frauenfelder, U. (1988). Coreference assignment during sentence processing. In M. Macken (Ed.), *Language structure and processing.* Stanford, CA: CSLI.

Tanenhaus, M., Carlson, G., & Seidenberg, M. (1985). Do listeners compute linguistic representations? In A. Zwicky, L. Kartunnen, & D. Dowty (Eds.), *Natural language parsing: Psycholinguistic, theoretical, and computational perspectives* (pp. 359–408). London and New York: Cambridge University Press.

Thompson, S., & Mulac, A. (1991). A quantitative perspective on the grammaticalization of epistemic parentheticals in English. *Approaches to grammaticalization*, *2*, 313–339.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Van Dyke, J. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 407-430.

Van Dyke, J., & Lewis, R. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316.

Van Dyke, J., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*, 157-166.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and anti-locality effects. *Language*, *82*, 767-794.

Warren, T., & Gibson, E. (2005). Effects of NP type in reading cleft sentences in English. *Language and Cognitive Processes*, *20*, 751-767.

Wells, J., Christiansen, M., Race, D., Acheson, D., & MacDonald, M. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology*, *58*(2), 250–271.

Wilkins, W. (1977). *The variable interpretation condition*. Unpublished doctoral dissertation, University of California-Los Angeles.

Wilkins, W. (1980). Adjacency and variables in syntactic transformations. *Linguistic Inquiry*, *11*, 709–758.

	$Estimate$	$HPD95lower$	$HPD95upper$	p
WHOBJ	0.21	0.14	0.30	$<.001$
WHSUBJ	0.47	0.38	0.57	$<.001$
WHOBJ X WHSUBJ	0.34	0.17	0.49	$<.001$
LOG(LIST POSITION)	0.13	0.05	0.22	$<.001$

Table 1: Fixed effects summary for Experiment I with participants as a random factor and by-participant random slopes for subject *wh*-complexity and log list position (WHOBJ = complexity of *wh*-object; WHSUBJ = complexity of *wh*-subject)

	$Estimate$	$HPD95lower$	$HPD95upper$	p
WHAT	-0.01	-0.08	0.06	.740
WHAT-N	-0.03	-0.11	0.04	.352
LOG(LIST POSITION)	0.06	0.00	0.12	.047

Table 2: Fixed effects summary of contrast sum coding for Experiment II (all conditions) with participants as a random factor

	$Estimate$	$HPD95lower$	$HPD95upper$	p
WHAT	-0.12	-0.20	-0.06	.001
WHAT-N	0.06	-0.01	0.13	.105
LOG(LIST POSITION)	0.05	-0.02	0.13	.141

Table 3: Fixed effects summary of contrast sum coding for Experiment II (SUV orders only) with participants as a random factor with by-participant random slopes for log list position

	$Estimate$	$HPD95lower$	$HPD95upper$	p
WHOBJ	-24.70	-35.82	-12.65	$<.001$
WHSUBJ	-19.97	-32.68	-8.83	$<.001$
WHOBJ X WHSUBJ	-5.60	-29.62	17.73	.695
SPILLOVER	0.08	0.00	0.16	.056
LOG(LIST POSITION)	-49.24	-61.27	-37.33	$<.001$

Table 4: Fixed effects summary for Experiment III with participants and items as random factors, by-participant random slopes for log list position, and by-item random slopes for spillover (WHOBJ= *wh*-object complexity; WHSUBJ = *wh*-subject complexity)

	BARE-BARE	BARE-WHICH	WHICH-BARE	WHICH-WHICH
VERB $-$ 1	-3.50 (7.89)	43.36 (13.62)	4.29 (7.44)	-6.82 (8.38)
VERB	50.09 (13.22)	60.15 (13.53)	17.34 (11.98)	15.27 (7.58)
VERB + 1	73.75 (14.90)	32.38 (10.51)	43.20 (11.59)	11.03 (7.90)
VERB + 2	42.02 (12.90)	11.12 (9.75)	26.36 (9.99)	-25.03 (5.27)
VERB + 3	-10.33 (9.25)	-9.29 (9.68)	-6.05 (9.99)	-23.10 (8.10)

Table 5: Mean residual readings time for Experiment III from word preceding head verb to three words after with standard errors in parentheses

	$Estimate$	$HPD95lower$	$HPD95upper$	p
WHOBJ	−25.17	−42.20	−6.66	.006
WHSUBJ	−43.08	−62.60	−25.47	<.001
WHOBJ X WHSUBJ	15.39	−19.38	53.23	.385
SPILLOVER	0.15	0.07	0.23	<.001
LOG(LIST POSITION)	−52.75	−64.10	−41.40	<.001

Table 6: Fixed effects summary for Experiment IV reading times with participants as a random factor and by-participant random slopes for spillover, *wh*-subject complexity, and log list position (WHOBJ = *wh*-object complexity; WHSUBJ = *wh*-subject complexity)

	BARE-BARE	BARE-WHICH	WHICH-BARE	WHICH-WHICH
VERB − 1	23.15 (17.83)	80.96 (21.77)	18.68 (15.59)	21.99 (14.30)
VERB	114.20 (26.16)	42.39 (16.12)	34.27 (16.96)	20.51 (16.40)
VERB + 1	57.39 (17.88)	29.76 (15.46)	37.18 (15.69)	52.82 (17.77)
VERB + 2	71.59 (26.22)	6.63 (16.06)	36.83 (18.25)	26.80 (11.16)
VERB + 3	−35.72 (11.37)	−51.97 (10.54)	−46.31 (12.02)	−68.11 (9.08)

Table 7: Mean residual readings time for Experiment IV from word preceding head verb to three words after with standard errors in parentheses

	$Estimate$	$HPD95lower$	$HPD95upper$	p
WHOBJ	0.384	.251	.507	<.001
WHSUBJ	0.428	.300	.563	<.001
WHOBJ X WHSUBJ	0.370	.107	.620	.004
LOG(LIST POSITION)	0.190	.065	.287	.001
RESIDUAL RTS	−0.001	−.002	.000	.036

Table 8: Fixed effects summary for Experiment IV acceptability judgments with participants and items as random factors and by-participant random slopes for WHOBJ, WHSUBJ, WHOBJ X WHSUBJ, log list position, and residual RTs (WHOBJ = *wh*-object complexity; WHSUBJ = *wh*-subject complexity)

WH1	WH2	WH1 RETRIEVAL	WH2 RETRIEVAL	AVG. SAMP. PROB.
which medicine	which patient	d = 10/12 s(X_1, WH2) = .43	d = 10/12 s(X_1, WH1) = .43	(.70 + .70)/2 = .70
which medicine	who	d = 11/12 s(X_1, WH2) = .40	d = 1/2 s(X_1, WH1) = .61	(.71 + .62)/2 = .67
what	which patient	d = 1/2 s(X_1, WH2) = .61	d = 11/12 s(X_1, WH1) = .40	(.62 + .71)/2 = .67
what	who	d = 1/2 s(X_1, WH2) = .61	d = 1/2 s(X_1, WH1) = .61	(.62 + .62)/2 = .62

Table 9: Similarity values and average sampling probabilities according to model specifications in Nairne (1990, 2001, 2006), assuming that $X_1 = E_1$

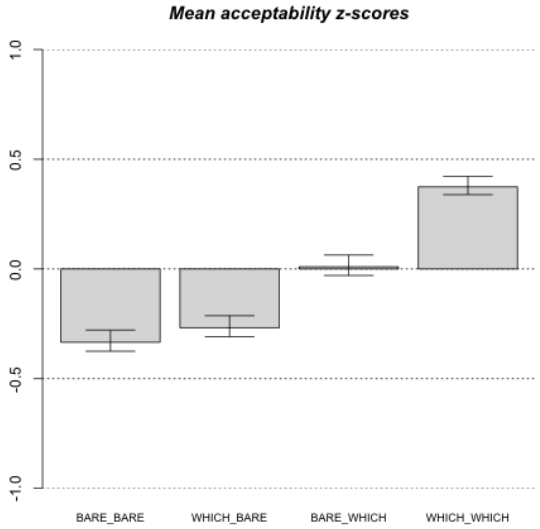

Figure 1: Mean z-scores of log normalized acceptability ratings from Experiment I, error bars show +/− one standard error

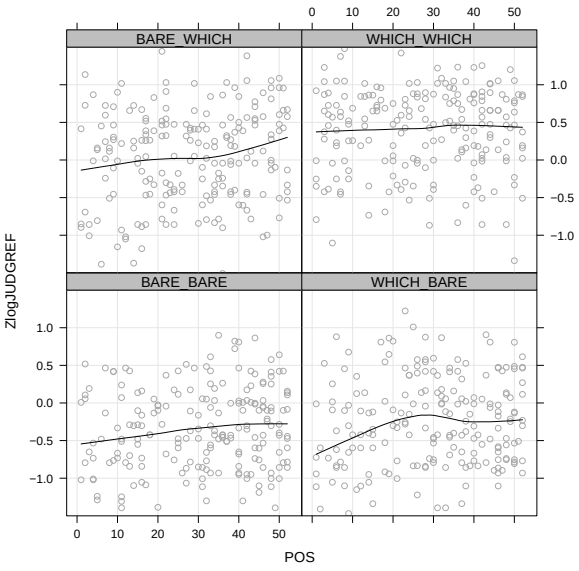

Figure 2: Scatterplot of acceptability z-scores by list position in Experiment I fitted with locally weighted scatterplot smoother (lowess curve).

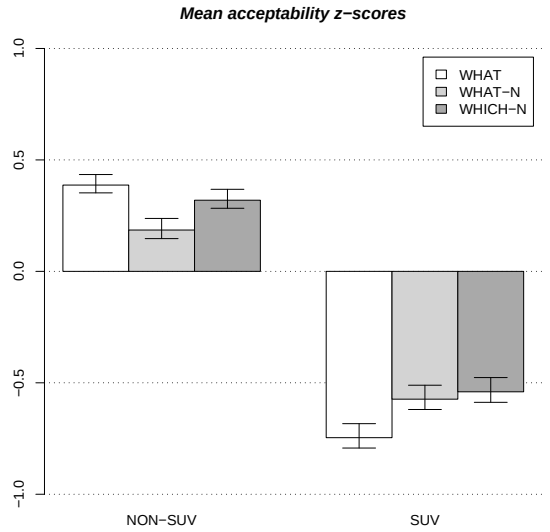

Figure 3: Mean z-scores of log-normalized acceptability ratings in Experiment II, error bars show +/− one standard error

Figure 4: Mean residual reading times in Experiment III, error bars show +/− one standard error

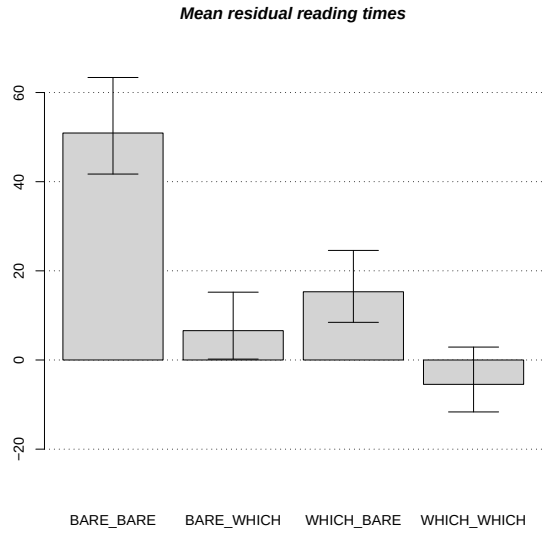

Figure 5: Mean residual reading times at verb + three-word spillover region in Experiment IV, error bars show +/− one standard error

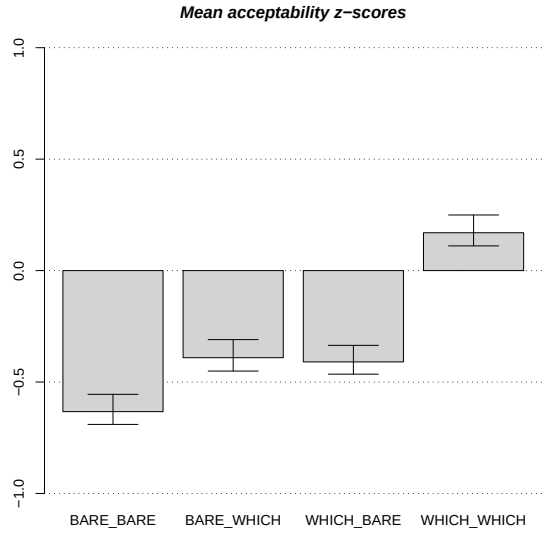

Figure 6: Mean acceptability z-scores in Experiment IV, error bars show +/− one standard error

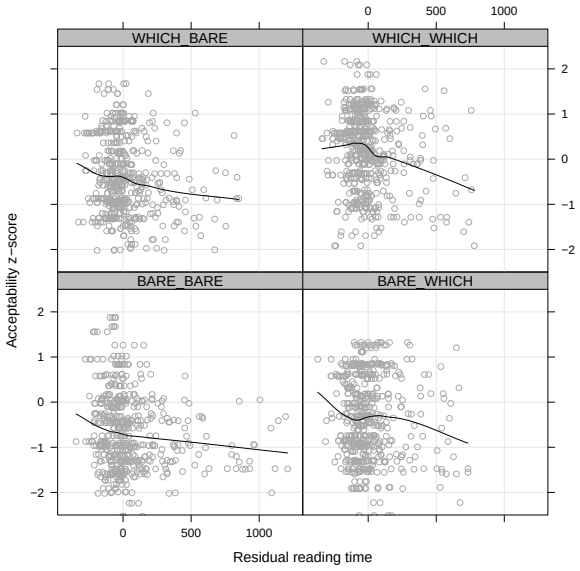

Figure 7: Scatterplot of relationship between mean reading times at critical word region and acceptability scores by condition in Experiment IV fitted with locally weighted scatterplot smoother (lowess curve)