Middle ratings rise regardless of grammatical construction:

Testing syntactic variability in a new repeated exposure paradigm

Jessica Brown, Gisbert Fanselow, Rebecca Hall, & Reinhold Kliegl

University of Potsdam

Address for correspondence:

Jessica Brown

Room 3.20, Haus 14

SFB 1287 Linguistic Variability

Department of Linguistics

Karl-Liebknecht Straße 24-25

University of Potsdam

14476 Potsdam, Germany

jmmbrown@cantab.net

ABSTRACT

People seem to perceive sentences more favourably after hearing or reading them many times. A prominent approach in linguistic theory claims that these types of exposure effects (satiation effects) show direct evidence of a generative approach to linguistic knowledge: only some sentences improve under repeated exposure, and which sentences do improve can be predicted by a model of linguistic competence that yields natural syntactic classes. However replications of the original findings have been inconsistent, and it remains unclear whether satiation effects can be reliably induced in an experimental setting at all. In this paper, we report the results of a new experimental paradigm (the Standardised Block paradigm) that reliably induces exposure effects in wh-question constructions across two languages. We report four new findings. First, the effects pertain to zone of well-formedness rather than syntactic class: all intermediate ratings, including calibrated distractor items, increase at the beginning of the experimental session regardless of syntactic construction. Second, ratings rise but do not satiate. Third, these effects are consistent across languages. Fourth, wh-question constructions show similar profiles in English and German, despite these languages being traditionally considered to differ strongly in whether they show effects on movement: in both languages violations of the superiority condition can be modulated to a similar degree by manipulating animacy or complexity of the wh-phrase. We demonstrate that the Standardised Block paradigm improves on classic satiation methods by allowing two crucial tests to be separated: whether repeated exposure effects exist at all is tested separately from whether exposure effects selectively target certain grammatical constructions. We conclude that repeated exposure effects can be reliably induced in rating experiments but exposure effects do not selectively target certain grammatical constructions. Instead repeated exposure effects are a phenomenon of gradient intermediate judgements.

INTRODUCTION

Linguistic theory relies crucially on how people perceive strange sentences. Strange sentences are important because they represent the border between the grammatical and the ungrammatical: determining the well-formedness of these sentences makes it possible to work

out what the output of a generative grammar should be by working out which sentences belong to the set of grammatical sentences and which sentences belong to the set of ungrammatical sentences. A large body of work demonstrates that people do indeed have consistently stable intuitions about even strange sentences, and that these intuitions can be reliably measured through well-formedness judgments [1-5].

Despite this overall stability, it is clear that the way that people perceive grammaticality changes with time and training and how many times someone sees a sentence. Early work starting in the eighties [6,7] claims that these changes can be reliably induced in a laboratory setting. In his seminal (2000) paper [8], Snyder goes further and suggests that such changes in fact yield direct evidence of a particular type of generative linguistic theory in action. Specifically, this approach claims that (i) people only change how they perceive some sentences, not how they perceive all sentences; and (ii) which sentences will be the ones to change under repeated exposure can be predicted by natural syntactic classes determined by the subjacency model of syntax [9].

However, attempts to replicate the findings in [8] using the classic satiation paradigm have yielded mixed results. Some studies find satiation effects for certain constructions but not others [10-14], whereas other studies [15-17] fail to find exposure effects at all (cf. Table 1). The picture is further complicated by arguments that rating data are inherently ambiguous with regard to the source of differences, regardless of whether rating data show exposure effects or not [18].

We argue that this mixed picture arises from the type of method used in these studies: (i) the classic block paradigm in satiation studies building on [8] contains only a very small number of items in each block and does not separate the test of whether exposure effects exist at all from the effect of grammatical construction; and (ii) tests for exposure effects with more items have been carried out either with non-standard components (e.g. a non-standard rating scale in the study reported in [6]), conflate the phenomenon of exposure effects with task-related effects like ordering effects that could instead be controlled for using individual counterbalancing (in

the studies reported in [18-19]), or do not use intermediate items as comparisons (in the study reported in [20]).

In this paper, we introduce a new method, the Standardised Block paradigm, that resolves these ambiguities, reliably induces exposure effects in rating data (against earlier conclusions that exposure effects do not exist [15]), and provides evidence against claims from the early 2000s that exposure effects selectively target certain grammatical constructions [8]. Specifically, the new method includes inbuilt comparisons to separate the question whether exposure effects exist from the question whether exposure effects depend on grammatical construction, and introduces a number of standardised components to ensure good design, such as individual counterbalancing of a high number of items to control for ordering effects [21-22].

In the rest of this introduction, we first lay out the properties of the classic satiation paradigm (section 1.1), highlighting the usefulness of blocks and considering which types of comparisons we need to test and falsify the effect of grammatical construction on exposure effects. We then introduce the phenomenon we take as test case in the present study: superiority effects in multiple wh-questions (section 1.2.).

Measuring repeated exposure effects

The classic block paradigm

How can we test whether ratings change across a single experimental session? The classic answer to this question is to take measurements at different points in an experiment and to compare them.

Early work [6] achieves these measures by running a rating experiment and then running it a second time immediately afterwards with the same participants. The means change between the two experiments, decreasing in well-formedness from an intermediate judgement to a lower intermediate judgement. These experiments show clear change in the intermediate zone of judgements, sometimes increasing, sometimes decreasing. However, determining the source of

this variability is impeded by some aspects of the design. For instance, filler items are crucial to providing context for judgements [2, 3, 15, 23], so are the changes just a result of a design without balanced filler items? Does the variability come from the type of rating task used and the type of scale used? Are the changes just a result of the type of target construction used? There is currently no clear answer to these questions in the literature.

Subsequent work to [6] revisits these changes with two new hypotheses [8]: (i) exposure effects selectively target certain constructions; and (ii) ratings for these constructions increase under repeated exposure until they approach the asymptote, i.e. they *satiate*. More recent work suggests that shifts in well-formedness do not necessarily imply satiation [20]. Therefore, in the rest of this paper we use *satiation effects* to refer specifically to previous studies in the literature, and otherwise use the term *exposure effects.*

To test the link between construction and exposure effect, the second wave of studies starting with [8] uses a block design [8, 10, 11, 12, 13, 14, 17]. Other studies introduce list as a factor into the linear model without using a block design [18,19,24] or with a block design [20]. Note that of the linear models, one study [18] tests for ordering effects post-hoc, and two studies [19, 24] do not use a block design or individual counterbalancing, meaning that these tests are testing ordering effects, not satiation effects per se.

Moreover, testing focuses on six types of ungrammatical constructions involving filler-gap dependencies. These constructions are illustrated in (1) and the gap is marked with (t) for 'trace'. In (1a), the argument *who* moves from within a want-for clause to the front of the sentence. In (1b), *who* is extracted from within a clause introduced by *whether*. In (1c), who is moved to the front of the sentence across *that.* In (1d), *what* is extracted from within a complex subject. In (1e), *who* is extracted from within a complex noun phrase. In (1f), *who* is extracted from within an adjunct, and in (1g), *how many* is extracted from within the noun phrase *how many books* without pied-piping the rest of its constituent.

1. Classic satiation constructions

   a. *Want-for*: *Who does John want for Mary to meet (t)?
   b. *Whether-island*: *Who does John wonder whether Mary likes (t)?
   c. *That-trace*: *Who does Mary think that (t) likes John?
   d. *Subject island*: *What does John know that a bottle of (t) fell on the floor?
   e. *Complex NP island*: *Who does Mary believe the claim that John likes (t)?
   f. *Adjunct island*: *Who did John talk with Mary after seeing (t)?
   g. *Left branch*: *How many did John buy (t) books?

[8: (2a), 576]

[8] uses the following steps in order to determine whether satiation has taken place: the number of "yes" responses is counted in the first two blocks and the last two blocks. If the number of "yes" responses in the final two blocks exceeds the number of "no" responses, a participant is deemed to have satiated. If the number of participants who satiated exceeds the number of participants who did not satiate, then the construction is deemed to satiate. Only participants who show change are included.

Mixed results across reported studies

Do the constructions in (1) show satiation? Table 1 summarises previous key results.

Table 1. Summary of satiation effects by previous study and construction

	Want-for	**Whether-island**	**That-trace**	**Subject island**	**Complex NP island**	**Adjunct island**	**Left branch extraction**
Braze (2002) [12]		**yes**				no	

Chaves and Dery (2014) [19]: exp. 1				**yes**			
Chaves and Dery (2014) [19]: exp. 2				**yes**			
Chaves and Dery (2019) [24]: exp. 1				**yes** (subject gap) no (object gap)			
Chaves and Dery (2019) [24]: exp. 3				**yes** (parasitic gap) **yes** (object gap) no (non-parasitic gap)			
Crawford (2011) [14]		**yes**		no		no	
Francom (2009) [11]: exp. 1	**yes**	**yes**	no	**yes**	no	no	no
Francom (2009) [11]: exp. 2			no	**yes**	no	no	no
Goodall (2011) [13]			no	**yes**	**yes**	**yes**	no
Hiramatsu (2000) [10]	**yes**	**yes**	**yes**	**yes**	no	no	no
Snyder (2000) [8]	no	**yes**	no	no	**yes**	no	no
Sprouse (2007a) [15]		no		no	no	no	
Sprouse (2007b) [16]: exp. 1	no	no	no	no	no	no	no
Sprouse (2007b) [16]: exp. 2	no	no	no	no	no	no	no
Sprouse (2007b) [16]: exp. 3	no	no	no	no	no	no	no
Sprouse (2009) [17]: exp. 1-5		no		no	no	no	
Sprouse (2009) [17]: exp. 6		no		no	no	no	no
Sprouse (2009) [17]: exp. 7		no			no	no	

The main observation to come out of Table 1 is how mixed previous results have been. The only consistent results are those for left branch extraction. Besides the classic constructions summarised in Table 1, some limited number of additional constructions have been tested in English (e.g. sentences with relative clauses and coordination in [16] and sentences with no inversion in [13], both showing no satiation). Additionally, some limited amount of cross-linguistic work has been undertaken. [13] reports satiation effects for sentences without

inversion and double psych fronting constructions in Spanish and no satiation effects for adjunct islands, subject islands, the complex NP constraint or left branch extraction. In Norwegian, satiation effects are reported for sentences with long movement, and no satiation for sentences with short movement or no movement at all or sentences with doubly-filled specifiers [25]. Sentences with crossing movement satiate in the first experiment but not in the second experiment [25].

Furthermore, even in studies that show no construction-specific satiation effects, researchers nonetheless report a high degree of individual variability. For instance, in the first experiment reported in [15], no satiation effects are reported in any constructions, but nonetheless a small percentage of participants are "satiaters" in all but left branch extraction: 9% in want-for constructions, 14% in whether-islands, 24% in *that*-trace islands, 24% in subject islands, 9% in complex NP islands, and 24% in adjunct islands. For comparison, the study in [15] also reports for the same experiment that 9% did not satiate in want-for constructions, 19% did not satiate in whether-islands, 14% did not satiate in *that*-trace islands, 9% did not satiate in subject islands, and 9% did not satiate in complex NP islands. In some constructions therefore, the quantity of satiating participants is actually higher than the quantity of non-satiating participants, e.g. in subject islands and sentences with *that*-trace effects. In all constructions, these results highlight how much of the variability and data is lost in the classic satiation paradigm [8, 15]: over 50% of participants in the first experiment in [15] showed no change, and therefore neither satiated nor failed to satiate.

Why are the results so mixed? According to one prominent hypothesis, the *Response Equalisation Strategy* [15], satiation effects simply result from task effects. Under this view, the reason that the results in the literature are so mixed is simply because some experiments use unbalanced designs and others do not. The study in [8] contained more ungrammatical fillers than grammatical fillers [15]. Therefore when participants were asked to make binary judgments about whether the sentence was well-formed, the overall set of stimuli would mean

that they were necessarily giving more 'no' responses than 'yes' responses. The Response Equalisation Strategy corrects for this by increasing the number of 'yes' responses across the experimental session, causing ratings to rise [17].

The Response Equalisation Strategy does not appear to wholly explain exposure effects. Results from an experiment with more grammatical than ungrammatical fillers show that participants do not lower their ratings if the unbalanced ratio of grammatical to ungrammatical fillers is inverted [19]. The one-way nature of the change in judgements indicates that exposure effects are not solely controlled by the desire to equalise "yes" and "no" responses [19].

A number of studies introduce balanced and unbalanced filler sets in order to balance the experiment. Fillers have been balanced between ungrammatical and grammatical [12, 15] experiments 1 to 5 in [17], the second experiment in [11, 14, 19, 24]. A number of further studies mix grammatical and ungrammatical fillers with either more grammatical than ungrammatical fillers (8, experiment 1 in 11, 13] or with more ungrammatical than grammatical fillers [16, 17].

Crucially however, in studies using balanced as well as unbalanced fillers, all the filler items fit into binary categories of grammatical and ungrammatical. None of the studies use intermediate fillers items. This choice of filler items does balance out the experiment in terms of the types of sentences that participants are presented with, but at the same time it also has an unintended negative consequence for testing the theory behind satiation effects: it means that the only intermediate items are the target constructions. Categorical sets of fillers therefore introduce a confound to the classic satiation paradigm: none of the previous studies allow exposure effects related to zone of well-formedness to be separated from construction-specific effects, simply because none of these studies since [8] include any intermediate items among the fillers. Therefore, satiation effects could be to do with construction, but they could also just be to do with zone of well-formedness.

Is it plausible that zone of well-formedness explains the distribution of satiation effects? One way to assess this hypothesis already with previous studies is to look at mean judgments. Unfortunately, many studies do not report means. One study that does report means fits with the zone of well-formedness explanation is illustrated in (2).

2. Means and satiation effects by construction in [20]
    a. Grammatical prepositional dative (dative-only verb): does not satiate, m=5.733s
    b. Garden path (unambiguous complement): does not satiate, m=5.443
    c. Garden path (ambiguous relative clause) : **satiates**, m=4.153
    d. Centre embedding (1 degree): **satiates**, m=3.853
    e. Ungrammatical double object construction (dative-only verb): **satiates**, m=3.817
    f. Grammatical binding: does not satiate, m=3.63
    g. Principle A binding violation: does not satiate, m=3.287
    h. Centre embedding (2 degrees): does not satiate, m=1.813

(2) shows that satiation effects also coincide with those constructions with intermediate means: satiation affects garden path constructions with relative clauses, centre embedding with one degree of embedding and ungrammatical double object constructions. All of these constructions involve means between 3.817 and 4.153. Constructions with both higher and lower means do not satiate. The authors in [15] attribute the changes they found to increased fluency of comprehension. We understand this hypothesis as fitting with a correlation between intermediate judgements and satiation.

The idea that exposure effects relate to intermediate judgments rather than natural syntactic classes is not new: already in the 80s, reports were made movement amongst intermediate judgments [6]. However, [8] dismisses this possibility, as does [13]. The initial claim in [8] came from asking an additional 10 subjects to take part in the same satiation experiment where the "yes"/"no" task was replaced with a rating scale. The mean ratings are given in (3), along with their status with regard to well-formedness.

3. Means from post-test (N=10) in [8]
    a. *Want-for*: no satiation (m=4.15)
    b. *Whether-island:* **satiation (m=3.35)**
    c. *That-trace:* no satiation (m=2.95)
    d. *Subject island:* no satiation (m=2.6)
    e. *Complex NP island:* **satiation (m=1.9)**
    f. *Adjunct island:* no satiation (m=1.65)
    g. *Left branch:* no satiation (m=1.05)

Snyder [8: 579] concludes from these means that "the satiability of a sentence type does not appear to correspond in any simple way to its initial well-formedness". [13] repeats and agrees with this assertion in [8], and claims that the results from the English experiment he reports agrees with the lack of correlation between well-formedness and satiation. Specifically, he points to two constructions that show similar initial percentages of "yes" responses in the first two blocks and different percentages of "yes" responses in the final two blocks (complex NP Constraint, 23.35% in the first two blocks compared to 41.1% in the final two blocks; and *that*-trace, 24.45% in the first two blocks compared to 24.45% in the final two blocks). However, the method to achieve these results was run post-hoc after a first analysis of the data using the [8] method and deviates from the standard method to evaluate satiation in [8]. More importantly, the two methods actually suggest different pictures: the results from the first, classic, method that counts the number of English-speaking subjects that satiate suggests that the complex NP constraint and *that*-trace conditions behave more alike than differently (the complex NP constraint has 10 satiaters, and the *that*-trace construction has 5 satiaters, compared to 4, 2, 1 and 0 for sentence types with no inversion, subject islands, adjunct islands and left branch extraction respectively). Under the second method, taking overall percentage of "yes" responses across participants, complex NP constraint shows a change between the initial and final blocks, whereas *that*-trace does not. What these results underline is not that the presence

of exposure effects does not correlate with well-formedness, but that a more reliable method is needed to test for exposure effects in rating data in the first place.

The main issue at the moment is that we do not have enough information to answer the question whether zone of well-formedness is sufficient to explain satiation effects. Part of the issue is a recurrent one in generative work in linguistics: experimental paradigms are not set up to disprove a hypothesis, meaning that they either provide evidence for a theoretical hypothesis or return a null effect.

In the case of repeated exposure effects, what we need to test the link between construction and increase in rating is a paradigm that (i) includes intermediate judgements as comparisons; and (ii) that keeps the block design. Keeping a block design allows principled points of measurement throughout the experiment to allow for a more principled way of controlling for individual variation. Such a paradigm can separate the test whether exposure effects exist from the test whether such effects target only certain constructions. We introduce such a method in this paper.

A new test case with high variability: superiority effects in multiple wh-constructions

What would be a good initial test case for such a paradigm? Ideally, we would want a construction that has not yet been tested, and that shows a high degree of variability from multiple sources. Superiority effects in multiple wh-questions [26] fit these requirements.

Current grammatical models attempt to derive the effects of the superiority condition from more general principles, such as the Minimal Link Condition [27]. Here we use Chomsky's original formulation for the superiority condition in (3) and examples of the construction are given in (4):

4. Superiority condition: definition

   a. No rule can involve *X*, *Y* in the structure . . . *X* . . . [$_{\alpha}$ . . . *Z* . . . –*WYZ* . . .] . . . where the rule applies ambiguously to *Z* and *Y* and *Z* is superior to *Y*

   [28: (73), 66]

   b. "the category *A* is "superior" to the category *B* in the phrase marker if every major category dominating *A* dominates *B* as well but not conversely"

   [28: 66]

5. Superiority condition: example

   a. John knows who saw what.

   b. *John knows what who saw ~~what~~.

   [28: (70), 66, formatting changed]

In the sentences in (3), the subject *who* is superior to (the object *what*, but the object *what* is not superior to the subject *who*. In (3a), the subject *who* and object *what* are both in their base order and the sentence is well-formed. (3b), in contrast, cannot be derived because the rule preposing wh-phrases cannot apply to *what* (=Y) because who (=Z) is superior to *what*.

Superiority effects show a particularly high degree of variability with multiple factors influencing the judgments (among them being the permissibility of *in situ* wh-subjects, the nature of the cues identifying grammatical functions and interference effects [29]) making them particularly suitable for this kind of study.

EXPERIMENT 1: RATING OF GERMAN SENTENCES

Experiment 1 tests exposure effects in German superiority constructions for multiple wh-questions using a new method. Specifically, we address the issues with the classic satiation paradigm in two ways.

First, we introduce two new types of comparisons to test whether any exposure effects result from grammatical construction. One type of comparison is between target items and standardised filler items. These filler items differ from balanced filler items used in previous studies in two ways: (i) they involve multiple levels of gradient rather than just acceptable versus unacceptable; and (ii) they are part of a standardised set that was developed and calibrated by an independent research group [3, 30]. The second type of comparison is a cross-linguistic comparison achieved by yoking the experiment across German and English (a discussion of the methodology of yoking experiments across languages with the example of superiority effects in German and English can be found in [31]).

Second, we bring the experiments in line with current standard psycholinguistic practice: we increase the quantity of items, and counterbalance these across blocks and participants, meaning that we can analyse the results using linear mixed models that take different sources of variability into account (unlike the studies in [8, 15]), whilst also controlling for confounding influences of ordering effects (unlike the studies in [18, 19]). This method provides a way of testing whether exposure effects exist at all that is independent of the construction used to test them.

Method

Subjects

A total of 55 students and employees of the University of Potsdam participated in the study. Only the results of German native speakers were included in the evaluation; two participants were excluded due to technical failures. Of the remaining 48 participants, 4 were male and 44

female and ranged between 19 and 40 years of age, with an average age of 24 years. Recruitment was carried out via flyers distributed on campus, internet advertisements on various platforms, and an e-mail distribution list from a university experimental laboratory. Participation was voluntary and was remunerated with study credit or eight euros. All participants were naive with regard to the questions and objectives of the study.

Apparatus

The experiment was conducted in the experimental psychology lab at the University of Potsdam. Three soundproof computer booths were used. The experiment was implemented using the Python software PsychoPy version 1.84.2 [32] and the stimulus material was presented on a computer screen; ratings of "1" to "7" of sentences were entered on a standard computer keyboard with a German keyboard layout.

Material

The material consisted of 120 target and 252 filler sentences. The target sentences represent indirect multiple wh-questions in subordinate clauses. Each target sentence was available in four conditions:

6. German target conditions

a. Condition 1: subject-initial; matching animacy (*wer-wen*)

Keinesfalls wusste die Haushälterin genau, **wer** bei der Gartenfeier **wen**
certainly.not knew the housekeeper exactly who by the garden.party who
ständig angesehen hat.
continuously looked.at had
*The housekeeper certainly did not know exactly **who** had kept on looking at **who(m)** at the garden party.*

b. Condition 2: subject-initial; mismatching animacy (*wer-was*)

Keinesfalls wusste die Haushälterin genau, **wer** bei der Gartenfeier **was**
certainly.not knew the housekeeper exactly who by the garden.party what
ständig angesehen hat.
continuously looked.at had
*The housekeeper certainly did not know exactly **who** had kept on looking at **what** at the garden party.*

c. Condition 3: object initial; matching animacy (*wen-wer*)

Keinesfalls wusste die Haushälterin genau, **wen** bei der Gartenfeier **wer**
certainly.not knew the housekeeper exactly who by the garden.party who
ständig angesehen hat.
continuously looked.at had
*The housekeeper certainly did not know exactly **who** had kept on looking at **who(m)** the garden party.*

d. Condition 4: object initial; mismatching animacy (*was-wer*)

Keinesfalls wusste die Haushälterin genau, **was** bei der Gartenfeier **wer**
certainly.not knew the housekeeper exactly what by the garden.party who
ständig angesehen hat.
continuously looked.at had
*The housekeeper certainly did not know exactly **who** had kept on looking at **what** at the garden party.*

A first factor manipulates the superiority/crossing movement variable: conditions 1 and 2 contain two wh-phrases and the subject precedes the object. These conditions therefore

illustrate instances where superiority is respected, and act as controls against which to compare the superiority violations. Conditions 3 and 4 contain the crucial superiority violations. In these conditions, the object wh-phrase crosses in front of the subject wh-phrase, meaning that superiority is violated.

A second factor manipulates an extra-grammatical feature, namely the animacy of the wh-phrases. Thus, subject and object match in animacy in two sentence types, namely (6a) and (6c), while the other two conditions, namely (6b) and (6d) have subjects and objects that do not match in animacy. We start here from the assumption that differences in subject and object increase the well-formedness of crossing movement [31, 33].

One half of the target sentences had, as in the example above, an adverb at the beginning of the main clause and the other half at the end of the main clause. In the subordinate clause there was an adverb after the second wh-word. We used adverbs because the goal of stimulus construction to create sentences that sound as natural as possible and it was certain that the use of adverbs would contribute significantly to this.

The distractor sentences represented six grammatical levels with 42 items each. For the first five gradations from (A) "almost not well-formed" to (E) "completely well-formed", we used calibration sets from [3] to create the fillers. That is, we started from single examples of each construction at each acceptability level (3 constructions x 5 levels) [3], and created 195 more fillers on the same template. We then added a sixth level (F) "uninterpretable and unacceptable" to provide a clearly ungrammatical level and better reflect the spectrum of grammaticality.

7. German filler examples (based on the Featherston fillers)

   a. Level A: Interpretable and highly acceptable

      In der Mensa essen viele Studenten zu Mittag.

      In the canteen eat.3PL many students to lunchtime

      Many students have lunch in the canteen.

   b. Level B: Interpretable but less acceptable than (A)

      Der Kaiser hat dem Fürsten den Maler empfohlen.

      The emperor has the prince the artist recommended

      The emperor recommended the artist to the prince.

   c. Level C: Interpretable but less acceptable than (B)

      Ich habe dem Kunden sich selbst im Spiegel gezeigt.

      I have the client himself.REFL himself in.the mirror shown

      I showed the client himself in the mirror.

   d. Level D: Interpretable but less acceptable than (C)

      Der Komponist hat dem neuen italienischen Tenor es zugemutet.

      the composer has the new Italian tenor it expected.of

      The composer expected it of the new Italian tenor.

   e. Level E: Interpretable but less acceptable than (D)

      Der Waffenhändler glaubt er, dass den Politiker bestochen hat.

      The arms.dealer believes he that the politician bribed has

      It is the arms dealer that he believes bribed the politician

   f. Level F: Uninterpretable and unacceptable

      Die Tinte wurde für vergossen.

      the ink was for spilled

Note that in (7), we follow the convention of providing literal translations for individual items in glosses, followed by the closest meaningful translation in English on a separate line (here in italics). Therefore, although the translations for filler levels (A) to (E) are all fully acceptable and well-formed in English, the German examples themselves from (B) through (E) are not fully acceptable: they decrease in well-formedness between the highest level (A) and the lowest level (F). The lowest filler (F) is intended to be uninterpretable, meaning that there is no meaningful way for all the words to be integrated into the interpretation of the whole sentence,

and therefore no full translation of the whole sentence to English. For filler (F), we therefore provide only literal translations for individual items in the gloss, and leave the translation blank.

Design and counterbalancing

The experiment comprised six blocks of 372 items. Each subject saw all target sentences once, an equal number in each of the four conditions, and also each of the filler sentences. Each block consisted of 62 items (4 x 12 target sentences and 24 filler sentences). The counterbalancing scheme also ensured that all items were seen equally often across the six blocks – a constraint imperative for the examination of satiation experiments, and particularly important when the number of items per block is higher than 1, as in Snyder's original paradigm. Also target sentences were seen equally often in the four conditions. Transitions between the four experimental conditions and the six types of fillers occurred equally often, that is we implemented a Williams design. Finally, the item sequence was subject to the following constraints: (a) no immediate repetition of target sentences (i.e., target sentences were bracketed by at least one filler sentence), (b) no immediate repetition of the same experimental condition, and (c) no immediate repetition of the same type of filler sentence. This design requires a multiple of 24 subjects and determined the total number of 48 subjects.

Procedure

After providing informed consent and collection of demographic information, subjects were instructed in the well-formedness rating procedure. Specifically, they were asked to judge the well-formedness of the sentences according to spoken—not written—language and to use the full spectrum of the seven-point scale. They practiced the procedure with nine examples spanning the scale of well-formedness, including two of the uninterpretable fillers (filler level F). Then, each subject worked through the six blocks of sentences; durations of breaks between blocks was under the subjects' control.

Statistical analysis

For data aggregation, graphics, and inferential statistics, the open source software R [34] was used, and especially package *lme4* [35] for linear mixed models (LMMs) as well as packages *tidyverse* [36], *cowplot* [37], *sjPlot* [38], and *broom.mixed* [39] for pre-processing of data and post-processing and visualization of results.

Target sentences

The primary LMM included subjects and items as crossed-random factors and subject- vs. object-initial word order (2), animate vs. inanimate objects (2), and six blocks of 42 trials as fixed factors. The three factors were varied within-subject and within-items ensuring that all sentence frames were rated equally often in the 24 conditions. Across blocks we expected satiation of well-formedness ratings, that is ratings would increase and eventually reach an asymptote. Therefore, a Helmert contrast was specified for block to capture the point at which the rating did no longer change. To this end the first contrast tested block 1 against the average of blocks 2 to 6, the second contrast the second block against the average of blocks 3 to 6, and so on.

Within-subject or within-item factors give rise to variance components (VCs) and correlation parameters (CPs) of the mean rating and of experimental effects. We selected a complex mixed model following a strategy outlined in [40]; see also [41]), leaving out contrasts for blocks beyond the first one. Specifically, we started with an LMM including the VCs and CPs for Grand Mean (GM), first-block contrast (*b1*), contrast for subject vs. object initial word order (*so*), contrast for animate vs. inanimate objects (*an*), and all their interactions. None of the estimates appeared at the boundary (although some of the VCs were quite small). Forcing CPs to zero (*zcpLMM*), led to a significant decrease in goodness of fit. We stayed with the complex LMM (*cmplxLMM*); note that fixed effects which are the focus here did not differ for *cmplxLMM* and *zcpLMM*.

Target in the context of filler sentences

A second, post-hoc LMM included ratings of both target and filler sentences. We specified a block (first block vs. average of blocks 2 to 6) x type of sentence (10) design; type of sentence comprised four types of target and six types of filler sentences. Sentence types were assigned to levels of a factor according to their overall mean rating of well-formedness. Then sequential-difference contrasts across these levels were specified for the factor and interactions between block and nine contrasts; five of these nine block x type-of-sentences contrasts tested the null hypotheses that increase in well-formedness ratings from block 1 to block 2 did not differ (C1 to C5 [21].

As the six types of filler sentences varied between items, only two of the nine contrasts for sentence types and their interactions with block yielded estimates for VCs and CPs; all nine sentence types varied within subjects. Forcing CPs to zero did not lead to a significant decrease in goodness of fit according to AIC and BIC criteria. The *zcpLMM* was still slightly overparameterized. VCs estimated at the zero boundary were removed in a third step. We report estimates for this parsimonious LMM (*prsmLMM*). Due to their complexity we fit these models with the JuliaStats/MixedModels.jl package [42].

Given the large number of subjects, items, and observations, the usual t-distribution approximates the normal distribution. Therefore, we report test-statistics (estimate / standard error) as z-values and interpret absolute values larger than 2 as significant.

Results and Discussion

Target sentences

The first analysis focuses on the target sentences involving multiple questions. The four types of sentences using constructions of (a)"wer-wen", (b) "wer-was", (c) "wen-wer", and (d) "wen-was" for the same sentence frames map onto a 2 x 2 design with the main effects subject-vs-object initial sentences (so; a+b-c-d) and animacy of object (an; a+c-b-d) as well as the interaction between these two effects (a-b-c+d). In addition, this 2 x 2 design was repeated

across six blocks of trials. Figure 1a displays the change in ratings of well-formedness for the four types of target sentences shown in red; the figure also shows performance for the types of filler sentences, but we initially focus only on the target sentences. Supplement Table S1 contains the LMM fixed-effect estimates and statistics.

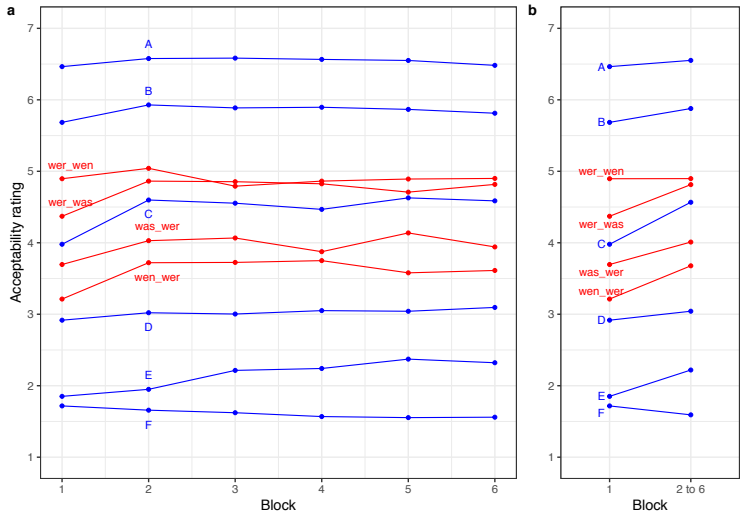

Figure 1. Well-formedness ratings (a) by block and (b) for first block and mean of later blocks for German target and filler sentences.

As expected, the two subject-initial sentences were rated as more acceptable than the two object-initial ones, yielding a significant main effect of order (b=0.52, z=8.97). Overall, there was no significant difference between animate and inanimate objects (b=-0.05, z=-1.62), but there was a significant interaction with order (b=0.13, z=5.00): there was a very clear preference for the inanimate "was-wer" construction over the animate "wen-wer" construction for object-initial sentences, whereas – at least in the first block-- the reverse preference held for subject-initial sentences.

Indeed, the pattern of change across blocks followed a very simple structure: overall, there was a significant increase in ratings from the first to the average of blocks 2 to 6 (b=0.31, z=3.24), and this increase was different for the interaction of order and animacy (b=-0.15, z=-2.60): As is clearly visible in the figure, ratings increased less strongly for "wer-wen" and more strongly for "wer-was" constructions than the other three conditions. None of the other contrasts defined for the change in ratings across blocks, nor – with one exception – none of the other interaction terms involving these contrasts were significant (all z-values < 2.00). The exception is a marked violation of parallelism between block 4 and 5: only the "was-wer" construction exhibited in increase in rating (b=0.10, z=2.34). We consider this interaction as spurious.

There are three main results. First, there is clear evidence for large differences in the judgement of well-formedness of multiple questions with a clear preference for subject-initial than object-initial sentences, in line with previous experiments [3, 31, 33] and at the same time there is no reliable evidence that these preferences changed much after one block of trials.

Second, for object-initial sentences, mismatched animacy is preferred to matched animacy. The condition with mismatched animacy is almost on par with subject-initial conditions, in line with the findings in [31, 33]. It is only in the condition with matched animacy that a superiority effect can be seen, and even here the condition is more acceptable than three of the filler levels (including filler levels D and E that are interpretable), suggesting that the sentence is not categorically ill-formed. This finding is consistent with the claim that German has no superiority condition in the grammar and shows only selective superiority effects in cases where processing difficulty is increased [29]. There is also a local ambiguity for the inanimate "was" which can serve both as subject or object of the sentence whereas the animate "wen" can only be interpreted as an object (i.e., it is unambiguously morphologically marked). One could assume that ambiguity leads to greater processing difficulty and therefore to reduced well-formedness. However, our results show that presumably the default expectation of a subject-initial sentence when encountering "was" confers an well-formedness advantage even if the initial interpretation as subject of the sentence must be revised later.

Finally, there is a general increase in well-formedness from the first to the second block of trials and this increase was stronger for "wer-was" constructions than the other three. Thus, it appears that the difference in well-formedness with an animacy mismatch in subject-initial sentences can be overcome with modest exposure with the mismatch condition.

Target sentences in the context of filler sentences

Is the increase in well-formedness from block 1 to block 2 related to sentence type – and therefore to natural syntactic classes - or is it rather a reflection of the general level of well-formedness? We use the ratings of filler sentences to address this question. As for target sentences, there was no interpretable change in well-formedness ratings from block 2 to block 6 for the types of filler sentences relevant for the comparison with targets. Therefore, we averaged the performance of the final five blocks for the ten types of sentences. Figure 1b exhibits the corresponding pattern of mean changes between block 1 and block 2.

The four types of target sentences were presented together with six different types of fillers A to E which were expected to cover the spectrum from clearly grammatical (A and B) to increasingly ungrammatical (C to F). Experience with "unusual" grammatical target sentences that are licensed by the grammar (i.e., object-initial target sentences) should lead to a larger gain in ratings of well-formedness than typical grammatical filler sentences like A and B (i.e., the latter should already be rated at a very high level – there is no or only little room for improvement) and also to a larger gain than for ungrammatical sentences like C to F (i.e., they should be rated at a low level and stay there because they are not licensed by the grammar).

Critical tests refer to interactions between block and neighboring sentence types. The contrast for the two subject-initial targets illustrates interactions in support of grammatical activation: "wer_wen" is rated higher than "wer_was" (b=0.30, z=4.26) and shows less of a gain between blocks (b=-0.44, z=-3.65). However, in the middle range of well-formedness ratings there is no

evidence for differential gains: “wer_was” is rated higher than filler type C (b=0.32, z=2.42), but there is no significant interaction with block (b=-0.15, z=-1.09. Filler type C is rated higher than the object-initial target sentences “was_wer” (b=0.42, z=2.91), but exhibits a small, but significantly larger gain (b=0.27, z=2.12). For the two object-initial targets we observe higher rating (b=0.40, z=4.24) for “was_wer” than “was_wen”, but no significant difference in gain (b=-0.15, z=-1.16). There is one ambiguity: “wen_wer” is rated significantly higher than filler type D (b=0.46, z=2.70) and significantly benefitted from exposure whereas filler type D stayed at the low level of acceptability (b=0.34, z=2.54). Possibly, D represents a “hard-core” ungrammatical filler type from the beginning. In general, the results do not allow us to reject the hypothesis that initial gains are due to the general level of well-formedness, but there is some ambiguity for low levels of well-formedness.

In summary, the German data shows that (i) regardless of exposure effects, the strength of superiority effects can be modulated by manipulating animacy (here we replicate previous findings in the literature, e.g. [31]); (ii) exposure effects can be reliably induced experimentally; and (iii) exposure effects may be a property of intermediate judgements rather than of certain types of syntactic constructions.

EXPERIMENT 2: RATING OF ENGLISH SENTENCES

For German, many authors assume that object-initial multiple questions are indeed grammatical because the superiority condition (or the more general constraints implying it) can be circumvented or fail to apply because of peculiarities of German sentence structure. When crossing movement is less acceptable, such differences in intuitions are taken to result from grammar-external factors such as increased processing complexity (e.g. [29]). Do patterns of exposure effects change in a language that has a grammatical superiority condition? In this section, we report a parallel (yoked) study to the German one in English, where it is still controversial whether superiority violations are ungrammatical or not (e.g. [18, 31] for recent discussion). We test whether we find the same or a different pattern of exposure effects to German.

Instead of varying the animacy of the wh-phrases, we instead varied how specific the wh-phrases were by using discourse-linking, a grammatical device that picks out a specific referent in the wider discourse context. In the discourse-linking factor, simple wh-phrases like *what* or *who* contrast with more complex wh-phrases with specific referents like *which book* or *which person*. Discourse-linking in English parallels animacy in German for two reasons. First, animacy does not influence the strength of superiority violations in English, meaning that this factor is not an appropriate choice for modulating island well-formedness [31]. Second it is known that discourse-linking changes the well-formedness of island constructions, including the well-formedness of multiple wh-questions by eliminating penalties for crossing movement [43, 44], cf. also [3] for experimental evidence supporting this assumption in the generative literature.

Method

Subjects

A total of 48 people participated in the study. Fourteen self-identified as male, 32 self-identified as female and 2 subjects selected “other” under gender. Ages ranged between 18 and 48, with an average age of 31 years. Recruitment was carried out through the web-based recruitment platform, Prolific. Participation was voluntary and was remunerated with £6.80. All participants were naive with regard to the questions and objectives of the study.

Apparatus

The experiment was conducted using Ibex software on the web-based Ibex Farm server (https://spellout.net/ibexfarm, developed by Alex Drummond). For the lab-based study, we had generated distinct questionnaires for each participant. To retain distinct counterbalancing in this web-based version, we created 48 distinct questionnaires with unique links. Participants first

clicked on a welcome page and then received a unique link, in randomized order. Clicking on the link led them to the questionnaire. The stimulus material was presented on a computer screen; ratings of "1" to "7" of sentences could either be entered on a keyboard or by pointing and clicking the mouse on the screen.

Material

As in the German study, material consisted of 120 target and 252 filler sentences. The target sentences themselves were translations of the German material with some adjustments. Instead of animacy we implemented discourse linking as a second factor. We also left out adverbs because including adverbs in English did not make the sentences sound more natural, and changed the content nouns in some sentences.

As for German sentences, target sentences represent indirect multiple wh-questions in subordinate clauses. Again, each sentence was available in four conditions:

8. English target conditions
    a. Condition 1: subject-initial; non-discourse-linked object: who – what
        *The housekeeper forgot **who** had dropped **what** during the party.*
    b. Condition 2: subject-initial; discourse-linked object: which N_{subj} – which N_{obj}
        *The housekeeper forgot **which guest** had dropped **which glass** during the party.*
    c. Condition 3: object-initial; non-discourse-linked object: what – who
        *The housekeeper forgot **what who** had dropped during the party.*
    d. Condition 4: object-initial; discourse-linked object: which N_{obj} – which N_{subj}
        *The housekeeper forgot **which glass which guest** had dropped during the party.*

The first factor 'superiority' can be seen by comparing (8a-b) with (8c-d). In the examples in (8a) and (8b), the subject appears before the object and thus fulfills the superiority condition. Examples (8c) and (8d) violate the superiority condition because the wh-object crosses in front of the wh-subject. The second factor of discourse-linking can be seen by comparing (8a) and (8c) with (8b) and (8d). In (8a) and (8c), the subject and object are the non-discourse-linked wh-words *who* and *what*, whereas in (8b) and (8d) the subject and object are the discourse-linked DPs *which guest* and *which glass*.

The distractor sentences represented six levels of well-formedness, as with German. Each level of well-formedness was made up of 42 items. The top five levels were each made up of three constructions that were identified in [30] as consistently rating at that gradient level. To create the fillers, we used the three example items from [30] for each level and created additional items until we reached the desired number of 42. We then added an additional sixth gradation (F) "not at all well-formed" to better reflect the lower end of the spectrum of grammaticality (see Supplement for the complete list of target and filler sentences).

9. English filler examples (based on Featherston fillers)
    a. Level A:
        The patient fooled the dentist by pretending to be in pain.
        There is a statue in the middle of the square.
        The winter is very harsh in the North.
    b. Level B:
        Before every lesson the teacher must prepare their materials.
        Jane does not boast about her being elected president.
        Jane cleaned her motorbike with which cleaning cloth?
    c. Level C:
        Hannah hates but Linda loves eating popcorn in the cinema.
        Most people like very much a cup of tea in the morning.
        The striker must have fouled deliberately the goalkeeper.
    d. Level D:
        Who did she whisper that had unfairly condemned the prisoner?
        The old fisherman took her pipe out of mouth and began story.

Which professor did you claim that the student really admires her?

e. Level E:

Historians wondering what cause is disappear civilisation.

Old man he work garden grow many flower and vegetable.

Student must read much book for they become clever.

f. Level F:

The ink was for spilled.

For the construction we used native speaker intuitions of naturalness from one of the investigators as the guiding principle. We again based the filler items on calibrated sets of distractor items developed by an independent research group, outlined for English in [30]. As in German, we started from the one example per construction cited in the paper for each of the top 5 levels of well-formedness, created further items on that same template, and then created a sixth uninterpretable filler level to reach a total of 252 filler items.

Design and counterbalancing

Counterbalancing was identical to the German study. In order to implement the individualized counterbalancing scheme over the internet, we created unique questionnaires for each participant ID, and assigned participants to IDs using a random link generator in Ibex Farm.

Procedure

Subjects were instructed to judge the well-formedness of the sentences according to their judgements of spoken language, rather than written language that they might find in a textbook, and to use the full spectrum of the seven-point scale. They practiced the procedure with ten examples. Then, each subject worked through the six blocks of sentences; durations of breaks between blocks was under the subjects' control.

Statistical analysis

Statistical analysis for the English study followed the same procedure as in Experiment 1, except that the order of levels for type of sentence was different and consequently also the contrasts resulting from the application of the sequential-difference contrast to this factor.

Results and Discussion

Target sentences

The four types of sentences using constructions of (a)"who-what", (b) "which N_{subj}-which N_{obj}", (c) "what-who", and (d) "which N_{obj}-which N_{subj}" for the same sentence frames map onto a 2 x 2 design with the main effects subject-vs-object initial sentences (so; a+b-c-d) and discourse linking of subject and object (dlink; a+c-b-d) as well as the interaction between these two effects (a-b-c+d). In addition, this 2 x 2 design was repeated across six blocks of trials. Figure 2a displays the change in ratings of well-formedness for the four types of target sentences shown in red; the figure also shows performance for the types of filler sentences. Supplement Table S2 contains the LMM fixed-effect estimates.

Here again, the two subject-initial sentences were rated as more acceptable than the two object-initial ones, yielding a significant main effect of order (b=0.67, z=12.25). Overall, there was also a significant difference between sentences with discourse-linked wh-phrases to sentences without discourse-linked wh-phrases (b=-0.35, z=-6.76) qualified by a significant interaction with order (b=0.43, z=9.49): Subject-initial word order was clearly preferred with or without discourse-linking to the other two conditions; for object-initial word order there was a clear preference for the discourse-linked construction ("which N_{obj} – which N_{subj}").

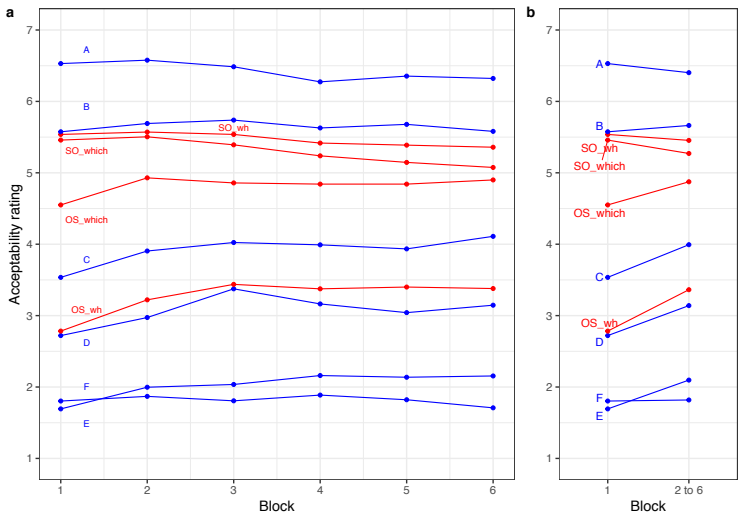

Figure 2. Well-formedness ratings (a) by block and (b) for first block and mean of later blocks for English target and filler sentences (web experiment).

The change in ratings across blocks showed an overall profile that was less clear than for German sentences: There was a nominal, but not significant increase from block 1 to the average of the following blocks (b=0.16, z=1.77) and, counter to expectations, a significant overall negative difference between block 3 and the final three blocks (b=-0.11, z=-2.47). However, there were two significant interactions between the first and second contrast for block with order. There was a significant increase in well-formedness from the first block to the average of the others for object-initial but not for subject-initial sentences (b=-0.29, z=-4.95) and a weaker effect for the change from the second to the average of the rest (b=-.14, z=-.3.17). The second interaction was "helped" to some degree by a small reduction of well-formedness of subject-initial sentences in the final blocks. Finally, there was also a significant interaction for the second contrast of block with discourse linking due to an increase in well-formedness for the non-discoursed linked sentences ("what-who"; b=0.10, z=2.28). These are the sentences with the lowest well-formedness of the four target sentences.

Target sentences in the context of filler sentences

As for German sentences, it is instructive to examine the well-formedness of English target sentences in the context of filler sentences (see Figure 2). Unlike for German sentences we do not see the prototype pattern for the two subject-initial target sentences; indeed, there is no evidence for differences in rating (b=0.14, z=1.23) and they do not show evidence for a differential change in rating (b=0.10, z=0.81); also filler type B is statistically not distinguishable from SO_wh (both z < 1.).

In contrast, the two types of object-initial target sentences clearly increased from block 1 to the average of the other blocks, but this increase was similar to filler sentences of type C and D, respectively. In the post-hoc LMM, the lines for these four sentence types were statistically parallel that is there was no interaction with block for the three pairwise contrasts (all |z-values| < 1.34). Thus, for English sentences in the middle range of well-formedness, we cannot rule out that the change associated with object-initial target sentences is simply reflecting the general level of well-formedness exhibited by filler sentences of comparable well-formedness. Unlike for German sentences, there is no ambiguity in this respect even for sentences of filler type D.

In summary, as for German target sentences, there is a large difference in the judgement of well-formedness of multiple wh-questions with a clear preference for subject-initial over object-initial sentences, in line with previous experiments [15, 18]. For object-initial sentences there was a very clear preference for discourse-linked than not-discourse-linked sentences. There was positive initial change for the well-formedness of object-initial sentences, but this change might simply reflect a general middle-raise in well-formedness also observed for filler sentences with a similar initial rating of well-formedness.

EXPERIMENT 3: Rating of English sentences (replication in lab)

In Experiment 3, we report a replication of the English experiment from section 3. Time-wise, this experiment was carried out after Experiment 1 and before the web-based Experiment 2.

We chose to run the Experiment 2 because the counterbalancing scheme was not rendered as intended in Experiment 3. Although the target items were not counterbalanced as intended across blocks in Experiment 3, some aspects of the counterbalancing were preserved: each participant still saw an individual questionnaire, as well as more than one target item per block. Fillers were counterbalanced across blocks as intended, and ordering restrictions such as making sure that target items were not adjacent were maintained. We submit that past studies counterbalanced target conditions, but, as far as we could tell from procedural descriptions, earlier studies did not counterbalance conditions across blocks either.

Most importantly, one important result of Experiments 1 and 2 was that the change in well-formedness between initial blocks of trials for target sentences might simply reflect their level of well-formedness. This argument was based on the absence of evidence for differences in change when compared to ungrammatical sentences with comparable ratings of well-formedness. The argument rests on arguing a null hypothesis of parallel changes. Such results are in need of replication.

Method

Subjects

A total of 48 subjects (28 female, 20 male) participated in the study. Ages ranged between 16 and 51, with an average age of 23 years. Recruitment was carried out through the participant pool (SONA) at University College London (Psychology and Language Sciences) and the University of Cambridge (Language Sciences). Participation was voluntary and was remunerated with £6. All participants were naive with regard to the questions and objectives of the study.

Apparatus

The experiment was conducted in the Speech and Language Sciences Lab at University College London and at Trinity Hall, Cambridge. The experiment was implemented using the Python software PsychoPy [32] version 1.84.2 and the stimulus material was presented on a computer

screen; ratings of "1" to "7" of sentences were entered on a standard computer keyboard with a UK keyboard layout.

Material

Material was identical to Experiment 2.

Design and counterbalancing

We did not counterbalance the target items across blocks in this study, due to an error in generating the questionnaires. Some of the features of the counterbalancing scheme used in Experiment 1 and Experiment 2 were nonetheless retained, such as different questionnaires for each participant, and counterbalancing of filler items across blocks.

Procedure and statistical analysis

Subjects were instructed to judge the well-formedness of the sentences according to their judgements of spoken language, rather than written language that they might find in a textbook, and to use the full spectrum of the seven-point scale. They practiced the procedure with ten examples. Then, each subject worked through the six blocks of sentences; durations of breaks between blocks was under the subjects' control. Statistical analysis for Experiment 3 followed the procedure for Experiment 2.

Results and Discussion

Target sentences

Figure 3a displays the change in ratings of well-formedness for the four types of English target sentences shown in red and the six types of filler sentences in blue. Supplement Table S3 contains the LMM fixed-effect estimates and statistics. Main effects of order (b=.74, z=12.93) and discourse-linking (b=-0.45, z=-8.87) as well as the interaction of these two factors (b=0.56, z=10.97) were replicated in the lab experiment: again, subject-initial word order was clearly preferred with or without discourse-linking to the other two conditions; for object-initial word

order there was a clear preference for the discourse-linked construction ("which N_{obj} – which N_{subj}").

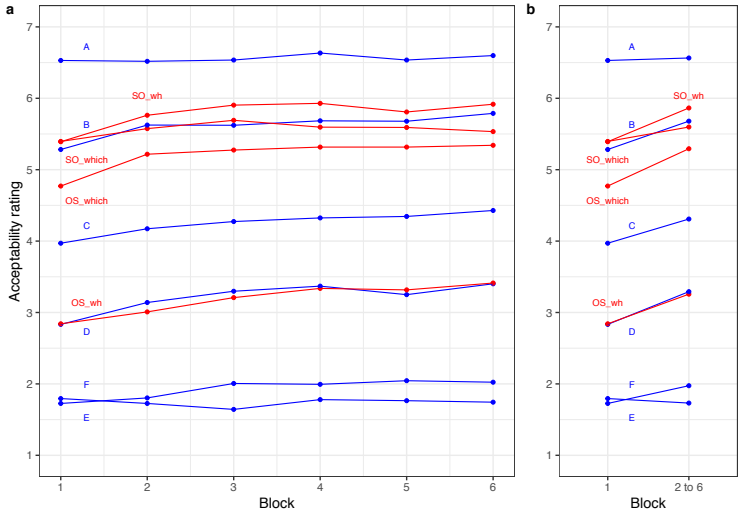

Figure 3. Well-formedness ratings (a) by block and (b) for first block and mean of later blocks for English target and filler sentences (lab experiment).

The change in ratings across blocks was much clearer than in Experiment 2: The first two block contrasts were significant (b1: b=0.39, z=4.54; b2: b=0.14, z=3.53) indicating an overall increase of well-formedness from the first to the average of the rest and a second, smaller increase from block 2 to the rest. Again theres were two significant interactions between the first and second contrast for block with order. There was a significant increase in well-formedness from the first block to the average of the others for object-initial but not for subject-initial sentences (b=-0.13, z=-2.24) and for the change from the second to the average of the rest (b=-.12, z=-.3.09). As in Experiment 2, the increase in well-formedness across the initial blocks was larger for object-initial than subject-initial target sentences (see Figure 3b). Different from Experiment 2 there was also an increase for subject-initial target sentences.

Target sentences in the context of filler sentences

The final question is whether the different increases for subject- and object-initial sentences are different than those of filler sentences of similar well-formedness. The results replicate Experiment 2. We have no evidence for differential change from block 1 to the average of the other five blocks for the three pairwise comparisons of the two object-initial target-sentence types and the two filler-sentence types C and D (z-values for three block x contrast interactions: 1.85, -0.68, and -0.25). Thus, again, we cannot rule out that the change across initial blocks is related to the general level of well-formedness exhibited by fillers of comparable well-formedness.

It is quite interesting that in Experiment 2 filler type B aligned with the absence of an initial raise of well-formedness of subject-initial sentences and in Experiment 3 it aligned with their increase. Moreover, aside from the core hypotheses relating to target sentences, the experiment also replicated the profiles for filler types A, E, and F. We also note that the unexpected negative effect and the interaction for the second contrast of block with discourse linking reported for Experiment 2 did not replicate.

GENERAL DISCUSSION

Summary of combined results

We found parallel patterns between German and English for both exposure effects, and modulating the strength of effects of crossing movement by factors related to semantics and discourse. We found:

g. A shift in initial block(s) only

h. A shift in intermediate zone irrespective of sentence type, i.e. shifts were found (i) in target conditions that respected superiority; (ii) target conditions that violated superiority; and (iii) filler levels with no wh-elements at all

i. Comparable shifts in German and English, both in initial blocks and in intermediate zones of well-formedness
j. Effects of D-linking in English that were on par with animacy effects in German, specifically:
    i. a decrease for German with superiority violations with like animacy that was comparable to regular (non-d-linked) superiority violations in English
    ii. an increase for English with superiority violations where both wh-phrases were D-linked that was comparable to acceptable (mismatched animacy) superiority violations in German

Satiation effects track zone of well-formedness, not grammatical construction

Our experiments sought to establish whether there are "satiation effects" in the sense of studies building on [8] caused by repeated exposure, and whether these effects are zone- or construction-sensitive. Are the shifts in well-formedness that we found "satiation effects"? Certainly, our results show satiation effects in the sense of repeated exposure effects that are induced experimentally. These effects are clearly visible between the first and the remaining blocks of the experiments. Our experiments provide strong evidence for an increase, resulting from the systematic increase in the number of items presented. If we look at the number of exposures per condition in our experiment, we have 12 per block, i.e. 48 target items per block. Previous studies vary between 5 and 7 blocks, each with 1 instance of the crucial constructions (some studies contain multiple islands (up to 6 in total), and the Hofmeister et al. study had a larger number (20) of items, but still less than in our first block). Our results are best thought of then as zooming out from the results in the literature: the results in the literature on the other hand are zooming in just on the initial part of the experiment (the first two blocks).

Demonstrating the existence of satiation effects leads to a number of further questions, some of which can be answered on the basis of our results. First, one would like to know if repeated

exposure eliminates the perception of non-well-formedness for certain constructions and for certain grammatical problems, or whether the penalty for violating certain laws of language is merely mitigated. The first interpretation may suggest itself if a certain violation is thought to be a result of the type of processing difficulty proposed in [18] that should no longer exist after a sufficient amount of exposure. In contrast, the second interpretation would be a natural consequence of the view developed in [20] for instance, according to which satiation effects result from improved comprehension strategies (while the perception of grammatical problems would be left intact).

Our results clearly go against the expectation that satiation effects show continuous rising, and that ratings finally reach some level of more or less full well-formedness in an asymptotic fashion. Our experiments clearly show an initial early increase followed by stability. Indeed, the most remarkable feature of the plots of judgments across blocks is the stability of the judgments across the experiment after the first block. Constructions do not change their well-formedness rating dramatically either. Constructions increase only by roughly .5 on a 7-point scale, meaning that repeated exposure does not change the quality of the judgment. This pattern can be detected in the experiments reported in the literature as well.

The stability we observe in our experiments could either be due to predominantly stable judgements, or to participants developing a rating pattern in the first two blocks that participants remember for the remainder of the experiment (entrenchment) - participants could have stopped deliberately rating the item they are presented with and instead produced memorised values to assign to types of sentences. There is currently no accepted way to control for memorising and entrenchment. Some limited discussion in previous works like [16] presents a potential argument that memorising and entrenchment undermine a balanced design and the use of sets of items of the same type, as is the case with control items and fillers. In this paper, we explored a way to identify entrenchment by looking at where response times stabilise. This technique is made possible specifically by the increased number of contrasts made available in the

Standardised Block paradigm. From a visual examination of the log-transformed response times, the time it took participants to respond stabilised at block 4, suggesting that this is the point at which responses became entrenched. Notably, response times stabilised after ratings stabilised (block 4 compared to block 2), suggesting that exposure effects increase but do not fully satiate. One could also speculate that once comprehension strategies of the sort envisaged by [20] have been developed there is simply no further tool available for further improvement if satiation indeed reflects such enhanced interpretation. Alternatively, participants may simply get more used to properly executing the unnatural task of assigning some numerical value from 1-7 to a sentence, because it involves identifying the range of constructions presented and their relative well-formedness. If participants are first reluctant to assign good ratings to any but the unquestionably acceptable items, some sort of 'satiation' effect would be expected.

It is interesting to note that one filler class behaves unusually in both its exposure pattern and in the type of judgment it receives: filler E, i.e. the lowest of the Featherston filler items and the second lowest of the fillers included. Speculatively, this may be because the lowest filler (filler F) completely rules out a parse of any kind, whereas filler E, although highly unacceptable, can still be parsed once the relevant errors are identified. It may be that participants are able to construct a parse that yields some kind of meaning over the course of the experiment for filler E, but cannot do this at all for filler F. This rising pattern in filler E is more like the classic expectations of a satiation effect.

The hope in some of the generative literature [8] that repeated exposure may make questionable sentences (nearly) completely acceptable may be unjustified in any event. Judgments are formed on the basis of various factors, among them the grammatical status of the construction in a narrow sense, the frequency of the construction, the ease with which the structure can be parsed, the relative quality of the construction as compared to grammatical alternatives, both in terms of what is offered in the language and what is present in the experimental material, the ease of computing the proper prosody, and the fit of the sentence into a conversational context

that is usually not presented in experiments. For obvious reasons, repeated exposure may enhance processability, but cannot for instance repair problems that result from the fact that a particular construction is bound to a particular conversational context that is not presented in the experiment. Thus, the processing difficulty possibly linked to *which book did which person read*? may have been eliminated in our experiment. However, the (unsatisfied) need for motivating the particular sorting key for answers (sorting them by books rather than people in the above example) that triggers object-initial sentences is not eliminated by repeated exposure.

A further question is whether the exposure effects are construction-specific or simply relate to a particular zone in the judgment space. Interest in satiation effects in the linguistic literature typically stems from the hope that specific constructional properties make a certain sentence pattern particularly prone to showing satiation effects. Relevant proposals relate to specific grammatical theories (attributing a particularly "unstable" status to constraints against crossing movement or movement out of grammatical islands) or address the grammar vs. processing divide [18].

Our results do not support construction-specific models of the weak-strong island divide at all, because, as we have shown, the satiation effect characterises all constructions within a certain range of judgments, irrespective of the precise choice of construction.

Our results for German and English demonstrate the link between satiation effects and range of judgment in two slightly different ways. German shows improvement of judgments not only for the two instances of crossing movement in multiples questions, but also for the grammatically perfect subject initial questions with inanimate objects and for filler items C which involve slightly unusual binding options and grammatically illicit order of dative and accusative objects. So, we observe upward movements of well-formedness for items that may have an unusual object choice, crossing movement, and an ordering constraint. There is no obvious common factor to these.

The situation in English is even more interesting. Crossing movement improves, but so do violations of verb-object adjacency, of the immobility of subjects following complementizers, and of gender agreement for pronouns. There is no obvious common factor to these, either. Interestingly, these effects pertain to "harder" grammatical constraints than what we see in German. What unifies all these constructions is simply their being judged in the intermediate zone of our experiments.

Finally, the long-term stability of the judgments in quantitative terms and the short term stability in qualitative terms underlines the reliability of judgments as an empirical foundation for studying language, whilst also providing a theoretically-informed way of getting consistent results on gradient judgments.

Conclusion

In conclusion, this paper revisited old questions about satiation, superiority effects and processing complexity with a new standardised method testing repeated exposure effects across a single experimental session.

The method validated claims of satiation effects, but demonstrated that the nature of these effects is different than claimed in previous studies: rather than a continuous rise, judgments rise only initially. The effects are sensitive to zone of well-formedness, not to natural syntactic classes.

Acknowledgements

We would like to thank Kim-Laura Speck for help developing the counterbalancing scheme used in the experiments, and Johannes Rothert for extracting the values in Table 1 from the relevant publications.

References

[1] Chomsky N. Aspects of the theory of syntax. MIT Press; 1965.

[2] Schütze C. The empirical base of linguistics: grammaticality judgments and linguistic methodology. Chicago: University of Chicago Press; 1996.

[3] Featherston S. Universals and grammaticality: wh-constraints in German and English. Linguistics. 2005;43:667–711.

[4] Weskott T, Fanselow G. On the informativity of different measures of linguistic acceptability. Language. 2011;87(2):249–273.

[5] Sprouse J, Almeida D. The empirical status of data in syntax: A reply to Gibson and Fedorenko. Language and Cognitive Processes. 2013;28(3):222– 228.

[6] Nagata H. The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. Journal of Psycholinguistic Research. 1988;17(1):1–17.

[7] Carroll JM. Complex compounds: phrasal embedding in lexical structures. Linguistics. 1979;17:863–877.

[8] Snyder W. An experimental investigation of syntactic satiation effects. Linguistic Inquiry. 2000;31:575–82.

[9] Chomsky N. Barriers. Cambridge, MA: MIT Press; 1986.

[10] Hiramatsu K. Accessing linguistic competence: evidence from children's and adults' acceptability judgments. University of Connecticut; 2000.

[11] Francom JC. Experimental syntax: Exploring the effect of repeated expo- sure to anomalous syntactic structure: Evidence from rating and reading tasks. University of Arizona. Tucson, AZ; 2009.

[12] Braze FD. Grammaticality, acceptability and sentence processing: A psycholinguistic study. University of Connecticut. Storrs, CT; 2002.

[13] Goodall G. Syntactic satiation and the inversion effect in English and Spanish wh-questions. Syntax. 2011;14(1):29–47.

[14] Crawford J. Using syntactic satiation to investigate subject islands. In: Choi J, Hogue EA, Punske J, Tat D, Schertz J, Trueman A, editors. Proceedings of the 29th West Coast Conference on Formal Linguistics. Cascadilla Proceedings Project; 2012. p. 38–45.

[15] Sprouse J. Continuous acceptability, categorical grammaticality, and experimental syntax. Biolinguistics. 2007;1:123–134.

[16] Sprouse J. A program for experimental syntax: finding the relationship between acceptability and grammatical knowledge. University of Maryland; 2007.

[17] Sprouse J. Revisiting Satiation: evidence for an equalization response strategy. Linguistic Inquiry. 2009;40:329–41.

[18] Hofmeister P, Jaeger TF, Arnon I, Sag IA, Snider N. The source ambiguity problem: distinguishing the effects of grammar and processing on accept- ability judgments. Language and Cognitive Processes. 2013;28:48–87.

[19] Chaves RP, Dery JE. Which subject islands will the acceptability of improve with repeated exposure? In: Santana-LaBarge RE, editor. Proceedings of the 31th West Coast Conference on Formal Linguistics. Cascadilla Proceedings Project; 2014. p. 96–106.

[20] Zervakis J, Mazuka R. Effect of repeated evaluation and repeated exposure on acceptability ratings of sentences. Journal of Psycholinguistic Research. 2013;42(6):505–525.

[21] Schad DJ, Vasishth S, Hohenstein S, Kliegl R. How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. Journal of Memory and Language. 2020;110.

[22] Rabe M, Vasishth S, Hohenstein S, Kliegl R, Schad D. hypr: an R package for hypothesis-driven contrast coding. Journal of Open Source Software. 2020.

[23] Cowart W. Experimental Syntax: applying objective methods to sentence judgments. Thousand Oaks: SAGE publications; 1997.

[24] Chaves RP, Dery JE. Frequency effects in subject islands. Journal of Linguistics. 2019;55(3):475–521.

[25] Christensen KR, Kizach J, Nyvad AM. Escape from the island: Grammaticality and (reduced) acceptability of wh-island violations in Danish. Journal of Psycholinguistic Research. 2013;42(1):51–70.

[26] Kuno S, Robinson J. Multiple wh-questions. Linguistic Inquiry. 1972;3:463– 487.

[27] Chomsky N. The Minimalist Program. Cambridge, MA: MIT Press; 1995.

[28] Chomsky N. Conditions on Transformations. In: Anderson SR, Kiparsky P, editors. A Festschrift for Morris Halle. New York: Holt, Rinehart and Winston; 1973. p. 232–286.

[29] Haider H. The superiority conspiracy: four constraints and a processing effect. In: Stepanov A, Fanselow G, Vogel R, editors. Minimality Effects in Syntax. Mouton de Gruyter; 2004. p. 147–176.

[30] Gebrich H, Schreier V, Featherston S. Standard items for English judgement studies: syntax and semantics. In: Featherston S, Hörnig R, von Wietersheim S, Winkler S, editors. Experiments in Focus: Information Structure and Semantic Processing. Berlin: de Gruyter; 2019. p. 305–328.

[31] Häussler J, Grant M, Fanselow G, Frazier L. Superiority in English and German: Cross-Language Grammatical Differences? Syntax. 2015;18:235– 265.

[32] Peirce JW. PsychoPy: Psychophysics software in Python. Journal of Neuroscience Methods. 2007;162(1–2):8–13.

[33] Fanselow G, Schlesewsky M, Vogel R, Weskott T. Animacy effects on crossing wh-movement in German. Linguistics. 2011;49:657–683.

[34] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. URL https://www.R-pro ject.org/.

[35] Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. Journal of Statistical Software. 2015.

[36] Wickham H, Averick M, Bryan J, Chang W, McGowan L, Fran¸cois R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(1686). Doi:10.21105/joss.01686.

[37] Wilke CO. Cowplot: streamlined plot theme and plot annotations for 'ggplot2'. https://cran.r-pro ject.org/web/packages/cowplot/index.html; 2019.

[38] Lüdecke D. sjPlot: data visualization for statistics in social science. https://cran.r-project.org/web/packages/sjPlot/; 2020.

[39] Bolker B, Robinson D. broom.mixed: tidying methods for mixed models. https://cran.rproject.org/web/packages/broom.mixed/index.html; 2020.

[40] Bates D, Kliegl R, Vasishth S, Baayen RH. Parsimonious mixed models. arXiv preprint arXiv:1506.04967.; 2017.

[41] Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D. Balancing type I error and power in linear mixed models. Journal of Memory and Language. 2017;94:305–315.

[42] Bates D, Alday P, Kleinschmidt D, Calder´on JBS, Noack A, Kelman T, et al. JuliaStats/MixedModels.jl: v2.3.0. https://doi.org/10.5281/zenodo.3727845; 2020.

[43] Pesetsky D. Wh-in-situ: movement and unselective binding. In: ter Meulen A, Reuland E, editors. The Representation of (In)definiteness. Cambridge, MA: MIT Press; 1987. p. 98–129.

[44] Pesetsky D. Phrasal movement and its kin. Cambridge, MA: MIT Press; 2000.

SUPPLEMENT: LMM parameter estimates

Table S1. LMM parameter estimates for ratings of German target sentences (Experiment 1)

	Complex LMM			**Zero-correlation LMM**		
Predictors	*beta*	*CI*	*z*	*beta*	*CI*	*z*
Grand mean	4.30	4.03 – 4.57	**30.85**	4.30	4.03 – 4.57	**30.85**
Order (so)	0.52	0.41 – 0.63	**8.97**	0.52	0.41 – 0.63	**8.93**
Animacy (an)	-0.05	-0.11 – 0.01	-1.62	-0.05	-0.11 – 0.01	-1.60
Block [2-6] – 1 (b1)	0.31	0.12 – 0.49	**3.24**	0.31	0.12 – 0.49	**3.27**
Block [3-6] – 2 (b2)	-0.08	-0.16 – -0.00	-1.98	-0.08	-0.16 – -0.00	-1.97
Block [4-6] – 3 (b3)	-0.03	-0.12 – 0.05	-0.82	-0.03	-0.12 – 0.05	-0.82
Block [5-6] – 4 (b4)	-0.00	-0.09 – 0.08	-0.11	-0.00	-0.09 – 0.08	-0.11
Block [6] – 5 (b5)	-0.01	-0.11 – 0.09	-0.22	-0.01	-0.11 – 0.09	-0.22
so x an	0.13	0.08 – 0.18	**5.00**	0.13	0.08 – 0.18	**5.10**
so x b1	-0.08	-0.19 – 0.02	-1.58	-0.08	-0.19 – 0.02	-1.59
so x b2	-0.04	-0.12 – 0.04	-1.01	-0.04	-0.12 – 0.04	-1.01
so x b3	0.05	-0.04 – 0.13	1.09	0.05	-0.04 – 0.13	1.09
so x b4	-0.01	-0.10 – 0.08	-0.22	-0.01	-0.10 – 0.08	-0.22
so x b5	0.07	-0.03 – 0.17	1.36	0.07	-0.03 – 0.17	1.36
an x b1	-0.07	-0.17 – 0.03	-1.45	-0.07	-0.17 – 0.02	-1.48
an x b2	-0.04	-0.12 – 0.04	-0.92	-0.04	-0.12 – 0.04	-0.92
an x b3	0.04	-0.04 – 0.12	1.00	0.04	-0.04 – 0.12	1.00
an x b4	-0.06	-0.14 – 0.03	-1.26	-0.06	-0.14 – 0.03	-1.25
an x b5	0.03	-0.07 – 0.13	0.63	0.03	-0.07 – 0.13	0.63
(so x an) x b1	-0.15	-0.26 – -0.04	**-2.60**	-0.15	-0.26 – -0.03	**-2.55**
(so x an) x b2	-0.02	-0.10 – 0.06	-0.55	-0.02	-0.10 – 0.06	-0.55
(so x an) x b3	0.04	-0.04 – 0.12	0.96	0.04	-0.04 – 0.12	0.95
(so x an) x b4	0.10	0.02 – 0.19	**2.34**	0.10	0.02 – 0.19	**2.33**
(so x an) x b5	-0.08	-0.18 – 0.02	-1.61	-0.08	-0.18 – 0.02	-1.60

Table S2. LMM parameter estimates for ratings of English target sentences (Experiment 2)

	Complex LMM			Zero-correlation LMM		
Predictors	*beta*	*CI*	*z*	*beta*	*CI*	*z*
Grand mean	4.71	4.45 – 4.97	**35.53**	4.71	4.45 – 4.97	**35.54**
Order (so)	0.67	0.56 – 0.78	**12.20**	0.67	0.56 – 0.78	**12.26**
D-linking (dl)	-0.35	-0.45 – -0.25	**-6.74**	-0.35	-0.45 – -0.25	**-6.74**
Block [2-6] – 1 (b1)	0.16	-0.02 – 0.33	1.78	0.16	-0.02 – 0.33	1.77
Block [3-6] – 2 (b2)	-0.08	-0.17 – 0.00	-1.91	-0.08	-0.17 – 0.00	-1.90
Block [4-6] – 3 (b3)	**-0.11**	**-0.20 – -0.02**	**-2.46**	-0.11	-0.20 – -0.02	**-2.46**
Block [5-6] – 4 (b4)	-0.03	-0.12 – 0.06	-0.67	-0.03	-0.12 – 0.06	-0.67
Block [6] – 5 (b5)	-0.02	-0.12 – 0.09	-0.29	-0.02	-0.12 – 0.09	-0.29
so x dl	0.43	0.34 – 0.52	**9.46**	0.43	0.34 – 0.52	**9.52**
so x b1	-0.29	-0.41 – -0.18	**-4.92**	-0.29	-0.41 – -0.18	**-5.01**
so x b2	-0.14	-0.22 – -0.05	**-3.16**	-0.14	-0.22 – -0.05	**-3.16**
so x b3	-0.08	-0.17 – 0.00	-1.90	-0.08	-0.17 – 0.00	-1.90
so x b4	-0.05	-0.15 – 0.04	-1.13	-0.05	-0.15 – 0.04	-1.13
so x b5	-0.03	-0.14 – 0.07	-0.63	-0.03	-0.14 – 0.07	-0.63
dl x b1	0.09	-0.01 – 0.19	1.76	0.09	-0.01 – 0.19	1.79
dl x b2	0.10	0.01 – 0.18	**2.27**	0.10	0.01 – 0.18	**2.27**
dl x b3	0.01	-0.08 – 0.10	0.19	0.01	-0.08 – 0.10	0.19
dl x b4	0.02	-0.08 – 0.11	0.36	0.02	-0.08 – 0.11	0.36
dl x b5	-0.01	-0.12 – 0.10	-0.17	-0.01	-0.12 – 0.10	-0.17
(so x dl) x b1	-0.04	-0.12 – 0.05	-0.89	-0.04	-0.13 – 0.06	-0.78
(so x dl) x b2	-0.02	-0.11 – 0.06	-0.58	-0.02	-0.11 – 0.06	-0.58
(so x dl) x b3	0.04	-0.05 – 0.12	0.81	0.04	-0.05 – 0.12	0.81
(so x dl) x b4	0.02	-0.07 – 0.12	0.52	0.02	-0.07 – 0.12	0.52
(so x dl) x b5	0.03	-0.08 – 0.14	0.55	0.03	-0.08 – 0.14	0.55

Table S3. LMM parameter estimates for ratings of English target sentences (Experiment 3)

	Complex LMM			Zero-correlation LMM		
Predictors	*beta*	*CI*	*z*	*beta*	*CI*	*z*
Grand mean	4.93	4.72 – 5.15	**44.48**	4.94	4.72 – 5.15	**44.33**
Order (so)	0.74	0.63 – 0.85	**12.93**	0.74	0.62 – 0.85	**12.81**
D-linking (dl)	-0.45	-0.55 – -0.35	**-8.87**	-0.45	-0.55 – -0.35	**-8.92**
Block [2-6] – 1 (b1)	0.39	0.22 – 0.56	**4.54**	0.38	0.21 – 0.56	**4.35**
Block [3-6] – 2 (b2)	0.14	0.06 – 0.22	**3.53**	0.14	0.06 – 0.22	**3.49**
Block [4-6] – 3 (b3)	0.03	-0.05 – 0.11	0.74	0.03	-0.05 – 0.11	0.67
Block [5-6] – 4 (b4)	0.01	-0.08 – 0.10	0.24	0.01	-0.08 – 0.10	0.22
Block [6] – 5 (b5)	0.05	-0.05 – 0.15	1.06	0.05	-0.05 – 0.15	1.02
so x dl	0.56	0.46 – 0.66	**10.97**	0.56	0.46 – 0.66	**11.02**
so x b1	-0.13	-0.24 – -0.02	**-2.24**	-0.12	-0.23 – -0.01	**-2.13**
so x b2	-0.12	-0.20 – -0.05	**-3.09**	-0.13	-0.21 – -0.05	**-3.17**
so x b3	-0.08	-0.16 – 0.00	-1.92	-0.08	-0.16 – 0.00	-1.87
so x b4	0.02	-0.07 – 0.10	0.37	0.02	-0.07 – 0.10	0.40
so x b5	0.05	-0.05 – 0.15	0.93	0.05	-0.05 – 0.15	1.03
dl x b1	0.01	-0.11 – 0.13	0.17	0.01	-0.11 – 0.13	0.16
dl x b2	0.07	-0.01 – 0.14	1.66	0.07	-0.01 – 0.14	1.62
dl x b3	0.03	-0.05 – 0.11	0.72	0.04	-0.04 – 0.13	1.05
dl x b4	0.00	-0.08 – 0.09	0.02	0.01	-0.07 – 0.10	0.29
dl x b5	0.06	-0.04 – 0.16	1.14	0.07	-0.03 – 0.17	1.36
(so x dl) x b1	0.10	-0.03 – 0.22	1.51	0.08	-0.05 – 0.20	1.19
(so x dl) x b2	-0.02	-0.10 – 0.06	-0.53	-0.02	-0.10 – 0.06	-0.57
(so x dl) x b3	-0.01	-0.09 – 0.07	-0.22	0.00	-0.08 – 0.09	0.09
(so x dl) x b4	-0.02	-0.11 – 0.07	-0.47	-0.00	-0.09 – 0.08	-0.10
(so x dl) x b5	0.01	-0.09 – 0.11	0.18	0.02	-0.08 – 0.12	0.41