Language and Cognition **6** (2014), 111–145. doi:10.1017/langcog.2013.7

Processing effects in linguistic judgment data: (super-)additivity and reading span scores*

PHILIP HOFMEISTER
Department of Language & Linguistics, University of Essex, UK

LAURA STAUM CASASANTO
Department of Linguistics, SUNY at Stony Brook, NY

AND

IVAN A. SAG
Department of Linguistics, Stanford University, Stanford, CA

(*Received* 06 *January* 2012 – *Revised* 28 *January* 2013 – *Accepted* 24 *April* 2013)

ABSTRACT

Linguistic acceptability judgments are widely agreed to reflect constraints on real-time language processing. Nonetheless, very little is known about how processing costs affect acceptability judgments. In this paper, we explore how processing limitations are manifested in acceptability judgment data. In a series of experiments, we consider how two factors relate to judgments for sentences with varying degrees of complexity: (1) the way constraints combine (i.e., additively or super-additively), and (2) the way a comprehender's memory resources influence acceptability judgments. Results indicate that multiple sources of processing difficulty can combine to produce super-additive effects, and that there is a positive linear relationship between reading span scores and judgments for sentences whose unacceptability is attributable to processing costs. These patterns do not hold for sentences whose unacceptability is attributable to factors other than processing costs, e.g., grammatical constraints. We conclude that tests of (super)-additivity and of relationships to reading span scores can help to identify the effects of processing difficulty on acceptability judgments, although these tests cannot be used in contexts of extreme processing difficulty.

KEYWORDS: sentence processing, acceptability judgments, grammar, individual differences, working memory

[*] Addresses for correspondence: Philip Hofmeister, Department of Language & Linguistics, University of Essex, Wivenhoe Park, CO4 3SQ, United Kingdom. e-mail: phofme@essex.ac.uk.

1. Introduction

Acceptability judgments are the primary source of evidence that linguists use to design theories of grammar. George Miller once noted in an address that "the form of the grammar is settled on for very good reasons, but for reasons that do not attempt to take account of any data other than primary linguistic intuitions" (Miller, 1975). What these intuitions or judgments imply about linguistic knowledge, however, is not straightforward. A rich history of research in linguistics and psychology, dating back to Miller and Chomsky (1963) (and predated by Saussure's, 1916, LANGUE–PAROLE distinction), makes it clear that judgments of linguistic acceptability are colored by 'performance' factors – limitations on cognitive resources and usage (Chomsky, 1965; Bever, 1970; Watt, 1970; Pylyshyn, 1973; Pritchett, 1992; *inter alia*). On the standard view that linguistic competence is a stable system of knowledge that is independent of performance factors, this makes judgments of acceptability naturally ambiguous. Any contrast in acceptability judgments between two sentences may reflect principles of grammar, limitations on sentence processing, or both.[1]

If linguists are to continue using judgments as the primary evidence for building grammatical theories, being able to identify and understand effects of processing difficulty on judgments serves an important function. For instance, if it was apparent that an acceptability contrast is largely attributable to processing differences, a grammatical constraint to explain that same contrast may well be otiose (Bever, Carroll, & Hurtig, 1976). Moreover, if the objective is to "see the grammar bare" (Gleitman & Gleitman, 1970), that is, unobscured by other confounding factors in acceptability judgment data, then there is a clear imperative to sharpen our understanding of how processing complexity bears on intuitions of well-formedness (Schütze, 1996).

Knowing when acceptability differences are at least partly attributable to contrasts in processing costs would also inform some long-standing debates in linguistics. For example, in the study of sentences with 'island violations' – sentences where a linguistic dependency in a particular syntactic configuration is judged to be unacceptable – both processing factors and grammatical constraints have been proposed to account for the unacceptability (Ross, 1967; Chomsky, 1973, 1977, 1981, 1986; Kluender, 1992, 1998; Kluender & Kutas, 1993; Phillips, 2006; Hofmeister & Sag, 2010; Sprouse, Wagers, & Phillips, 2012; Hofmeister, Jaeger, Arnon, Sag, & Snider, 2013; Hofmeister, Staum Casasanto, & Sag, 2013; Sprouse & Hornstein, 2013). Generally speaking, knowing when processing differences are at play in

[1] Exogenous factors may also play a role in judgment variation, as detailed by Schütze (1996).

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

acceptability contrasts is an essential ingredient to a more comprehensive understanding of the relationship between formal and functional factors.

In this vein, our objective here is to follow the admonition of Schütze (1996, p. 9): "linguists ought to study their methodology … Eliminating or controlling for confounding factors requires us to have some idea of what those factors might be, and such an understanding can only be gained by a systematic study of the judgment process." In the present case, even though there is a general consensus that processing costs enter into estimations of acceptability, relatively little else is known about how processing costs map onto acceptability judgments. To take the easy way out, we could blanketly assume that effects of processing costs on judgments will parallel their effects in other empirical domains. That is, we might take the bold but unsupported position that wherever processing differences exist (however we measure them), acceptability differences will as well, and that the direction and magnitude of processing differences will be faithfully reflected in acceptability differences. But this overlooks the very real possibility that judgment tasks differ not only in their degree of sensitivity to processing costs but also the kinds of processing cost they reflect, and the existing evidence already hints that greater processing complexity may not always be realized as lower acceptability (Frazier, 1985; Gibson & Thomas, 1999; Fanselow & Frisch, 2004; Sprouse, 2009).

This leads to the following question: How can we know whether processing differences contribute to an acceptability contrast? Assuming that processing differences are realized as acceptability differences because of PROCESSING LIMITATIONS, we arrive at a slightly modified version of the preceding question that becomes the central focus of this paper: How are processing limitations manifested in acceptability judgment tasks? By 'limitations', we refer to either the finiteness of a set of available resources or processing bottlenecks, i.e., some processes may be constrained to begin only after others have finished (Welford, 1952; Pashler, 1994; Pashler & Johnston, 1998; Ferreira & Pashler, 2002). A consequence of these limitations is that a sufficient level of processing difficulty can (temporarily) exhaust these resources, leading to a processing breakdown and/or significant delays (Gibson, 1991).

We approach these questions from the following perspective: there are general indicators of limitations to processing resources which apply across various methodologies. One is the observation of so-called super-additive effects, where two stimulus properties or tasks combine to have an effect on a dependent variable that surpasses the sum of their independent effects. According to the logic of the additive factors model (Sternberg, 1969), super-additive effects indicate that two processes draw on the same limited pool of resources. For instance, Fedorenko, Gibson, and Rohde (2007) found that reading harder-to-process sentences (e.g., sentences with object relative

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

clauses vs. subject relative clauses) while simultaneously doing complex arithmetic tasks (made harder than easier arithmetic tasks by making the addends larger) slows reading rates down super-additively. They conclude that the two tasks draw on the same pool of cognitive resources. Evidence of super- additivity, therefore, suggests that the task demands stress the limits of the relevant cognitive system, whether those limitations involve a finite set of resources, processing bottlenecks, or both. To be clear, though, the absence of super-additive effects does not imply that the relevant pool of resources is UNLIMITED. Even if some set of cognitive resources are limited, simultaneous demands may together make proportionately little demand on the system. Signs of processing limitations thus only emerge in the form of super-additive effects when the combined demands exceed a critical threshold. Moreover, because the cognitive demands must interact with one another, they must overlap to some significant extent within a sufficiently narrow window of time.

Second, we assume that correlations between measures of individual differences in neuropsychological assessments, such as one of the variety of memory span tasks, and performance on a secondary task (e.g., a reading or acceptability task) indicate the extent of cognitive limitations on the secondary task. The more that the secondary task calls upon the resources measured by the neuropsychological assessment, the stronger the correlation will be. The nature of those limitations depends upon what aspects of cognition (e.g., memory, attention, task switching, spatial reasoning, etc.) the neuropsychological assessment actually measures. For language processing, numerous researchers have proposed that the resources available for language processing differ from one individual to the next, and that this variation can explain the magnitude of syntactic complexity effects (King & Just, 1991; Just & Carpenter, 1992; King & Kutas, 1995).[2] As with super-additivity, the absence of a correlation between scores from some neuropsychological test and a psycholinguistic task clearly does not entail unlimited processing resources or that functional factors are irrelevant for the task.

Our strategy here is to assess whether and how these indicators of processing limitations relate to judgments for sentences with varying degrees of complexity. In light of the wealth of preceding work on the role of memory constraints in determinations of processing complexity, we focus primarily on processing costs standardly attributed to memory retrieval in language processing. We leave aside processing costs due to probabilistic factors such

[2] There is a more nuanced and orthogonal debate about whether the resources used for language processing overlap with the resources used in other cognitive tasks (Just & Carpenter, 1992; Caplan & Waters, 1999). We opt to not enter these deep waters.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

as word or phrase expectancy, even though these factors may be as, if not more, important than considerations of memory retrieval in real-time language processing. It is a matter for future research to determine whether the results we report here generalize to other sources of processing difficulty besides memory costs.

As a comparison set for the complex sentences, we examine cases where acceptability differences are not generally taken to reflect online processing costs. More specifically, we look at cases similar to (1) below:

(1) a. I embarrassed him.
    b. I embarrassed he.

The acceptability difference between (1a) and (1b) is widely seen as being causally related to factors OTHER THAN PROCESSING COMPLEXITY. The standard characterization here involves the notion of 'grammar'; however, in what follows, our primary aim is not to set up a contrast between grammar and processing. Instead, the aim is to juxtapose phenomena where the causal agent in acceptability differences IS OR IS NOT online sentence processing complexity. Beyond the fact that examples like (1) and other similar cases we test below are classically treated in terms other than processing complexity, there are principled reasons for presuming that processing costs are not what separates the examples in (1).[3] The examples above are short, involve frequent lexical items, express plausible events, and, while case-marking acts as an indicator of thematic role, such information is independently recoverable from word order cues in English. Thus, a possible meaning for (1b) is not difficult to surmise. In the end, readers will be convinced of our claims about how processing complexity affects judgments to the extent that they believe that the acceptability difference between examples like (1a) and (1b) pertains to considerations besides differential processing complexity.

For each type of stimulus – those varying in complexity and those varying in grammaticality – two properties of the corresponding judgment data are examined in accord with the above discussion. First, we consider what happens when multiple sources of unacceptability combine in the same sentence. Logically, three distinct outcomes could result from combining multiple sentence features that each individually lower acceptability ratings:

– a penalty significantly smaller than the sum of the two individual penalties, a result which we refer to as UNDER-ADDITIVE;

[3] This leaves open the possibility that the realization of case-marking distinctions, historically speaking, depends upon functional considerations.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

- a penalty statistically indistinguishable from the sum of the two individual penalties, which we refer to as ADDITIVE;
- a penalty significantly larger than the sum of the two individual penalties, which we call SUPER-ADDITIVE.

Accordingly, we test here how online processing costs (PCs) combine with each other to lower judgments, how grammatical constraint violations (GCVs) combine with each other, and how the two sources of unacceptability affect judgments when they co-occur in the same sentence.

Some preceding work already confirms that participants take into account multiple GCVs in their judgments of acceptability. Sorace and Keller (2005), for instance, illustrate how the ratings for a sentence become progressively lower as more GCVs are added (based on data from Keller, 2000):

(2) a. Which friend has Thomas painted a picture of?
b. Which friend Thomas has painted a picture of?
c. Which friend Thomas have painted a picture of?
d. Which friend Thomas have painted a picture of her?

In (2b), the non-inverted auxiliary lowers judgments, and this penalty is added to by the agreement error in (2c), and by the presence of a resumptive pronoun in (2d). Although Sorace and Keller (2005) do not report directly on the issue of super-additivity, the data suggest that such GCVs do not combine super-additively, i.e., there is no steep drop-off in acceptability in 2 vs. 1 or 3 vs. 2 GCVs. It thus remains to be seen whether sources of unacceptability in judgment tasks ever combine to yield super-additive effects. The experimental studies described below are accordingly aimed at determining whether and under what conditions such effects occur.

The second component of our investigation relates individual cognitive differences to judgment data. Here, the relevant question is how, if at all, do measures of these cognitive capacities relate to judgments for hard-to-process or complex sentences? In principle, there are several distinct ways that any particular measure of individual cognitive differences might relate to judgments for such sentences:

- individuals with greater (or faster) resources provide higher judgments for complex sentences than their low-resource counterparts;
- individuals with greater resources rate sentences with any perceived abnormalities (i.e., anything that triggers an acceptability penalty) as being worse than their lower-resource counterparts would;
- no systematic relationship exists between the measure of cognitive capacity and judgments.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

A recent exploration of this topic by Sprouse et al. (2012) found that two measures of individual cognitive differences – the *n*-back task and a serial recall task – bore no systematic relationship to acceptability judgments for the magnitude of syntactic island effects. The absence of a correlation led Sprouse et al. to conclude that island effects do not derive from processing complexity. As noted in the response of Hofmeister, Staum Casasanto, and Sag (2012a), a drawback of this study is that there is no body of evidence to indicate how these two measures relate to judgments for hard-to-process sentences generally. It could be, for instance, that no matter what kind of sentence is tested, these two measures of memory lack any relationship to the supplied judgments (for further criticisms, see Hofmeister et al., 2012a, and Hofmeister, Staum Casasanto, and Sag, 2012b). In short, we simply do not know if acceptability judgments ever systematically relate to measures of individual cognitive differences (and there may well be significant differences across measures). While it is possible that individuals who score higher on neuropsychological assessments of cognitive properties like memory will be less strained by the demands of complex sentences and thus provide higher acceptability judgments, this hypothesis requires confirmation with each measure of individual cognitive differences and an array of sentences differing in complexity. A major objective of this research, therefore, is to determine if and how individual differences in processing limitations influence judgments for different types of sentences.

1.1. ON THE PROCESSING/GRAMMAR DIVIDE

While examining how processing limitations are manifested in judgment data, we remain agnostic about how grammatical constraints and processing limitations relate. We do not assume that all grammatical constraints have their origin in functional considerations, or that the two are fundamentally distinct. Instead, we are sampling from the kind of examples linguists standardly characterize in terms other than processing costs (i.e., grammar) and from those characterized in terms of processing costs to identify what distinguishing effects processing limitations have on acceptability judgments. If there are no observable differences, we could not draw strong conclusions about the uniformity of 'performance' and 'competence' factors, as such a conclusion would hinge upon null results. Conversely, the observation of contrastive patterns by itself is not confirmation of categorical differences, even if consistent with such an interpretation. Such observations leave open the possibility that two sets of stimuli fall along a single spectrum but with sufficient distance that they appear to belong to entirely separate categories, much like the difference between voiced and non-voiced consonants with the same place and manner of articulation, e.g., /p/ vs. /b/. In short, our work does not speak to the degree of autonomy between grammatical constraints and processing-related constraints.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

In the following sections, we describe experiments testing how PCs combine to affect judgments (Experiment 1), how GCVs combine to affect judgments (Experiments 2a and 2b), and how the two interact (Experiment 3). In the final experiment, we assess how generalizable the results are by examining a case of extreme processing difficulty.

2. Gathering acceptability judgments

To acquire acceptability ratings, we used the thermometer judgment (TJ) methodology described in Featherston (2008), which resembles the Magnitude Estimation (ME) technique of gathering judgments. In ME experiments, participants are asked to rate the magnitude of acceptability difference between test items and a reference sentence (e.g., twice as good, three times as good, half as good, etc.; Bard, Robertson, & Sorace 1996; Sorace & Keller 2005).

There are several differences, however, between the ME and TJ methodologies. In the latter, participants are not instructed to evaluate test items in terms of the magnitude of acceptability compared to the reference item, as evidence shows that participants ignore these instructions and rate sentences in terms of their linear distance from the reference (Featherston, 2008). In TJ studies, participants judge items relative to two reference sentences in terms of linear distance. One of these references is quite good and the other quite bad, and we follow Featherston (2008) in assigning these sentences the arbitrary values 20 and 30. For all of our experiments, we used the same reference sentences:

(3) a. The way that the project was approaching to the deadline everyone wondered. = 20
b. The architect told his assistant to bring the new plans to the foreman's office. = 30

Test sentences were presented to participants on a computer screen one word at a time for a fixed duration via the DMDX software package (Forster & Forster, 2003). The duration varied with the number of characters in the word (250 ms + 33.34 * number of characters), so that longer words remained visible for longer periods. We chose word-by-word presentation over full sentence presentation to prevent participants from excessive introspection about the test sentences, and auto-paced presentation rather than self-paced presentation to prevent differences in how long each participant studied a given stimulus.

Each participant also completed a reading span task during the same session to assess their memory span (Daneman & Carpenter, 1980). We used this memory span task largely because of the rich history of its use, the extensive body of literature on the underlying cognitive constructs,

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

and its strong relationship to measures of listening and reading comprehension (Just & Carpenter, 1992; Towse, Hitch, & Hutton, 2000; Vos, Gunter, Schriefers, & Friederici, 2001; Whitney, Arnett, Driver, & Budd, 2001; Friedman & Miyake, 2004; Conway et al., 2005; Daneman & Hannon, 2007).[4]

We scored each test using the partial credit method outlined in Conway et al. (2005): successful recall of a word in a study list counts toward the final reading span score, even if the entire item set was not recalled correctly. This method provides a greater range of reading span scores and differentiates individuals more than methods that only describe the maximum recall level reached.

Prior to statistical analysis, we computed *z*-scores for each subject on the basis of all data in the experimental dataset (except practice items), including fillers. This reduces the impact of varying uses of the interval scale by subjects. Finally, we excluded data points with *z*-scores more than 2.5 standard deviations from each condition mean. For Experiment 1, this outlier removal process affected 2.3% of the data. The resulting *z*-scores constitute the data on which we conducted statistical analyses.

For all experiments, we used linear mixed effects models to estimate the effects of the experimental manipulations (Baayen, 2008; Baayen, Davidson, & Bates, 2008). This method of statistical analysis also allows for the evaluation of additional factors such as reading span score alongside effects due to direct experimental manipulation. Prior to analysis, all predictors were centered – higher-order variables (interactions) were also based on these centered predictors.

For each experiment, we utilized the maximal random effect structure that converged. That is, for a design with two factors, F_1 and F_2, the random effect structure included random intercepts for participants and items, as well as by-participant and by-item random slopes for each factor and the interaction ($F_1 \times F_2$). This type of design essentially parallels the logic of classical ANOVAs, as it acknowledges that the effect of treatment conditions may vary across experimental participants and items. In our studies, all such models successfully converged, making it unnecessary to drop any terms from the random effect structure specifications. Although models with nested random effect structures do not directly yield *p*-values, significance at the .05 level can be conservatively estimated for fixed effects coefficients with *t*-values which have absolute values at or above 2 (Baayen, 2008; Baayen et al., 2008; Pinheiro & Bates, 2000).

[4] The reading span task most likely taps other aspects of cognition besides memory, including attention (Whitney et al., 2001; Conway et al., 2005; inter alia). Our findings here thus potentially speak to individual differences in cognitive ability besides memory. We take this as an advantage of our approach, however, as we do not treat all processing difficulty in sentence processing as being reflective of memory retrieval difficulty.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

3. Experiment 1: processing difficulty

In this experiment, we evaluate how judgments of acceptability are affected by increasing the number of distinct sources of processing difficulty. We also ask if and how individual differences, as measured by the reading span task, relate to judgments for these sentences with varying degrees of processing costs.

3.1. PARTICIPANTS

Thirty-two Stanford University students participated in exchange for payment. All self-identified as native speakers of English.

3.2. MATERIALS

We utilized twenty-four items from Grodner and Gibson (2005), who manipulated the distance between two dependent arguments and their syntactic head. In these items, the hierarchical distance between a subject and object noun phrase and their subcategorizing verb was varied. This was achieved by varying (i) the presence/absence of a relative clause between the subject and verb (4a)/(4c) vs. (4b)/(4d) and (ii) positioning the object NP immediately after the verb or before the subject NP by relativizing it:

(4) a. [SHORT–SHORT] The nurse from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room.
b. [LONG–SHORT] The nurse who was from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room.
c. [SHORT–LONG] The administrator who the nurse from the clinic supervised scolded the medic while a patient was brought into the emergency room.
d. [LONG–LONG] The administrator who the nurse who was from the clinic supervised scolded the medic while a patient was brought into the emergency room.

These items were selected because reading time evidence from Grodner and Gibson (2005) shows that increasing the hierarchical distance in examples like these leads to slower processing at the critical integration sites (the verbs *scolded* and *supervised* in (4) above). Moreover, increasing subject and object distance simultaneously, as in the LONG–LONG condition, led to the slowest overall reading times. The twenty-four experimental items appeared with seventy-two fillers (twenty-four of which were the items for Experiment 3). Each participant saw only one condition of each item. The order of the materials was pseudo-randomized by DMDX.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

3.3. PROCEDURE

Each session began with four practice trials to familiarize participants with the rating task and the TJ scale. In the practice and main experimental session, each sentence was presented word-by-word, after which a new screen repeated the reference sentences. A comprehension question followed this judgment stage as a further motivator for participants to read the test sentences. Immediately after the acceptability task, participants completed the reading span task. All participants saw exactly the same sentences in the reading span task in the same order. Moreover, each participant completed all levels of difficulty on the reading span task, regardless of their recall accuracy.

3.4. RESULTS

Additivity

Both distance manipulations produce main effects – subject distance and object distance lower judgments. In addition, these factors interact significantly (see Table 1). As Figure 1 depicts, the acceptability decrement produced by two processing costs is greater than the sum of the decrements produced by each cost in isolation.

Individual differences

Reading span score is also a highly significant predictor of acceptability scores. In particular, higher reading span scores predict higher judgment *z*-scores. As Figure 2 shows, this effect is driven by the conditions with longer dependencies between the object and verb – the most difficult to process conditions, according to Grodner and Gibson (2005) and corroborated by Bartek, Lewis, Vasishth, and Smith (2011) – which is reflected by the significant interaction of reading span score and the object distance manipulation.

TABLE 1. *Fixed effect summary for Experiment* 1

	Coefficients	Standard Error	*t*-value
SubjDistance	−0.320	0.063	−5.05
ObjDistance	−0.255	0.061	−4.21
SubjDistance × ObjDistance	−0.432	0.103	−4.18
Reading span	0.057	0.015	3.82
Reading span × SubjDistance	0.002	0.025	0.07
Reading span × ObjDistance	0.094	0.026	3.66
Reading span × SubjDistance × ObjDistance	−0.005	0.046	−0.12

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

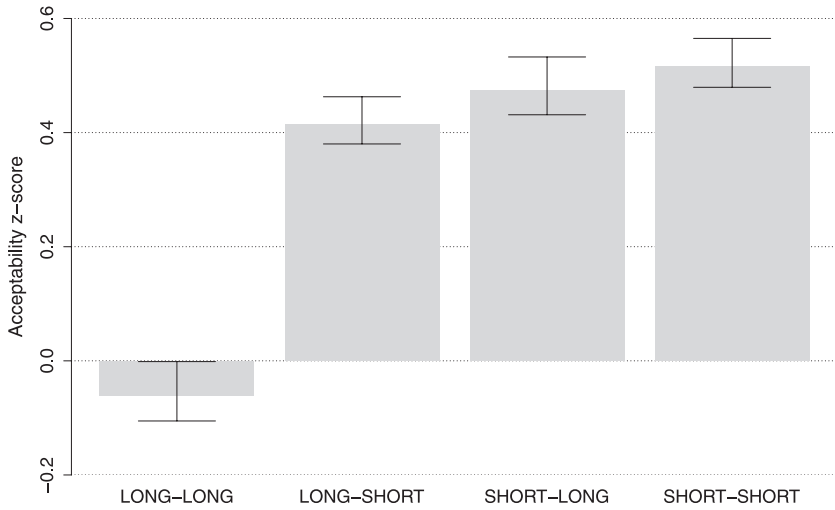

Fig. 1. Mean acceptability z-scores from Experiment 1. Error bars show ±1 standard error.

3.5. DISCUSSION

According to the results, processing costs may have only minor effects on acceptability in isolation, yet have highly significant effects on judgments when combined. Increasing the distance between a single dependent argument and its head only lowered judgments slightly. But when we simultaneously increased the hierarchical distance of both dependents to their syntactic head, a sharp drop in acceptability occurred. Consequently, these data provide positive evidence that there can be super-additive consequences for judgments when multiple PCs co-occur, at least under some circumstances. This means that judgment data can reflect processing limitations, and not simply processing costs, as evidenced by the super-additive effects. As far as we know, this is the first data to show that unambiguous processing costs can yield super-additive effects in acceptability data.

The second finding is that individuals with higher reading spans provided higher acceptability judgments for the most difficult-to-process conditions, based on self-paced reading and eye-tracking data from Grodner and Gibson (2005) and Bartek et al. (2011). In sentences where the processing demands were less, individual differences played little role in judgment variation. This suggests that sentence processing complexity modulates the relationship between reading span scores and acceptability scores: greater processing complexity leads to a more positive slope between reading span and acceptability scores (but see the results of Experiment 4).

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

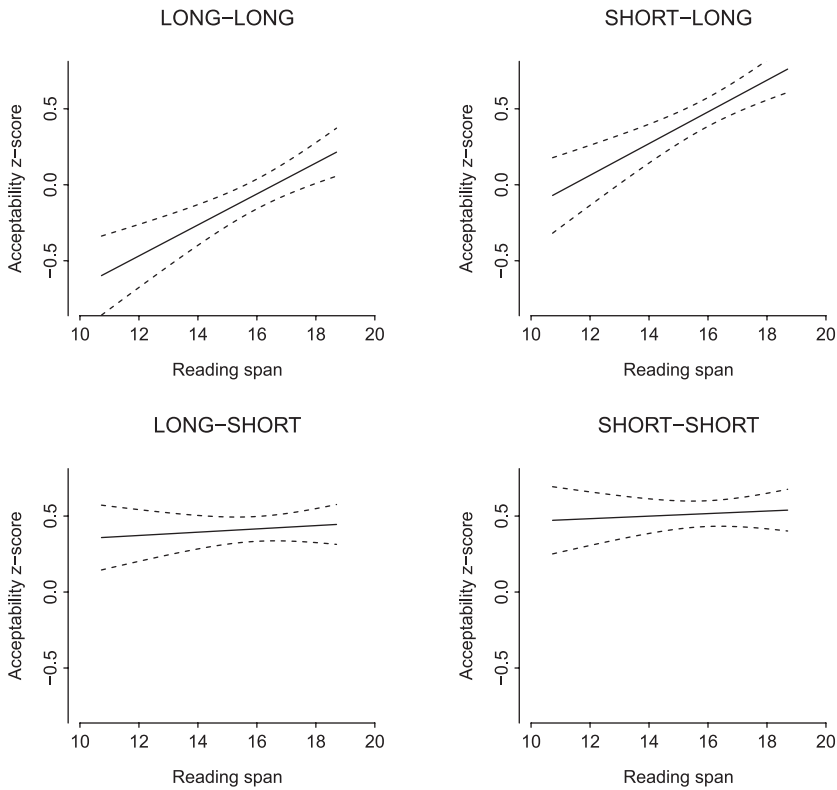

Fig. 2. Effects of reading span score on acceptability *z*-score for each condition in Experiment 1, according to ordinary least squares regression modeling.

Whether or not these features are specific to sentences whose unacceptability relates to processing costs, however, depends on whether similar relationships appear in sentences with GCVs. This is the subject of the next series of experiments.

4. Experiment 2a: separate grammatical violations

Experiment 2a evaluates how multiple GCVs affect judgments when they co-occur in the same sentence. The objective here is to compare such a scenario with the effects of multiple PCs within a single sentence, as well as the relationship between reading span scores and judgments for sentences with varying numbers (and types) of GCVs.

Some long-standing assumptions about the nature of GCVs, in fact, lead us to anticipate a different configuration of results in Experiment 2. Unlike

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

the case of PCs, we know of no claims that GCVs combine in a super-additive fashion. Such an account would imply that a GCV is intensified (i.e., violations become more egregious) in the context of another GCV. Moreover, to the extent that GCVs lower ratings for reasons other than processing/memory costs, there is little reason to expect that the severity of GCVs varies across speakers of differing memory capacities. Thus, if sentences with GCVs are unacceptable for reasons other than processing costs, estimates of cognitive ability such as the reading span should relate differently to judgments of sentences with GCVs compared to those with PCs.

4.1. PARTICIPANTS

Stanford University students ($n = 28$) who had not participated in Experiment 1 completed this experiment in exchange for payment.

4.2. MATERIALS

The twenty-four experimental items in Experiment 2a contained either zero, one, or two GCVs. We manipulated the grammaticality of two separate but nearby constituents to yield a 2 × 2 design. The first manipulation targeted the morphological form of a verb in a subject relative clause. Subjects either saw the correct form (5a)/(5b) or they saw a form that was missing the appropriate inflectional morphology (5c)/(5d). Additionally, participants either read an object pronoun with the proper case-marking (5b)/(5d) or they read a pronoun with unlicensed nominative case-marking (5a)/(5c):

(5) a. [GOOD−BAD]: The friend who visited Sue asked she whether the value of the house had dropped since the recession began.
b. [GOOD−GOOD]: The friend who visited Sue asked her whether the value of the house had dropped since the recession began.
c. [BAD−BAD]: The friend who visit Sue asked she whether the value of the house had dropped since the recession began.
d. [BAD−GOOD]: The friend who visit Sue asked her whether the value of the house had dropped since the recession began.

Seventy-two filler items appeared along with the critical items. As in the previous experiment, all items were followed by comprehension questions.

4.3. PROCEDURE

The procedure was identical to Experiment 1. Data were analyzed using the same methods as in Experiment 1. Outlier removal affected 0.89% of the data.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

TABLE 2. *Fixed effect summary for Experiment 2a*

	Coefficients	Standard Error	*t*-value
Finiteness	−0.506	0.082	−6.16
Case	−0.726	0.106	−6.86
Finiteness × Case	0.268	0.134	1.99
Reading span	−0.008	0.040	−0.21
Reading span × Finiteness	0.048	0.039	1.23
Reading span × Case	−0.103	0.047	−2.17
Reading span × Finiteness × Case	0.009	0.089	0.10

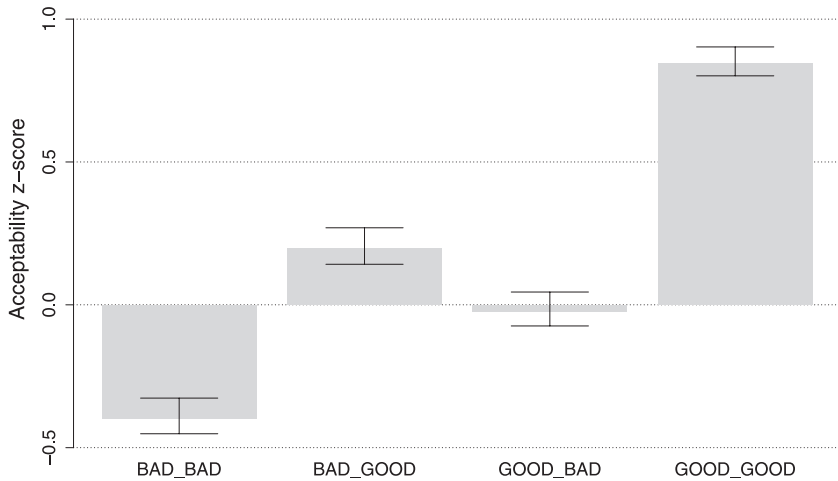

Fig. 3. Mean acceptability *z*-scores from Experiment 2a. Error bars show ±1 standard error.

4.4. RESULTS

Additivity

Both inflectional morphology and case errors significantly lower acceptability judgments, as Table 2 indicates. These factors also interact marginally, because the two GCVs in combination yield an acceptability decrement that is less than the sum of the decrements caused by each error in isolation, as seen in Figure 3.

Individual differences

No main effect of reading span was found for these stimuli. For the conditions judged the worst by participants (those with either a case error or a case error and an inflectional error), however, reading span scores exhibit a negative linear relationship with *z*-scores (see Figure 4). That is, individuals judge these conditions as being even less acceptable than individuals with lower reading span scores do, relative to the other items in

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

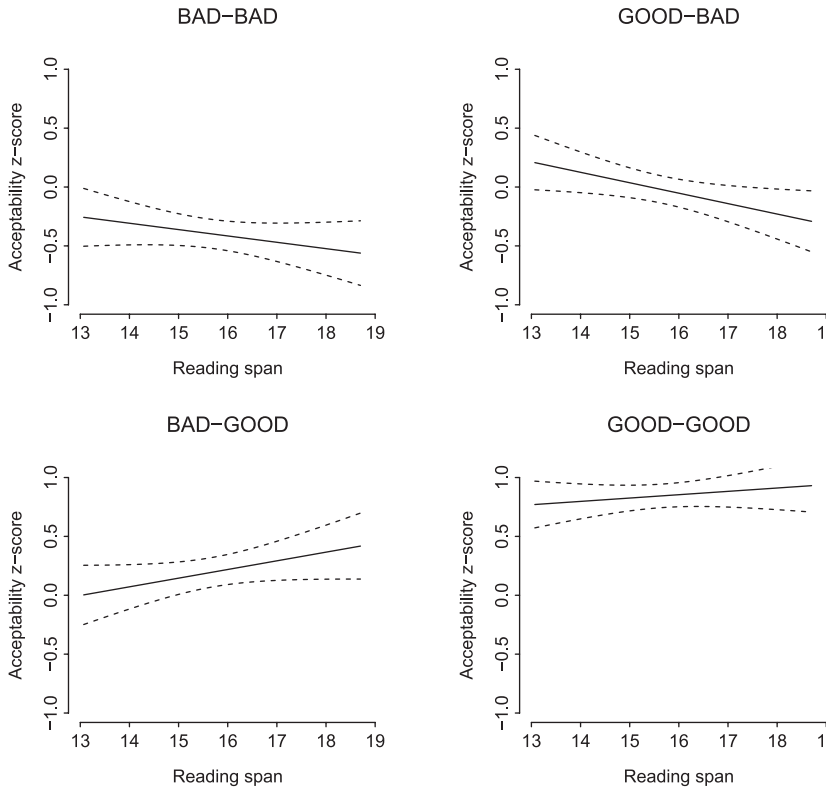

Fig. 4. Effects of reading span score on acceptability *z*-score for each condition in Experiment 2a, according to ordinary least squares regression modeling.

the experiment. This difference between the conditions leads to a statistically reliable interaction of reading span score and the case manipulation.

4.5. DISCUSSION

In contrast to the results of combining PCs in Experiment 1, combining GCVs did not result in super-additive effects; the effect of two co-occurring, proximal violations did not reduce judgments further than expected on the basis of each violation in isolation. In fact, the results suggest the opposite: a slightly smaller decrement than expected when the two GCVs co-occur. As Sorace and Keller (2005) note, similar findings of cumulativity (i.e., GCVs 'stacking' up) occur with constraints on word order and gapping (Keller, 2000), as well as selectional restrictions and subcategorization requirements (Chapman, 1974). It thus appears to be true in a number of

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

cases that respondents factor multiple GCVs into their judgments. But in none of the cases cited is there evidence that combining GCVs results in a super-additive acceptability penalty – all known cases result in either additive or under-additive decrements.

The other important contrast between the first two experiments involves the relationship between reading span scores and acceptability scores. We found a positive linear relationship between the two for sentences with relatively high processing difficulty in Experiment 1. In Experiment 2a, higher reading spans were associated with lower judgments for the conditions receiving the lowest mean judgments (those with a case error), while there was essentially no relationship between reading spans and judgments for sentences with inflectional errors.[5]

In order to interpret the results of Experiment 2a as being truly contrastive with those of Experiment 1, it is necessary to rule out a skeptical interpretation. The two violations occur on different words in Experiment 2a, whereas the processing manipulations affected the processing of the same word in Experiment 1. Experiment 2b is consequently designed to evaluate what happens when the violations are triggered by the same word.

5. Experiment 2b: contemporaneous grammatical violations

5.1. PARTICIPANTS

University of California – San Diego undergraduate students (n = 24) who had not participated in Experiment 2a completed this experiment for course credit.

5.2. MATERIALS

To create conditions where multiple grammatical violations could emerge at once, we manipulated (i) the agreement between a dislocated *wh*-phrase (e.g., *which manufacturers*) and a verb heading a complement clause (e.g., *make* vs. *makes*) and (ii) the presence of the complementizer *that* at the beginning of the complement clause. The overt complementizer's adjacency to the empty embedded subject (or subject trace) position incurs a grammatical penalty (i.e., this violates the Empty Category Principle (Chomsky, 1981, 1986) or some principle with the same force). In (6d), the verb *makes* not

[5] Because of the fact that this inflectional error results from MISSING material, it is possible that the weaker acceptability effect is connected to the saliency or perceptibility of the 'error'. This possibility is a further motivation for Experiment 2b, where both GCVs follow from illicit ADDITIONAL material.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

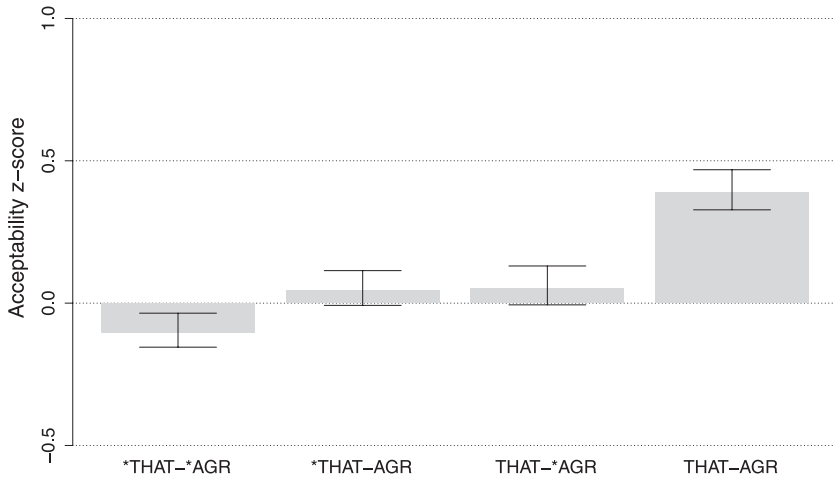

Fig. 5. Mean acceptability *z*-scores from Experiment 2b for conditions with varying processing complexity. Error bars show ±1 standard error.

only disagrees in number with *which manufacturers*, but it also triggers a *that*-trace violation ('*' indicates a violation of a grammatical constraint).

(6) a. [THAT-AGR]: I was shocked to see which manufacturers the consumer report indicated make reliable and safe automobiles.
b. [*THAT-AGR]: I was shocked to see which manufacturers the consumer report indicated that make reliable and safe automobiles.
c. [THAT-*AGR]: I was shocked to see which manufacturers the consumer report indicated makes reliable and safe automobiles.
d. [*THAT-*AGR]: I was shocked to see which manufacturers the consumer report indicated that makes reliable and safe automobiles.

Seventy-six filler items appeared along with these items.

5.3. PROCEDURE

Procedure was identical in all aspects to that used in the previous experiments. Outlier removal affected 0.69% of the data.

5.4. RESULTS

Additivity

Unsurprisingly, both types of GCV lower acceptability judgments – agreement errors and *that*-trace violations, as shown in Figure 5. These variables did not significantly interact (see Table 3).

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

TABLE 3. *Fixed effect summary for Experiment 2b*

	Coefficients	Standard Error	*t*-value
that-trace	−0.248	0.105	−2.36
Agreement	−0.240	0.070	−3.44
that-trace × Agreement	0.185	0.132	1.40
Reading span	0.009	0.028	0.31
Reading span × *that*-trace	−0.053	0.058	−0.91
Reading span × Agreement	−0.077	0.042	−1.84
Reading span × *that*-trace × Agreement	−0.064	0.078	−0.82

Individual differences

Reading span scores only show a marginal relationship to acceptability scores, according to the results. Specifically, reading span and the grammaticality of the number agreement marginally interact. This is due to the fact, as seen in Figure 6, that only when another error is present – that is, the least acceptable condition with both a *that*-trace error and an agreement error – is there a negative linear relationship between reading span scores and judgments.

5.5. DISCUSSION

As in Experiment 2a, each GCV lowered judgments and the violations combined additively when they co-occurred. This pattern emerged despite the violations being triggered by the same lexical item. In contrast to Experiment 2a, though, GCVs combined additively in this experiment – there was no strong evidence of an interaction. So, both additivity and under-additivity are possible results from combining GCVs. Of primary importance here, however, is that there is no indication that multiple GCVs lead to super-additive decrements.

This experiment also provides suggestive evidence of a relationship between reading span scores and acceptability judgments similar to the one found in Experiment 2a: individuals with higher reading span scores judge sentences with the least acceptable GCVs more harshly than individuals with lower reading span scores. Again, the critical point is that sentences with GCVs yield a pattern with reading span scores that is the opposite of what sentences with PCs demonstrate.[6]

[6] Because our focus is on the manifestation of processing limitations in acceptability judgments and not how grammatical constraints relate to judgment data, we do not dwell on the question of why some sentences with GCVs are rated lower by those with higher reading span scores. Nonetheless, we offer some speculation in the final 'General discussion' section.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

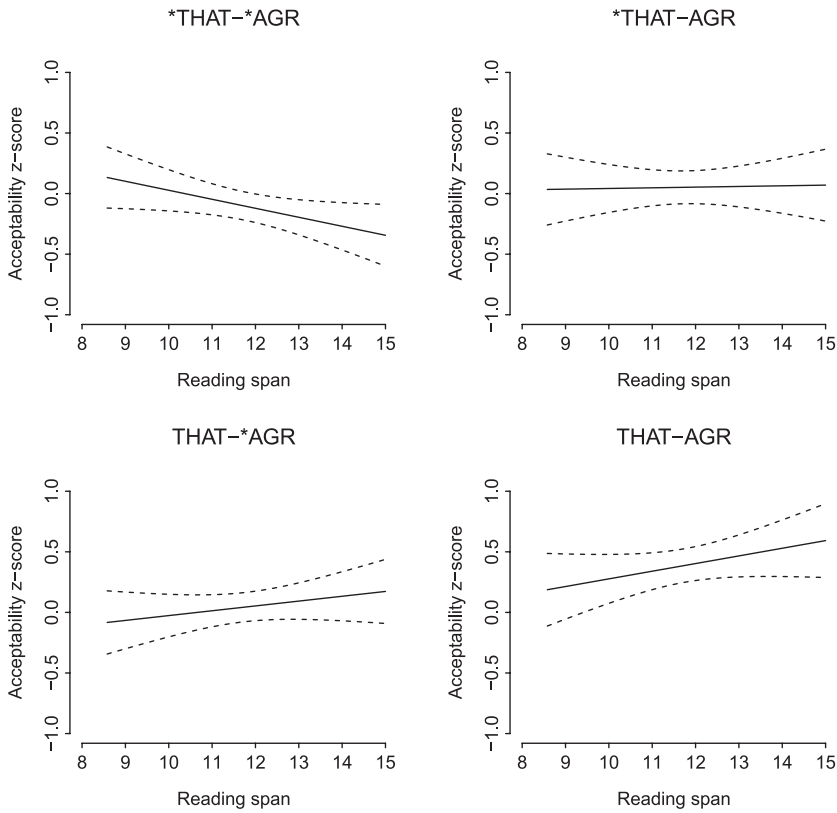

Fig. 6. Effects of reading span score on acceptability *z*-score for each condition in Experiment 2b, according to ordinary least squares regression modeling.

6. Experiment 3: grammar and processing

Because the two types of manipulations — PCs vs. GCVs — were investigated in separate experiments, the high- and low-reading span participants were different individuals across experiments and because of the limitations of making conclusions based on effects with entirely different items, Experiment 3 consequently looks at how reading span scores relate to judgments for contrasting items (PCs vs. GCVs) with the same participants and items.

6.1. PARTICIPANTS

The materials for this experiment appeared in the same session as Experiment 1 and were rated by the same thirty-two Stanford University students.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

6.2. MATERIALS

Experimental items appeared with either a correctly inflected verb (7a, 7b) or an incorrectly inflected verb (7c, 7d). Dependency locality was utilized again to vary processing difficulty; the *wh*-dependencies in (7b) and (7d) are shorter than those in (7a) and (7c) and are consequently presumed to be easier to process.

(7) a. [HARD–GOOD] They couldn't remember which lawyer that the reporter interviewed had defended the elderly man at the courthouse.
b. [EASY–GOOD] They couldn't remember which lawyer had defended the elderly man that the reporter interviewed at the courthouse.
c. [HARD–BAD] They couldn't remember which lawyer that the reporter interviewed had defending the elderly man at the courthouse.
d. [EASY–BAD] They couldn't remember which lawyer had defending the elderly man that the reporter interviewed at the courthouse.

6.3. PROCEDURE

Procedure and data analysis were the same as in the previous experiments. Removal of outliers affected 1.4% of the dataset.

6.4. RESULTS

According to the results, improperly inflected verbs significantly lower judgments (see Table 4).[7] In contrast, the effect of processing difficulty on judgments is not statistically significant; however, there is a significant interaction between processing difficulty and grammaticality. As Figure 7 illustrates, this interaction arises because processing difficulty lowers judgments in sentences without GCVs, but it does not do so in sentences with GCVs.

While reading span does not emerge as a significant predictor for judgments across all condition types, this is because the grammatical and ungrammatical conditions pattern in different ways, as displayed in Figure 8. Individuals with higher reading span scores assign relatively lower ratings to sentences with GCVs, but in the grammatical conditions, higher reading span scores are associated with higher acceptability judgments, leading to a significant interaction of reading span score and grammaticality, as shown in Table 4. In other words, reading span scores only show a positive linear relationship with judgments in the absence of GCVs.

[7] As an anonymous reviewer notes, participants may perceive these sentences as being unnatural, not because of inflectional problems, but because they parse 'defending the elderly man at the courthouse' as an NP, e.g., 'The lawyer had defending the elderly man at the courthouse on his calendar'. Whether this is the preferred or less preferred parse, it still yields an ungrammaticality as an obligatory constituent would still be missing in our items.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

TABLE 4. *Fixed effect summary for Experiment 3*

	Coefficients	Standard Error	*t*-value
Difficulty	−0.097	0.058	−1.66
Grammaticality	−0.738	0.074	−9.97
Difficulty × Grammaticality	0.310	0.082	3.79
Reading span	−0.008	0.015	−0.53
Reading span × Difficulty	−0.006	0.020	−0.29
Reading span × Grammaticality	−0.087	0.030	−2.95
Reading span × Difficulty × Grammaticality	−0.019	0.038	−0.50

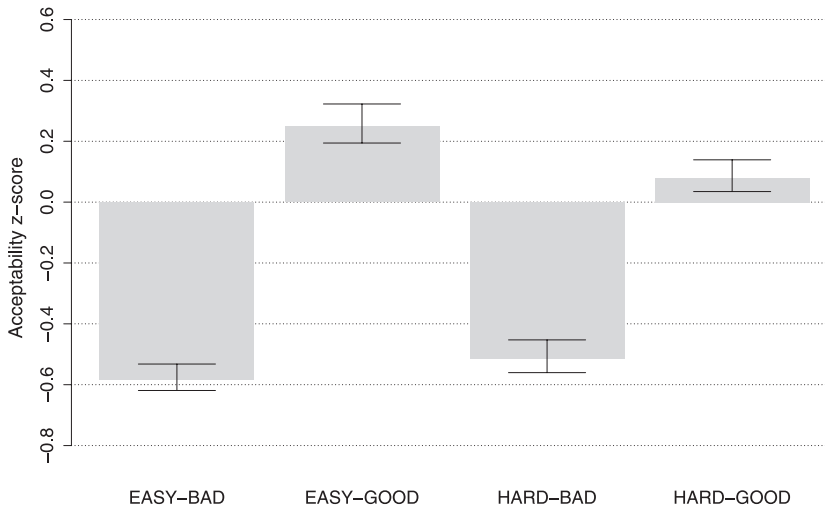

Fig. 7. Mean acceptability *z*-scores from Experiment 3.

6.5. DISCUSSION

The data show that PCs and GCVs combine under-additively: combining the two sources of unacceptability yields something less than expected on the basis of each in isolation. A likely explanation for this under-additivity is that the GCV effectively drowns out effects of the PC. In general, if a grammatical constraint does not depend on processing difficulty, combining this constraint with processing challenges should not result in super-additivity, according to the logic of the additive factors model. The present results support this hypothesis.

Echoing the findings of the previous experiments, participants with higher reading span scores find ungrammatical sentences worse, but difficult sentences better, compared to their low-span counterparts. The documentation of these contrasting effects for the same set of subjects adds support to the similar contrasts found in Experiments 1 and 2a/2b:

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

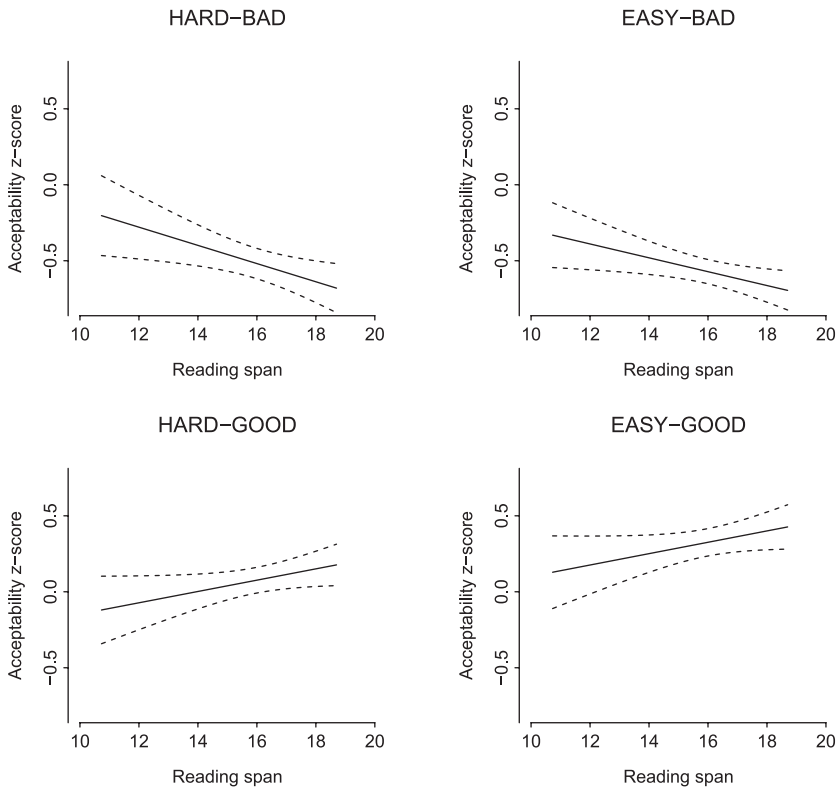

Fig. 8. Effects of reading span score on acceptability *z*-score, according to ordinary least squares regression modeling.

individuals with higher reading span scores do not indiscriminately provide higher judgments for sentences with features that lower acceptability compared to some baseline – only sentence features linked to processing difficulty trigger this relationship.

7. Experiment 4: extreme sentence processing difficulty

Based on the evidence from Experiments 1–3, reading span tests seem to systematically relate to judgments for sentences with varying degrees of processing complexity. However, these experiments do not tell us whether the observed patterns hold for sentences of all degrees of processing difficulty, particularly extreme PCs. That is, do all acceptability contrasts that emerge from processing-related sources demonstrate this sensitivity to individual characteristics like reading span? To address this question, we consider

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

constructions that give rise to severe processing difficulty, which may impose such high cognitive demands that even individuals with quite high reading spans would encounter serious parsing difficulty.

7.1. PARTICIPANTS

Twenty-eight Stanford University undergraduates, naive to the purposes of the study, received cash for their participation.

7.2. MATERIALS

The materials (n = 24) for the experiment varied in two respects: (i) the distance between a *wh*-phrase and its subcategorizing head, and (ii) the presence of either a subject or object relative clause, as illustrated in (8):

(8) a. [SHORT–SRC] Someone figured out which politician wrote that Robert bribed a reporter that trusted Nancy without thinking about it.
b. [SHORT–ORC] Someone figured out which politician wrote that Robert bribed a reporter that Nancy trusted without thinking about it.
c. [LONG–SRC] Someone figured out which politician a reporter that trusted Nancy wrote that Robert bribed without thinking about it.
d. [LONG–ORC] Someone figured out which politician a reporter that Nancy trusted wrote that Robert bribed without thinking about it.

Thus, in the long conditions, the *wh*-dependency crosses a nested object relative clause. In contrast, the dependencies are non-overlapping in the short conditions.

7.3. PROCEDURE

The procedure was identical to that used in the previous experiments. Outlier removal affected 1.49% of the data.

7.4. RESULTS

As Figure 9 depicts, higher reading span scores are associated with higher acceptability z-scores in the two relatively easy conditions with short dependencies, (8a) and (8b). But in the more difficult conditions with long dependencies, (8c) and (8d), no evidence of a relationship between reading span scores and judgments appears. This pattern accounts for the interaction between reading span scores and dependency length in the LME model of acceptability judgments (see Table 5). Such findings thus do not reveal a relationship between judgments and individual cognitive differences, despite the clear fact that it is the processing difficulty of these items that yields the low acceptability ratings.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

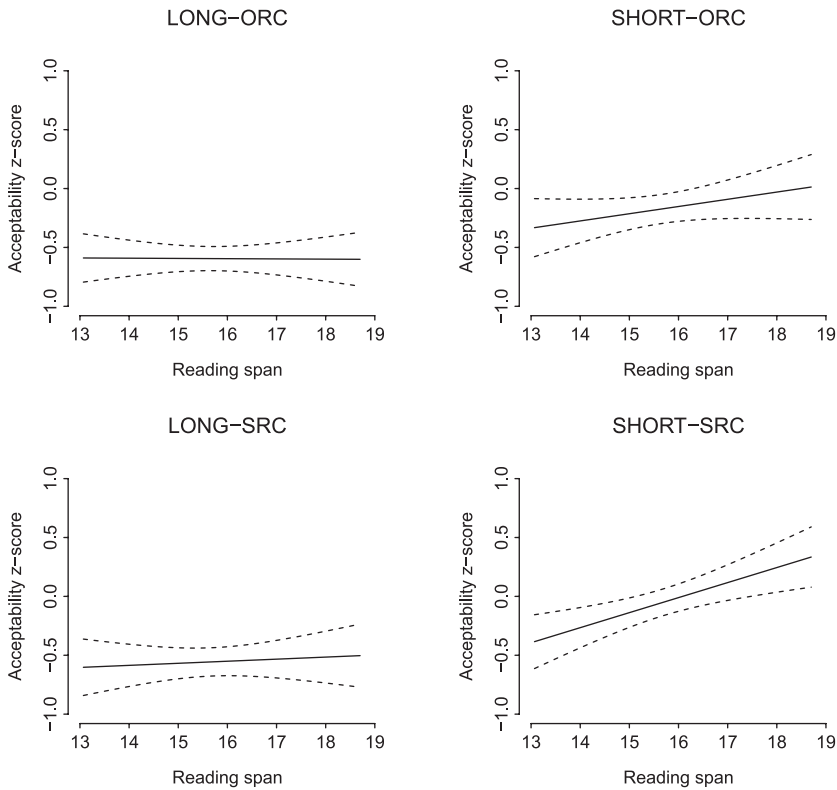

Fig. 9. Effects of reading span score on acceptability *z*-score for each condition in Experiment 4, according to ordinary least squares regression modeling.

TABLE 5. *Fixed effect summary for Experiment* 4

	Coefficients	Standard Error	*t*-value
Length	−0.465	0.075	−6.19
RC-type	−0.079	0.052	−1.52
Length × RC-type	0.076	0.106	0.71
Reading span	0.054	0.049	1.11
Reading span × Length	−0.100	0.049	−2.03
Reading span × RC-type	−0.039	0.034	−1.15
Reading span × Length × RC-type	0.026	0.069	0.38

7.5. DISCUSSION

According to the results, individuals with relatively high reading span scores do not always rate hard-to-process sentences as being more acceptable than individuals with lower reading span scores. Thus, harder-to-process sentences

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

will not necessarily show a stronger relationship with individual-level memory characteristics than easier-to-process sentences.

Items with long, syntactically complex, dependencies here seem to produce such extreme processing difficulty that individual differences have little impact. Items with short dependencies, in contrast, are comparatively easier to process, leaving room for differences due to individual variation to emerge. Thus, the relationship between memory capacity and the judgment of items with PCs may be absent at the edges of the difficulty spectrum. Judgments for trivially easy or extremely difficult items may show little relationship to individual differences because all or almost all individuals behave in a virtually identical fashion at these extremes.

In sum, memory estimates and processing difficulty are not uniformly related across all constructions in acceptability judgment tasks. Even if memory measures have the potential to identify processing limitations at work in some acceptability judgment datasets, the absence of correlations or linear relationships cannot license the conclusion that such limitations do not influence acceptability judgments.

8. General discussion

The purpose of these experimental studies was to augment our understanding of how processing limitations are reflected in acceptability judgments. To do this, we explored acceptability datasets differing with respect to the role of processing complexity in creating judgment contrasts. In Experiment 1, the critical items differed along a spectrum of processing complexity. The most difficult sentences had a syntactic structure that complicated the retrieval and integration of two key argument phrases. As compared to examples where the retrieval of only one such argument was complicated, the simultaneous demands imposed by two relatively difficult long-distance dependencies resulted in super-additive decrements in the acceptability judgment data. Variation in the judgment data for the most difficult conditions also related to performance on a reading span test: individuals who scored higher on this test provided higher ratings for the difficult items. Subsequent experiments indicated that these two effects (super-additivity and positive interactions between reading span and sentence complexity) do not extend to other datasets where the source of unacceptability can be attributed to factors other than processing complexity. In Experiments 2a and 2b, we saw that multiple grammatical errors, whether simultaneously triggered or not, failed to produce super-additive effects, and that the only significant interactions with reading span were negative, i.e., higher reading span scores were associated with lower ratings for the least acceptable sentence conditions. These patterns in the first two experiments were replicated in Experiment 3, where the two

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

sources of unacceptability combined under-additively, and showed contrasting relationships to reading span scores.

On the view that processing costs are not central to the unacceptability of sentences with GCVs, the evidence suggests that how sources of unacceptability combine and how reading span scores relate to judgments depend on the role of processing limitations in judgment variation. Evidence of super-additive interactions, together with a positive relationship between judgments and reading span scores, strongly imply that processing difficulty plays a large role in a given acceptability judgment contrast. In contrast, additive or non-additive cumulativity, together with either a flat or negative relationship between judgments and reading span scores, suggest a relatively small role for processing limitations. This is consistent with a gradient perspective of the observed findings: the greater the magnitude of super-additive interactions, and the greater the positivity of the relationship between reading span scores and judgments, the more that processing limitations play a role in the judgment data.

In sum, these data supply novel evidence that (i) processing costs can produce super-additive effects on acceptability judgments and (ii) unlike other factors that lower acceptability, processing costs can show a positive relationship to individual measures of processing resources like the reading span. Taken together, these telltale signs of processing limitations can be used as a tool in interpreting judgment data, particularly when the primary driver of acceptability contrasts is ambiguous or unknown. Like most tools, though, this one too is limited, as Experiment 4 showed. Indeed, because super-additive effects are limited to contexts where resource demands are overlapping, this particular indicator of processing limitations will apply only to a specific sort of 'hard-to-process' sentence. Relatedly, only a small subset of the many types of proposed grammatical constraints and sources of processing difficulty have been considered here. The generalizability of the present findings thus ultimately depends on whether further research confirms these findings in the consideration of other sentence types.

8.1. INDIVIDUAL DIFFERENCES AND ACCEPTABILITY JUDGMENTS

The evidence obtained here includes the finding that higher reading span scores are associated with higher judgments for complex but grammatically well-formed sentences. In Experiment 1, this trend emerged starkly in the most difficult conditions. In the easier conditions, where there was no embedding of the subject NP, there was essentially no significant variation ascribable to individual differences on the reading span task (see Figure 2). This pattern is consistent with the view that individuals with higher reading spans encounter the same absolute amount of difficulty, but generally have more or faster resources to cope with these processing costs. As a consequence of being

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

taxed less, proportionately speaking, these individuals provide higher ratings for the critical sentences.

In several respects, these findings resemble those of Just and Carpenter (1992). They also obtained reading span scores from their participants and found that the magnitude of the processing difference between subject vs. object relative clauses depended on the reading span bracket each participant fell into. In brief, they discovered an interaction between reading span group (high vs. low) and syntactic complexity, much as we observed in Experiment 1. On the other hand, the findings of Caplan and Waters (1999) and Waters and Caplan (1996) point in the other direction: Caplan and Waters were unable to replicate many of the key results of Just and Carpenter (1992), and, in other investigations, Caplan and Waters found no relationship between performance on the reading span task and judgments for complex sentences. In particular, Waters and Caplan (1996) tested how high-, medium-, and low-span participants judged the acceptability of various garden path sentences under whole-sentence visual presentation or rapid serial visual presentation in a forced-choice ('good' or 'bad') task. All groups responded more slowly and less accurately to garden path sentences, compared to non-garden path sentences. However, the magnitude of these differences did not vary across the groups. Perhaps most compellingly, individuals with severe memory impairments, as reflected by reading spans of 0 or 1, behaved no differently than control groups in the reading of subject and object relative clause sentences.

There are several possible explanations for these conflicting findings. One concerns the coding and categorization of the individual participants. Waters and Caplan employed the traditional method of scoring the reading span according to the maximum level a participant reached on the task. As Conway et al. (2005) observe, this all-or-nothing scoring strategy obliterates useful information about individual variation. Moreover, by lumping individuals into high, medium, or low working memory bins, Waters and Caplan reduce the statistical probability of finding effects. A second concerns the type or source of processing difficulty. The Waters and Caplan (1996) study cited above centers on garden path effects, where the observable processing difficulty can be attributed to reanalysis and expectations given the bottom-up input. This situation differs markedly from the cases considered where the online difficulty is standardly attributed to memory retrieval difficulty. Thus, it remains a possibility that different sources of processing difficulty have notably different signature effects on acceptability judgments and that they have different relationships to measures of memory capacity, such as the reading span task.

Lastly, although one measure of individual differences, the reading span task, relates systematically to judgments for some complex sentences, this by no means implies that similar results will obtain with other measures. As noted earlier, other researchers have looked for a relationship between memory measures and

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

judgments for particular types of sentences, and have failed to uncover any reliable patterns (Tokimoto, 2009; Sprouse et al., 2012). However, these null results are reconcilable with the view that inappropriate measures of individual differences were chosen, that the participant sample reflects an insufficiently wide range of memory spans, that the materials themselves are too extreme to allow individual differences to emerge, or that other confounds obscure the critical effects. Thus, the current results and conclusions about individual differences and their relationship to judgments apply only to the reading span task.

8.2. THE PROCESSING OF GRAMMATICAL CONSTRAINT VIOLATIONS

The strength of the conclusions made here depends on the assumption that certain sentences are unacceptable for reasons other than on-line processing difficulty. This assumption is unlikely to provoke much outcry. But it leaves the lingering question of why multiple GCVs do not combine super-additively, since there must be some mental effort used in identifying and processing them. The explanation for this depends on how one conceptualizes the link between grammatical and general cognitive constraints. Tanenhaus, Carlson, and Seidenberg (1985), for instance, state that at least some linguistic processes depend on automatic processes that are "extremely rapid, they are sealed off from awareness and not subject to strategic control, and they do not draw on processing resources" (p. 367). From this highly modular perspective, the processing of GCVs is free from the sort of limitations that affect the processing of PCs. Super-additive effects are consequently predicted to be absent, as linguistic modules recruit different kinds of cognitive processes and do not suffer from the same sort of limitations as the general system, making it effectively impossible for the module to be overtaxed.

On the contrasting view that grammar and processing difficulty lie on a continuum, the lack of super-additive effects from combining GCVs can be taken as an indication of minimal processing costs. That is, if grammatical constraints reflect highly overlearned generalizations about regularities in the structure of linguistic input based on previous experience (MacWhinney, 1998; Kemmer & Barlow, 2000; Langacker, 2000; Bod, 2006, 2009; Goldberg, 2006; Bybee, 2007), then multiple grammatical errors are unlikely to combine in a super-additive fashion simply because they are trivial to process.

Similar reasoning can be applied to the results on individual differences. It is not surprising that judgments rise for individuals with higher reading span scores if these individuals experience less sensitivity to processing costs (up to some threshold of difficulty). On either perspective outlined above, such a linear relationship is predictably absent when grammatical errors appear in the input: the assessment of grammatical constraints,

whether learned over a lifetime or part of a discrete language module, does not significantly tax general processing resources. This is either because the relevant language processing module operates without the limitations the general cognitive system abides by or because the relevant morphosyntactic regularities are highly overlearned and trivial to evaluate.[8]

8.3. CONCLUSION

It is useful to conclude with what our results do and do not show. What they tell us is how some processing-related sources of unacceptability affect judgments and relate to reading span scores, and, relatedly, they allow us to roughly gauge the extent of processing limitations in acceptability judgment variation. What they do not do, however, is tell us how to know that processing-related sources of unacceptability are NOT present in an acceptability contrast, that unacceptability follows solely from grammatical constraints, or that grammar has no role to play in a given contrast.

Even when we find super-additivity and/or a positive linear relationship between judgments and measures of memory/verbal abilities, these findings cannot definitively rule out the possibility that grammar or other factors account for some observable part of the variation in judgments. Given our current state of knowledge, it appears to be impossible to rule out grammar as a contributor to an acceptability contrast.[9] There may be cases where we

[8] We consider it probable that grammatical violations are not homogenous with respect to their associated processing costs. Some violations, for instance, may be more obvious, or more 'repairable' than others. Thus, the comments above apply specifically to the sorts of violation we have looked at here.

[9] A reviewer makes a similar point by noting that super-additive effects are ambiguous between cases of interacting demands on cognitive resources and cases where general cognitive demands additively summate but a third factor, i.e., grammar, contributes an additional source of variation (see Sprouse, 2007; Sprouse et al., 2012). We concur with this logic, but we would make several observations in this regard. To the best of our knowledge, no acceptability data exist yet that clearly portray the latter scenario, and our results would suggest that, even in such a case, the extent of processing costs on judgments would be reflected in the correlation with reading span scores. In addition, there is another way of addressing whether super-additive effects can be understood in terms of grammatical constraint effects stacking on top of functional constraint effects. On such a scenario, two sentence features, A and B, which independently lower judgments (for reasons due to processing complexity) happen to combine in such a way that the sentence now violates a grammatical constraint. In other words, the ungrammaticality and super-additivity results from the specific combination of A and B. To test whether this interpretation is sound, A and B could each be combined with other established sentence processing factors that lower judgments in a separate series of experiments to look for signs of super-additive effects on acceptability. If A and B combine with other sentence features to yield super-additive effects in judgment datasets, then it is unlikely that the original super-additive effects of combining A and B are idiosyncratic and specific to the co-presence of those two sentence features.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

have no theoretical reason to suppose that GCVs are responsible for low judgments of acceptability, e.g., center embeddings, garden path sentences, etc. But this is quite different from having empirical evidence that proves a negative – a difficult task in any scientific endeavor.

On the other hand, finding support for the role of factors unrelated to general cognitive costs is more straightforward. We would need data showing that judgment ratings get lower as processing gets easier. If evidence from processing difficulty measures and acceptability tasks shows that sentences become less acceptable in conditions where processing becomes easier (relative to some baseline condition), then online sentence processing difficulty becomes an unlikely explanation for the acceptability pattern. Assuming that grammar and processing difficulty are the primary or only candidates to explain such acceptability contrasts, then such data strengthen the case for grammatical explanations.

As an example of a study that reaches such a conclusion, Staum Casasanto and Sag (2008) find that repetition of the word *that* lowers acceptability judgments in sentences like (9):

(9) I truly wish *that* if something like that were to happen *that* my children would do something like that for me.

However, when the distance between the first and second complementizer *that* is greater, the acceptability decrement is less severe. Hypothesizing that the second *that* may have some functional value, Staum Casasanto and Sag looked at reading times for similar items. In this case, reading times at the relative clause subject (*my children*) were faster after a repeated *that* compared to sentences with only the initial *that*. These findings support the idea that the second *that* facilitates processing, but they do not parallel the acceptability findings: the cases where the extra *that* helps processing still receive lower acceptability judgments than the single-*that* cases. Hence, it makes little sense to suppose that processing effort makes sentences with a repeated *that* unacceptable. Instead, this data pattern supports the conclusion that the repeated *that* violates a grammatical principle, although the severity of the accompanying acceptability decrement can be modulated by functional considerations. In short, the same factor that introduces a grammatical error may simultaneously aid online sentence processing.

The motivation to continue with such investigations is aptly summarized by Bever (1970): "Linguistic intuitions do not necessarily reflect the structure of a language, yet such intuitions are the basic data the linguist uses to verify his grammar. This fact could raise serious doubts as to whether linguistic science is about anything at all, since the nature of the source of its data is so obscure" (p. 346). Forty years of subsequent research has unfortunately

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press

witnessed sparingly few discoveries about the factors that contribute to and shape acceptability judgments, or about how to distinguish them from one another. In response, there are many who would abandon intuitions altogether as a primary source of linguistic data. We are not among them. But at the very least, if introspective judgments of sentence acceptability are to remain part of the data linguists use to construct theories of grammar, it is essential that we explore explanations of complex patterns of graded acceptability in terms of the interaction of grammatical constraints, limitations of processing resources, and other factors outside the domain of grammar.

REFERENCES

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59** (4), 390–412.

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, **72** (1), 32–68.

Bartek, B., Lewis, R., Vasishth, S., & Smith, M. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **37** (5), 1178–1198.

Bever, T. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: John Wiley & Sons.

Bever, T., Carroll, J., & Hurtig, R. (1976). Analogy; or, ungrammatical sequences that are utterable and comprehensible are the origins of new grammars in language acquisition and linguistic evolution. In T. Bever, J. Katz, & D. T. Langendoen (Eds.), *An integrated theory of linguistic ability* (pp. 149–182). New York: Crowell.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, **23** (3), 291–320.

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, **33** (5), 752–793.

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

Caplan, D., & Waters, G. (1999). Verbal working memory and sentence comprehension. *Brain and Behavioral Sciences*, **22** (1), 77–126.

Chapman, R. (1974). *The interpretation of deviant sentences in English: A transformational approach*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A festschrift for Morris Halle* (pp. 232–286). New York: Holt, Reinhart & Winston.

Chomsky, N. (1977). On *wh*-movement. In P. Culicover, T. Wasow, & A. Akmajian (Eds.), *Formal syntax* (pp. 71–132). New York: Academic Press.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, **12** (5), 769–786.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, **19** (4), 450–466.

Daneman, M., & Hannon, B. (2007). What do working memory span tasks like reading span really measure? In N. Osaka, , R. Logie, & M. D'Esposito (Eds.), *The cognitive neuroscience of working memory* (pp. 21–42). New York: Oxford University Press.

Fanselow, G., & Frisch, S. (2004). Effects of processing difficulty on judgments of acceptability. In G. Fanselow, C. Fery, M. Schlesewsky, & R. Vogel (Eds.), *Gradience in grammar* (pp. 291–316). Oxford: Oxford University Press.

Featherston, S. (2008). Thermometer judgments as linguistic evidence. In M. Claudia & A. Rothe (Eds.), *Was ist linguistische Evidenz?* (pp. 69–89). Aachen: Shaker Verlag.

Fedorenko, E., Gibson, E., & Rohde, D. (2007). The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language*, **56** (2), 246– 269.

Ferreira, V., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28** (6), 1187–1199.

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods*, **35** (1), 116–124.

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language processing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge: Cambridge University Press.

Friedman, N., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language*, **51** (1), 136–158.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh.

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, **14** (3), 225–248.

Gleitman, L., & Gleitman, H. (1970). *Phrase and paraphrase: Some innovative uses of language*. New York: W.W. Norton and Company.

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

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, **29** (2), 261–290.

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I., & Snider, N. (2013). The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes*, **28** (1), 48–87.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, **86** (2), 366–415.

Hofmeister, P., Staum Casasanto, L., & Sag, I. A. (2012a). How do individual cognitive differences relate to acceptability judgments? A reply to Sprouse, Wagers, & Phillips. *Language*, **88** (2), 390–400.

Hofmeister, P., Staum Casasanto, L., & Sag, I. A. (2012b). Misapplying working memory tests: A reductio ad absurdum. *Language*, **88** (2), 408–409.

Hofmeister, P., Staum Casasanto, L., & Sag, I. A. (2013). Islands in the grammar? Standards of evidence. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and the islands debate*, (pp. 42–63). Cambridge: Cambridge University Press.

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, **99** (1), 122–149.

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

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

King, J., & Just, M. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, **30** (5), 580–602.

King, J., & Kutas, M. (1995). Who did what and when? Using word-and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience*, **7** (3), 376–395.

Kluender, R. (1992). Deriving islands constraints from principles of predication. In H. Goodluck & M. Rochemont (Eds.), *Island constraints: Theory, acquisition and processing* (pp. 223–258). Dordrecht: Kluwer.

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., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes*, **8** (4), 573–633.
Langacker, R. (2000). A dynamic usage-based model. In S. Kemmer & M. Barlow (Eds.), *Usage-based models of language* (pp. 1–63). Stanford, CA: CSLI.
MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, **49** (1), 199–227.
Miller, G. (1975). Some comments on competence and performance. *Annals of the New York Academy of Sciences*, **263** (1), 201–204.
Miller, G., & 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.
Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, **116** (2), 220−244.
Pashler, H., & Johnston, J. (1998). Attentional limitations in dual-task performance. In H. Pashler (Ed.), *Attention* (pp. 155–189). Hove: Taylor & Francis.
Phillips, C. (2006). The real-time status of island phenomena. *Language*, **82** (4), 795–823.
Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.
Pritchett, B. (1992). *Grammatical competence and parsing performance*. Chicago: University of Chicago Press.
Pylyshyn, Z. (1973). The role of competence theories in cognitive psychology. *Journal of Psycholinguistic Research*, **2** (1), 21–50.
Ross, J. R. (1967). *Constraints on variables in syntax* (Unpublished doctoral dissertation). MIT, Cambridge, MA. [Published in 1986 as *Infinite Syntax!* by Ablex, Norwood, NJ.]
Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, **115** (11), 1497–1524.
Sprouse, J. (2007). *A program for experimental syntax* (Unpublished doctoral dissertation). University of Maryland, College Park.
Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, **40** (2), 329–341.
Sprouse, J., & Hornstein, N. (Eds.) (2013). *Experimental syntax and island effects*. Cambridge: Cambridge University Press.
Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working memory and syntactic island effects. *Language*, **88** (1), 82–123.
Staum Casasanto, L., & Sag, I. A. (2008). The advantage of the ungrammatical. In B. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the* 30*th Annual Meeting of the Cognitive Science Society* (pp. 601–606). Austin, TX: Cognitive Science Society.
Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, **57** (4), 421–457.
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.
Tokimoto, S. (2009). *Island phenomenon in Japanese and working memory: Syntactic constraints independent from working memory constraints*. Poster presented at the 22nd Annual CUNY Sentence Processing Conference.
Towse, J., Hitch, G., & Hutton, U. (2000). On the interpretation of working memory span in adults. *Memory & Cognition*, **28** (3), 341–348.
Vos, S., Gunter, T., Schriefers, H., & Friederici, A. (2001). Syntactic parsing and working memory: The effects of syntactic complexity, reading span, and concurrent load. *Language and Cognitive Processes*, **16** (1), 65–103.

Waters, G. S., & Caplan, D. (1996). Processing resource capacity and the comprehension of garden path sentences. *Memory & Cognition*, **24** (3), 342–355.
Watt, W. (1970). On two hypotheses concerning psycholinguistics. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 137–220). New York: John Wiley & Sons.
Welford, A. (1952). The 'psychological refractory period' and the timing of high-speed performance – a review and a theory. *British Journal of Psychology. General Section*, **43** (1), 2–19.
Whitney, P., Arnett, P., Driver, A., & Budd, D. (2001). Measuring central executive functioning: What's in a reading span? *Brain and Cognition*, **45** (1), 1–14.

https://doi.org/10.1017/langcog.2013.7 Published online by Cambridge University Press