How do individual cognitive differences relate to acceptability judgments? A reply to Sprouse, Wagers, & Phillips

Philip Hofmeister
University of Essex

Laura Staum Casasanto
Stony Brook University

Ivan A. Sag
Stanford University

Abstract

Sprouse, Wagers, & Phillips (in press) carried out two experiments in which they measured individual differences in memory to test processing accounts of island effects. They found that these individual differences failed to predict the magnitude of island effects and construe these findings as counterevidence to processing-based accounts of island effects. Here, we take up several problems with their methods, their findings, and their conclusions.

First, the arguments against processing accounts are based on null results using tasks that may be ineffective or inappropriate measures of working memory (the *n*-back and serial recall tasks). The authors provide no evidence that these two measures predict judgments for other constructions that are difficult to process and yet are clearly grammatical. They assume that other measures of working memory would have yielded the same result, but provide no justification that they should. We further show that whether a working memory measure relates to judgments of grammatical, hard-to-process sentences depends on how difficult the sentences are. In this light, the stimuli used by the authors present processing difficulties other than the island violations under investigation and may have been particularly hard to process. Second, the Sprouse et al. results are statistically in line with the hypothesis that island sensitivity varies with working memory. Three out of the four island types in their Experiment 1 show a significant relation between memory scores and island sensitivity, but the authors discount these findings on the grounds that the variance accounted for is too small to have much import. This interpretation, however, runs counter to standard practices in linguistics, psycholinguistics, and psychology.*

Acceptability judgments are inherently ambiguous. Judging a sentence's acceptability potentially involves the assessment of its syntactic well-formedness, the online processing difficulty encountered while parsing it, as well as other factors (Chomsky & Miller, 1963; Miller & Chomsky, 1963; Bever, 1970; Schütze, 1996; Cowart, 1997; Staum Casasanto *et al.*, 2010; Hofmeister *et al.*, in press). Hence, it is unsurprising that there is disagreement about the explanation for some acceptability contrasts. It is even less surprising when one considers that there are no agreed upon diagnostics distinguishing grammatically-based acceptability contrasts from those based on processing cost.

Sprouse *et al.* (in press; hereafter, SWP) suggest that the relationship between measures of working memory and acceptability judgments may be one such diagnostic. They attempt to use it to resolve a long-standing debate regarding the source of island effects. The term "island effects" here refers to the low acceptability ratings given to sentences with a dependency between a phrase and a syntactic position inside select syntactic environments. For instance, the *wh*-element, *what*, in (1) is linked to an argument position inside a complex noun phrase (in brackets):

(1) What did Jim repeat [the rumor that Spock loved ___]?

Ross (1967) lists an array of such syntactic environments that block dependency formation. Since that time, the associated island constraints have been entrenched in linguistic theory as instances of universal grammatical principles and have acted as standard tests for overt and covert phrasal movement (Chomsky, 1973, 1977, 1981, 1986).

Numerous researchers, however, have noted that island violations often co-occur with features known to produce processing difficulty (Kuno, 1973; Deane, 1991; Kluender, 1991, 1998, 2005; Kluender & Kutas, 1993; Hofmeister, 2007; Hofmeister *et al.*, 2007; Sag *et al.*, 2007; Hofmeister & Sag, 2010), such as long-distance dependencies spanning multiple new discourse referents, opportunities for garden-pathing or misanalysis, syntactically and semantically similar discourse referents, vague or non-specific filler phrases, etc. For instance, the dependency in (1) not only spans multiple discourse references, but the parser is likely to attempt to integrate the *wh*-item at the first verb (*repeat*), leading to reanalysis when the subsequent NP is encountered, or even later. Notably, the island violation is less severe when *what* is replaced with *who*, which is explainable on the hypothesis that *who* is not a particularly plausible argument of *repeat*. These facts, and others like them, motivate the alternative hypothesis that island effects follow from the increase in processing difficulty that such constructions engender directly and typically co-occur with in examples found in the literature.

SWP set out to evaluate such a processing-based theory of islands, what they label a "reductionist" account[1], by examining how individual differences in processing resources relate to acceptability

judgments. The intuition at play is a reasonable one: if judgments of island-violating sentences reflect processing limitations, then individuals with more processing resources should be less likely to "run out" of resources while processing a sentence with an island violation. The acceptability contrast between island-violating and minimally different, non-violating sentences should accordingly be smaller for such individuals than it is for those with fewer language processing resources.

To assess the relationship between individual working memory (WM) resources and island sensitivity, SWP calculated differences-in-differences (DD) scores for each participant. These scores are calculated by first subtracting a participant's mean acceptability judgment for sentences with a dependency into an island (2b) from their mean judgment for sentences with a dependency into an embedded, non-island structure (2a). This difference is termed D1. Next, the participant's mean judgment for sentences with an island structure but which the *wh*-dependency does not enter (2d) is subtracted from the mean for sentences with an embedded non-island constituent and a *wh*-dependency that does not enter into the constituent (2c). This difference, termed D2, is subtracted from D1 to yield the overall DD score.

(2) a. What do you think that John bought?
b. What do you wonder whether John bought?

c. Who thinks that John bought a car?
d. Who wonders whether John bought a car?

In essence, these scores reflect "how much greater the effect of an island structure is in a long-distance dependency sentence than in a sentence with a local dependency" (Sprouse *et al.*, in press). The main findings from SWP show that how respondents performed on the memory assessment tasks accounted for only 0-6% of the overall variance in the magnitude of island effects (DD scores). The authors interpret these facts, whose statistical reliability is well-established, as counter to the predictions of a processing-based perspective of island effects.

There are several notable obstacles, however, that stand in the way of interpreting the results as the authors do, which we briefly summarize below:

The Relationship Between Individual Working Memory Measures and Acceptability Judgments

i. The null results leave open the possibility that the authors have selected an inappropriate WM measure, as the authors acknowledge;

ii. There is no extant data on how the chosen memory measures relate to acceptability judgments for uncontroversially hard-to-process sentences of English;

Interpretation of Goodness-of-Fit Statistics (R^2)

> The authors assume that resource limitation theories must account for some arbitrarily large amount of overall variance above that found in their studies. Because that threshold is not reached, they interpret as "not particularly meaningful" the fact that 3 out of the 4 island types in their Experiment 1 (in the data of participants who show island effects at all) exhibit a statistically significant relationship between their memory measures and island sensitivity, as predicted by a resource limitation theory.

We will discuss both of these issues in turn.

To assess working memory, SWP use two different measures: the *n*-back task and the serial recall task. In the *n*-back task, participants see sequences of letters or pictures and must answer whether the current stimulus matches with the stimulus seen *n* turns before. SWP's version of the serial recall task asks participants to recite back the same 8 words in the precise order in which they were presented, using a total of 10 different orderings. Besides these two simple span tests (simple because there is no secondary task involved), there exist numerous other measures of working memory, including complex memory span tasks, which require participants to store stimuli while performing a highly distracting secondary task (Daneman & Carpenter, 1980). For instance, in the reading span task, participants read a series of sentences and must store and eventually recite the final word from each sentence following the final sentence in the sequence. This task has been used by numerous researchers in investigations of linguistic processing and performance on the task is highly predictive of reading and listening comprehension skill, as well as general fluid intelligence (Daneman & Carpenter, 1980; King & Just, 1991; Whitney *et al.*, 1991; Just & Carpenter, 1992; Daneman & Merikle, 1996; Whitney *et al.*, 2001; Friedman & Miyake, 2004, *inter alia)*.

Clearly, the validity of the SWP findings hinges on the choice of WM measure. SWP explicitly acknowledge that "it is logically possible that a different capacity measure could be found that does indeed correlate with the acceptability of island effects" but counter that "many working memory measures share common variance" and similarly that "there is a large component of shared variance between simple span tasks and complex span tasks", citing Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle (2005). They subsequently cite evidence from Kane *et al.* (2007) that the *n*-back and serial recall task results are uncorrelated, and take this to mean that "the likelihood of finding a new measure that correlates with neither is very small indeed." They conclude that the choice of another WM task is unlikely to change the results, even if the task was a complex span task: "The two tasks we have chosen jointly correlate with most other popular WM tasks ... this does not eliminate the possibility that we did not test the correct component of the working memory system. However it does substantially decrease the likelihood of this error, as it is highly unlikely that any

other extant working memory task would yield results that are significantly different than the results of the two tasks employed here." They assume, therefore, that a different memory measure would be highly likely to correlate with either the results from the *n*-back or serial recall task, and hence produce similar findings.

There are several problems with this reasoning. First, Conway et al. (2005) only show that scores from three complex span tasks are highly correlated: reading span, counting span, and operation span tasks. None of these WM tasks are used by SWP. In fact, there is reason to question whether the *n*-back and serial recall tasks should be considered WM tasks at all as they do not involve a secondary processing task. Kane *et al.* (2004, p. 190), for instance, treat such simple span tasks as measures of short-term memory, rather than working memory tasks, noting that complex span tasks like the reading span have "strongly predicted comprehension abilities in ways that simple short-term memory (STM) storage tasks did not". Similarly, in their only mention of *n*-back tasks, Conway et al. (2005, pp. 780-1) note that *n*-back tasks "present quite different cognitive demands" from complex span tasks and that "the *n*-back task may be a more appropriate indicator of the construct measured by STMC [short term memory capacity], rather than by WMC [working memory capacity] tasks". Conway et al. (2005, p. 780) also observe that correlations between the three above-mentioned complex span tasks range from .40 to .60, "suggesting that they are indeed tapping some common process or ability but also suggesting that they are not identical." So even for these "highly correlated" complex span tasks, one cannot presume that identical or even qualitatively similar results would necessarily follow from substituting one task for the other. Consequently, it does not follow that the choice of some alternative WM measure would yield identical or even highly similar results.

This is a critical issue, illustrated by the following example. Assume that the amount a person smokes (inversely) correlates with the amount of time they exercise on a weekly basis, but that time spent exercising fails to correlate strongly with dental hygiene. It would be imprudent to conclude that because smoking rates and exercise are correlated, that smoking habits will similarly fail to show a strong relationship to dental hygiene. Generally speaking, even if there are significant correlations between WM measures, one cannot strongly conclude that they will produce qualitatively similar results with a given task: it depends on what aspect of cognition (e.g. attention, inhibition of distractors, strategic encoding, etc.) they commonly capture, on what aspects they capture independently, and on what aspects of cognition the critical task itself taps.

A final note on SWP's choice of memory measures: the argument they present is essentially that the combination of the *n*-back task and serial recall task captures every relevant component or aspect of WM. Although it is true that simple span tasks like the *n*-back and serial recall task share common variance with multiple other memory tasks, including some complex span tasks, it does

not follow that these two tasks are sufficient by themselves to capture all components of variance in tasks that recruit WM resources. A central point of Kane *et al.* (2004) is that, even though simple span tasks and complex span tasks share substantial variance with each other, they are nevertheless "separable", differing in their ability to tap domain-general cognitive processes and their ability to predict general fluid intelligence. In other words, the two simple span tasks chosen by SWP are ultimately distinct from other memory measures, even if they tap some common properties. Indeed, memory researchers would be well-advised to abandon all other memory measures if these two were sufficient by themselves to account for all differences in verbal working memory. We therefore consider it improbable that these two combined memory measures alone capture any and all variation due to differences in working memory resources.

Beyond this issue, SWP's conclusions depend upon a critical assumption about how WM measures generally relate to acceptability judgments. Specifically, they assume that their WM measures can predict ratings for sentences that impose varying degrees of processing difficulty: respondents with more resources, as measured by the *n*-back and serial recall tasks, should show less sensitivity to manipulations of processing difficulty. While plausible in theory, there is no evidence that scores from the *n*-back and serial recall tasks strongly correlate with judgments for any sentences with clear processing difficulty, let alone with the sources of processing difficulty that processing-based accounts attribute to island structures. In other words, they may have chosen WM measures that do not actually correlate with processing-based acceptability decrements for *any* type of sentence structure. If judgments for sentences with varying degrees of uncontroversial processing difficulty show no relationship to WM scores, then there is little reason to suspect that judgments for sentences containing island violations would either. Moreover, even if their measures correlated with judgments on some hard-to-process constructions, not all types and degrees of processing difficulty will necessarily show the same relationship to measures of memory, particularly if not all sources of sentence processing difficulty are memory-related.

Is this merely a theoretical obstacle, or is there reason to doubt the assumed correlation between judgments for sentences of high difficulty and individual WM measures? In a series of recent experiments, Hofmeister *et al.* (to appear) examine how acceptability judgments for sentences with varying degrees of processing difficulty relate to individual assessments of working memory. WM capacity was assessed by a reading span task along the lines of that described in Daneman & Carpenter (1980).

In one of these studies, the critical items varied in terms of two factors known to affect processing difficulty: dependency length and relative clause type (subject- vs. object relative clause). Research by Gibson & colleagues has established that increased dependency length increases processing difficulty, due to storage of syntactic predictions and retrieval costs (Chen *et al.*, 2005; Gibson, 1998,

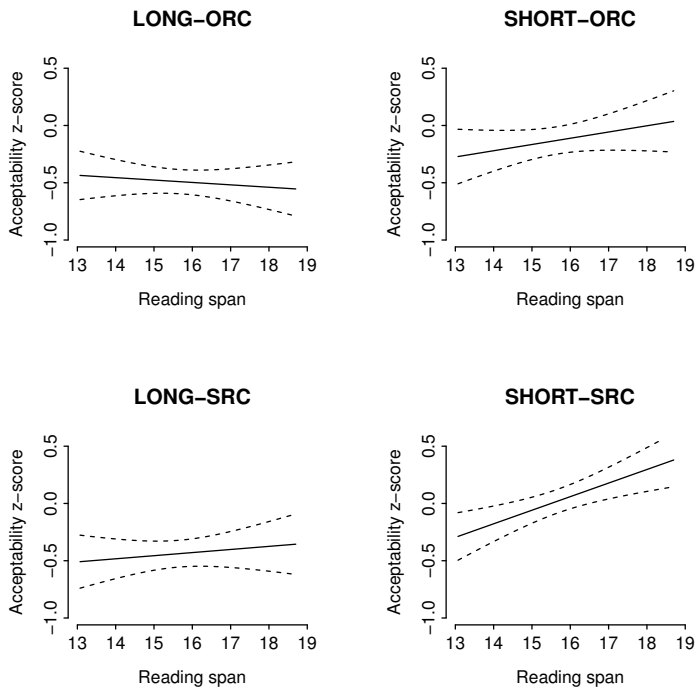

Figure 1: Linear estimates of the relationship between reading span and acceptability z-scores for sentence types with varying degrees of difficulty. Dotted lines show 95% confidence intervals.

2000; Grodner & Gibson, 2005). Object relative clauses are also well-known to impose more processing difficulty than subject relative clauses (King & Just, 1991; Just & Carpenter, 1992; King & Kutas, 1995). An example item is shown below:

(3) a. Someone figured out which politician wrote that Robert bribed a reporter that trusted Nancy without thinking about it. [SHORT-SRC]

b. Someone figured out which politician wrote that Robert bribed a reporter that Nancy trusted without thinking about it. [SHORT-ORC]

c. Someone figured out which politician a reporter that trusted Nancy wrote that Robert bribed without thinking about it. [LONG-SRC]

d. Someone figured out which politician a reporter that Nancy trusted wrote that Robert bribed without thinking about it. [LONG-ORC]

As Figure 1 depicts, higher scores on the reading span task are associated with higher acceptability z-scores in the two relatively easy conditions with short dependencies, (3a) & (3b). But in the difficult conditions with long dependencies, (3c) & (3d), there is no evidence of a relationship

between span scores and judgments. This pattern accounts for the significant interaction between WM and dependency length in the linear mixed effects model of acceptability judgments (see Hofmeister *et al.* (to appear) for details). In addition, the R^2 between z-score and reading span score in the conditions with long dependencies (or high processing difficulty) is just .012. Individual cognitive differences thus account for little variation in judgments, despite the clear fact that it is the processing difficulty of these items that yields the low acceptability ratings.

What this demonstrates is that the relationship between acceptability judgments and at least some measures of WM is not straightforward. Some sentences may be so hard to process that even individuals with high spans experience a resource shortfall or processing breakdown. However, sentences with mild to moderate processing difficulty leave room for individual differences to emerge in the acceptability ratings (for further evidence of this, see Staum Casasanto *et al.* (2010)). The point here is not that the reading span task necessarily taps the resources relevant for island processing. Rather, increased processing resources, as measured by span tasks, does not uniformly decrease sensitivity to difficulty manipulations across all sentence types. The same WM measure can be shown to relate to variation in judgments for some sentence types, but not others, particularly when these latter sentences are extremely hard to process. Thus, if islands represent extreme cases of processing difficulty (the central claim being made by processing-based accounts), individual WM differences may be irrelevant for judgments of acceptability.

This point is particularly relevant given the nature of the stimuli used by SWP. For the critical items,[2] the sentence structures had confounding sources of difficulty beyond the island structure itself. A representative island-violating stimulus is *What do you wonder whether John bought?* This is a decontextualized direct question with an odd pragmatic import, i.e. it is difficult to imagine why someone would ever ask this question, it has a vague *wh*-filler, a referential NP (*John*) with no discourse antecedent, and an opportunity for misanalysis at the verb *wonder* where the parser may strategically attempt to integrate the *wh*-filler, leading to a potentially costly reanalysis phase at the next word. According to the item lists, all experimental items contained similar hindrances to processing. Consider, by contrast, the much more natural attested examples of *wh*-island violations from the internet:

(i) While there are many beneficial aspects to child adoption, there are a number of disadvantages that you should consider and decide whether you are comfortable with before committing time, energy and resources to the process.[3]

ii) Insul-knife is one of those time-saving tools that you will wonder how you ever lived without.[4]

Note that in the attested examples, the only references to individuals between the filler and its gap site are made with the high accessibility-marking second person pronoun *you*. Moreover, these

real life examples do not pose pragmatically odd questions with non-specific *wh*-words. One of the primary points of Hofmeister & Sag (2010), in this regard, is that prior research on island effects has not systematically controlled for factors affecting acceptability that are orthogonal to the island structure, i.e. many prior examples in the literature have features besides the island structure that lower acceptability judgments. The unnecessary complexity of their items may thus be yet another factor in the null results of SWP. This point highlights the danger of drawing strong conclusions on the basis of null results: even if the experiment has sufficient statistical power to find an effect, the null results may plausibly stem from design features in the materials.

To establish that individual WM differences truly matter for judgments of island violations, we need a measure that shows a reliable relationship to judgments for sentences with uncontroversial difficulty of the appropriate sort. In the case of island violating sentences, Hofmeister & Sag (2010) point to a number of distinct, complicating factors in the processing of islands. It is not any of these factors alone that produces island effects, but the simultaneous pressure of multiple processing demands, as in the case of Complex Noun Phrase Constraint violations:

> ". . . a dependency into a complex noun phrase in English requires processing at least three nominal references inside the dependency and crossing a clause boundary:
>
> (46) Which politician did you$_1$ read reports$_2$ that we$_3$ had impeached?
>
> Added to this, a syntactic ambiguity arises after processing *reports that*. At this point, both a sentential complement and a relative clause parse (as in *Which politician did you read reports [$_{RC}$ that we had written] in front of?*) are theoretically possible. Even if one parse is more likely from a top-down perspective, these parses may nevertheless compete with one another"

Accordingly, to know whether a given WM measure should predict sensitivity to CNPC violations, it would be necessary to look at uncontroversial examples that have similar features. Garden-path sentences, for instance, are characterized by the high potential for misanalysis. Theoretically, combining such constructions with multiple discourse references would yield examples that resemble CNPC constructions. If judgments for garden path sentences with high reference processing costs (and even better, a clause boundary) fail to show any relationship to performance on WM tasks, there is little reason to suspect something different for island-related judgments. Indeed, Waters & 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. This provides cautionary evidence that

standard measures of WM may not adequately capture differences in the processing and judging of garden path sentences, and to the extent that some island types share similar features, these measures may be similarly limited in the domain of island effects.

The second major obstacle to interpreting the SWP findings as the authors do relates to their reliance on R^2 values as a means of hypothesis testing, rather than p-values. In their first experiment, SWP find a significant negative linear relationship ($p < .05$) between DD scores for 3 out of the 4 island types and the participants' WM estimates for the subset of participants with positive DD scores. That is, these participants (more than 80% of the total sample in Experiment 1) found that the negative (acceptability-lowering) effect of an island structure (e.g. a *wh*-island) on judgments was greater when a dependency entered into it than when the dependency did not. Among these participants, the effect of islandhood was significantly weaker for individuals with higher spans. This evidence, however, is not taken as support for a processing-based theory, or a refutation of a grammar-based theory. Rather, SWP conclude that, "because the goodness-of-fit of the lines is so low, these results are not particularly meaningful." That is, because the WM estimates only account for a small amount of variance, the fact that t-tests indicate the statistical significance of the linear fit does not constitute good evidence for a processing-based account of islands.

This argument critically assumes that a theory can be proved or disproved on the basis of how much overall variance it accounts for. In this respect, SWP say that

> "[u]nlike p-values, there are no broadly agreed-upon conventions for interpreting R^2 values; however, it is safe to assume that the extremely small R^2 values found for each of the island types (even after removing potentially noisy DD scores) are not at all what one would predict under a theory like the resource-limitation theory, which relies heavily on a single factor for its explanatory power."

Again, there are numerous reasons to question this interpretation of the results.

First, WM estimates may account for little overall variance, even in situations where processing difficulty clearly plays an integral role in judgment patterns. This is evident from the Hofmeister *et al.* (to appear) experiment described above, as well as the Waters & Caplan (1996) results. Processing difficulty is clearly the reason why examples like (3c) & (3d) are judged to be unacceptable, yet the measure of WM – the reading span — fails to show any relationship to the corresponding judgments.

Second, while some experimental manipulations in psycholinguistic studies may account for a large amount of variance, this is by no means standard and does not act as a benchmark for reporting significance in any linguistics, psycholinguistics, or psychology journal. Often, the lion's share of variance in acceptability and psycholinguistic studies can be attributed to differences in participants and items, not the manipulation of interest.

Third, the reasoning in the passage cited above assumes that the memory measures they have chosen adequately capture variation in resource limitations, particularly as they apply to the processing of island structures. But as we have already seen, there is no evidence that the n-back and serial recall task scores correlate with judgments for any hard-to-process sentence structures in English. The low R^2 values could thus reflect the fact that the two memory tasks only weakly overlap with the processing of island violations in terms of cognitive resources and processes. In general, no WM measure has been claimed to perfectly capture differences in verbal working memory, so any R^2 for the amount of overall variance of acceptability scores explained by a WM measure will be based on imperfect snapshots of the participants.

Fourth and most importantly, no matter what the accompanying R^2 value is, there are no established grounds in the scientific community for claiming that statistically significant tests lack meaning. Their meaning is, in fact, unambiguous: the predictors of interest have a statistically reliable impact on the dependent variable at the α-level of .05 used in these studies, i.e. there is a probability of .05 or less that the finding of significance is spurious. SWP's conclusions, therefore, depend on post hoc assumptions about how to interpret R^2 values and the puzzling assertion that some statistically significant effects are not meaningful. Indeed, on the standard method of hypothesis testing (using p-values rather than R^2 values as the arbiter of statistical significance), the SWP results argue, if anything, in favor of a processing-based account.

The absence of clear insights into how WM differences relate to acceptability judgments therefore makes it difficult, if not impossible, to draw meaningful conclusions from the SWP findings. We have suggested several reasons why no relationship between these variables is evident in the case of island-violating structures. First, the null effects established with the use of the n-back and serial recall tasks may be dependent upon the use of those memory measures. The arguments adduced in favor of their adequacy rely on the unwarranted assumption that correlated measures of memory will return qualitatively similar results. Furthermore, for any measure that we wish to use to draw conclusions about grammar vs. processing, there needs to be some precedent showing that differences in this measure can predict variation in judgments for sentences of the appropriate processing difficulty. We have shown that for at least one measure (reading span), performance on the task only correlates with judgments for items which are moderately difficult; it does not correlate with items whose reduced acceptability is attributable to extreme processing difficulty. Third, the very evidence which SWP present is at least partially in line with the hypothesis that island sensitivity will vary with individual differences in resource limitations (as measured by the n-back and serial recall tasks).

In sum, we agree that finding diagnostics for separating the effects of grammatical and processing constraints on acceptability judgments is a necessary and worthwhile enterprise in linguistics.

However, in order to support inferences about controversial cases, these diagnostics must be validated in terms of cases where little controversy exists. Only once this is done can we begin to properly understand how grammar and processing differ in their effects on acceptability judgments.

Notes

*For stimulating feedback and discussion of the ideas presented here, we thank Daniel Casasanto, Herb Clark, Ted Gibson, and Tom Wasow. The usual provisos apply.

[1]The term "reductionist" suggests a perspective where a complex phenomenon is simplified to the point of minimizing it or missing important details. Accordingly, we opt for more neutral terms like "processing-based accounts" or "emergent accounts" of island effects.

[2]Items are available at http://www.socsci.uci.edu/ jsprouse/

[3]http://www.ehow.com/info_7854078_disadvantages-child-adoption.html accessed 20 March 2012

[4]http://www.cepcotool.com/insulknife/ accessed 30 Dec 2011

References

Bever, Thomas. 1970. The cognitive basis for linguistic structures. *Cognition and the Development of Language*, ed. by John R. Hayes, 279–362. New York: John Wiley & Sons.

Chen, Evan, Edward Gibson, & Florian Wolf. 2005. Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language* 52.144–169.

Chomsky, Noam. 1973. Conditions on transformations. *A Festschrift for Morris Halle*, ed. by Stephen Anderson & Paul Kiparsky, 232–286. New York: Holt, Reinhart & Winston.

Chomsky, Noam. 1977. On *wh*-movement. *Formal Syntax*, ed. by Peter Culicover, Thomas Wasow, & Adrian Akmajian. New York: Academic Press.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, Noam. 1986. *Barriers*. Cambridge: MIT Press.

Chomsky, Noam, & George Miller. 1963. Introduction to the formal analysis of natural languages. *Handbook of Mathematical Psychology*, ed. by R. Duncan Luce, Robert R. Bush, & Eugene Galanter, 269–321. New York: Wiley.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods in Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

Daneman, Meredyth, & Patricia Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior* 19.450–466.

Daneman, Meredyth, & Philip M. Merikle. 1996. Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review* 3.422–433.

Deane, Paul. 1991. Limits to attention: A cognitive theory of island phenomena. *Cognitive Linguistics* 2.1–63.

Friedman, Naomi P., & Akira Miyake. 2004. The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General* 133.101–135.

Gibson, Edward. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68.1–76.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, ed. by Alec Marantz, Yasushi Miyashita, & Wayne O'Neil, 95–126. Cambridge: MIT Press.

Grodner, Daniel, & Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science* 29.261–290.

Hofmeister, Philip, 2007. *Representational Complexity and Memory Retrieval in Language Comprehension*. Stanford, CA: Stanford University dissertation.

Hofmeister, Philip, T. Florian Jaeger, Inbal Arnon, Ivan A. Sag, & Neal Snider. in press. The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes* .

Hofmeister, Philip, T. Florian Jaeger, Ivan A. Sag, Inbal Arnon, & Neal Snider. 2007. Locality and accessibility in *wh*-questions. *Roots: Linguistics in Search of its Evidential Base*, ed. by Sam Featherston & Wolfgang Sternefeld, 185–206. Berlin: Mouton de Gruyter.

Hofmeister, Philip, & Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language* 86.366–415.

Hofmeister, Philip, Laura Staum Casasanto, & Ivan A. Sag. to appear. Islands in the grammar? Standards of evidence. *Experimental Syntax and Island Effects*, ed. by Jon Sprouse & Norbert Hornstein. Cambridge: Cambridge University Press.

Just, Marcel, & Patricia Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 99.122–149.

Kane, Michael J., Andrew R.A. Conway, Timothy K. Miura, & Gregory J.H. Colflesh. 2007. Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.615–622.

Kane, Michael J., David Z. Hambrick, Stephen W. Tuholski, Oliver Wilhelm, Tabitha W. Payne, & Randall W. Engle. 2004. The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General* 133.189–217.

King, Jonathan, & Marcel Just. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language* 30.580–602.

King, Jonathan W., & Marta Kutas. 1995. Who did what and when? Using word-and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience* 7.376–395.

Kluender, Robert, 1991. *Cognitive Constraints on Variables in Syntax*. La Jolla, CA: University of California, San Diego dissertation.

Kluender, Robert. 1998. On the distinction between strong and weak islands: A processing perspective. *Syntax and Semantics 29: The Limits of Syntax*, ed. by Peter Culicover & Louise McNally, 241–279. San Diego, CA: Academic Press.

Kluender, Robert. 2005. Are subject islands subject to a processing account? *Proceedings of the 23rd West Coast Conference on Formal Linguistics*, ed. by Vineeta Chand, Ann Kelleher, Angelo J. Rodríguez, & Benjamin Schmeiser, 475–499. Somverville, MA: Cascadilla Press.

Kluender, Robert, & Marta Kutas. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8.573–633.

Kuno, Susumu. 1973. Constraints on internal clauses and sentential subjects. *Linguistic Inquiry* 4.363–385.

Miller, George A., & Noam Chomsky. 1963. Finitary models of language users. *Handbook of Mathematical Psychology, Volume 2*, ed. by R. Duncan Luce, Robert R. Bush, & Eugene Galanter, 419–492. New York: Wiley.

Ross, John R., 1967. *Constraints on Variables in Syntax*. Cambridge, MA: MIT dissertation. Published in 1986 as *Infinite syntax!* by Ablex, Norwood, N. J.

Sag, Ivan A., Philip Hofmeister, & Neal Snider. 2007. Processing complexity in Subjacency violations: The complex noun phrase constraint. *Proceedings of the 43 Annual Meeting of the Chicago Linguistic Society*, ed. by Malcolm Elliott, James Kirby, Osamu Sawada, Eleni Staraki, & Suwon Yoon, 215–229, Chicago. University of Chicago.

Schütze, Carson. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Sprouse, Jon, Matt Wagers, & Colin Phillips. in press. A test of the relation between working memory and syntactic island effects. *Language* .

Staum Casasanto, Laura, Philip Hofmeister, & Ivan A. Sag. 2010. Understanding acceptability judgments: Distinguishing the effects of grammar and processing on acceptability judgments. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, ed. by Stellan Ohlsson & Richard Catrambone, 224–229. Austin, TX: Cognitive Science Society.

Waters, Gloria S., & David Caplan. 1996. Processing resource capacity and the comprehension of garden path sentences. *Memory & Cognition* 24.342–355.

Whitney, Paul, Peter A. Arnett, Amy Driver, & Desiree Budd. 2001. Measuring central executive functioning: What's in a reading span? *Brain and Cognition* 45.1–14.

Whitney, Paul, Bill G. Ritchie, & Matthew B. Clark. 1991. Working-memory capacity and the use of elaborative inferences in text comprehension. *Discourse Processes* 14.133–145.